

**Investigating reliability and construct validity of a source-based academic writing test for
placement purposes**

by

Phuong Thi Tuyet Nguyen

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Applied Linguistics and Technology

Program of Study Committee:
Carol Chapelle, Major Professor
Evgeny Chukharev-Hudilainen
Amy Froelich
Volker Hegelheimer
Gary Ockey

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2021

Copyright © Phuong Thi Tuyet Nguyen, 20121. All rights reserved.

DEDICATION

To my parents, for always loving and supporting me.

Remember the dream I told you in our back yard when I was 12, Mom?

I did it!

TABLES OF CONTENTS

LIST OF FIGURES	vii
LIST OF TABLES	viii
ACKNOWLEDGEMENTS	ix
ABSTRACT.....	xii
CHAPTER 1: INTRODUCTION.....	1
Goals of the Study.....	5
Significance of the Study	7
Dissertation Outline	7
CHAPTER 2: LITERATURE REVIEW	10
Research on Source-Based Writing Constructs	11
Relationship between Source-based Writing Scores, Independent Writing Scores, and Independent Reading Scores.....	12
Test takers’ Cognitive Processes During Completion of Source-Based Writing Tasks.....	14
Discourse Analysis of Source-based Writing Task Responses.....	16
Reliability Research in Source-Based Writing Assessment	18
Strengths and Weaknesses of Eye-Tracking Technology.....	24
Eye-Tracking Technology	25
Eye-Tracking Measures	26
Eye-Tracking Technology in L2 Assessment Research	27
Test takers’ cognitive processes and behaviors during task completion	28
Raters’ cognitive processes and behaviors	32
An Argument-based Validity for Validation Research.....	34
Overview.....	34
Structure of an Argument-based Validity Inference.....	35
Strengths of Argument-based Validity for Test Validation.....	36
Research Goals and Research Questions	38
The EPT Writing Construct	39
Theoretical perspective adopted for the definition of the EPT Writing construct.....	40
Description of the EPT Writing construct	41
An Interpretive/Use Argument for a Source-based Academic English Writing Test for Placement Purposes	46
The Interpretive/Use Argument for the EPT Writing.....	48
The three inferences for reliability and construct validity	50

Research Questions	54
Chapter Summary	56
CHAPTER 3: METHODOLOGY	57
Research Design.....	57
The EPT Writing Tasks	60
Participants.....	60
Materials and Instruments.....	61
Essay Task (Task 2) Responses	61
Rating Rubric	62
Web-based Training Materials.....	64
Eye-Tracking Equipment and Software.....	64
Interview Protocol.....	65
Procedures.....	65
Operational EPT Writing Ratings.....	66
Experimental Ratings, Stimulated Recalls, and Interviews	68
Data Preparation and Analyses	73
RQ1: What Did Experts Think About the Appropriateness of The Rating Scale for Evaluating Test takers' Performance?	75
RQ2: Was There Statistical Evidence Supporting the Number of Score Bands?.....	76
RQ3: Were Raters Comparable and Consistent in Their Rating?.....	78
RQ4: To What Extent Did Raters Exhibit Bias Against Certain Test Tasks?	79
RQ5: What Was the Score Reliability When Two Tasks Are Included, and Each Response Double-Rated?	80
RQ6: What Writing Features Did Raters Attend to When Rating Source-Based Writing Task Responses?	81
Effects of the Rating Medium (Paper-based vs. Computer-based) on the Raters	81
Writing Features Raters Attended to When Rating Source-Based Writing Task Responses.....	82
Chapter Summary	87
CHAPTER 4: RESULTS AND DISCUSSION	88
Raters' Opinions about the Appropriateness of the Rating Scale.....	89
Descriptor Clarity.....	90
Criteria Relevance.....	93
Number of Score Bands	97
Score Band Labeling.....	101

Weighting of Rubric Categories	102
Statistical Evidence Supporting the Number of Score Bands.....	105
Raters' Comparability and Consistency in Their Rating	109
Rater Comparability.....	109
Rater Consistency	114
Rater Bias.....	115
Score Reliability with Two Tasks x Two Raters Design.....	119
Raters' Attention to Writing Features.....	124
Rating Medium and Rater Severity.....	124
Writing Features Attended to by Raters.....	126
Grammar and lexis	129
Arguments and organization.....	132
Discourse synthesis.....	136
Citation quality.....	136
Source text understanding	140
Style and conventions	141
Other features.....	144
Summary.....	145
Chapter Summary	146
CHAPTER 5: CONCLUSION.....	149
A Partial Validity Argument for the Source-based Academic Writing Test for Placement Purposes	149
The Test and Its Uses.....	150
The Partial Validity Argument.....	150
Evaluation inference	151
Raters' opinions on the appropriateness of the rating rubric	152
Adequacy of the rubric for distinguishing among the writing proficiency levels.....	153
Raters' comparability and consistency.....	154
Raters' bias against task type	155
Generalization inference	155
Explanation inference	157
Implications.....	158
A Construct of Source-based Academic Writing.....	158
Data Collection Methods in Cognition Research.....	162

Validation Research on Language Assessment	163
Recommendations for the EPT Writing.....	166
Revision of the Test Construct and Rating Rubric	167
Rater Training	168
Directions for Future Studies	169
Rater Cognition in Source-Based Writing Assessment	169
Validation Issues for The Source-Based Academic Writing Test for Placement Purposes	170
Conclusion	177
REFERENCES.....	178
APPENDIX A: WRITING PROMPT USED IN THE EYE TRACKING SESSION	189
APPENDIX B: RATING RUBRIC	191
APPENDIX C: RATER INTERVIEW PROTOCOL	192
APPENDIX D: TRANSCRIPTS OF RATERS' INTERVIEWS	193
APPENDIX E: TRANSCRIPTS OF RATERS' VERBAL REPORTS	212
APPENDIX F: INTERVIEWS CODING SCHEME	372
APPENDIX G: FINAL LIST OF WRITING FEATURES RATERS ATTENDED TO ...	375
APPENDIX H: WALD STATISTICS COMPARING PAIRS OF RATERS	378

LIST OF FIGURES

Figure 2.1. Structure of an inference in argument-based validity.....	35
Figure 2.2. The EPT Writing construct of source-based academic writing ability.....	42
Figure 2.3. Components of the language knowledge and fundamental processes measured by the EPT Writing	44
Figure 2.4. Structure of the IUA of the EPT Writing	47
Figure 2.5. Structure of the evaluation inference in the IUA for the EPT Writing	51
Figure 2.6. Structure of the generalization inference in the IUA for the EPT Writing	52
Figure 2.7. Structure of the explanation inference in the IUA for the EPT Writing	53
Figure 3.1. Mixed-methods multiphase research design	58
Figure 3.2. Procedure for collecting Phase 2 data	69
Figure 3.3. Illustration of the eye-tracking and essay-rating setup.....	70
Figure 3.4. Screen capture of a video recording Gazepoint GP3.....	71
Figure 3.5. Steps for analyzing verbal reports	82
Figure 4.1. Probability curves for individual score bands from MFRM analysis.....	108
Figure 4.2. Wright map output from MFRM analysis	110
Figure 4.3. Rater severity estimates and 95% confidential intervals	111
Figure 4.4. Interaction between raters and tasks from MFRM analysis	116
Figure 4.5. Dependability change relative to numbers of tasks and raters	123
Figure 4.6. Interaction between raters and rating medium from MFRM analysis.....	125
Figure 5.1. Evaluation inference with two assumptions and backing.....	152
Figure 5.2. Generalization inference with one assumption and backing	156
Figure 5.3. Structure of the explanation inference in the IUA for the EPT Writing	157
Figure 5.4. A revised source-based academic writing construct	160
Figure 5.5. A partial validity argument for the EPT Writing.....	164

LIST OF TABLES

Table 2.1. Warrants, Assumptions and Research Questions Associated with the Evaluation, Generalization, and Explanation Inference	55
Table 3.1. Raters' Background Information (N = 9)	61
Table 3.2. Rating Design for Fall 2018 EPT Writing in Fall 2018.....	67
Table 3.3. Summary of Analyses Conducted to Address the Research Questions.....	74
Table 3.4. Aspects of the Test Construct, Their Definitions, and Coding Scheme Categories	84
Table 4.1. Summary of Raters' Comments on The Rubric.....	89
Table 4.2. Score Band Statistics	106
Table 4.3. Wald Statistics for Pairs of Raters	112
Table 4.4. Measurement Results for the Rater Facet	114
Table 4.5. Analysis of Raters' Bias towards Tasks from Facets	117
Table 4.6. Descriptive Statistics for the Ratings Used in the G-study.....	120
Table 4.7. Univariate G-Study Results for One Task and One Rater	120
Table 4.8. D-Study Results for Various Numbers of Tasks and Raters	122
Table 4.9. Analysis of Raters' Bias against Rating Medium from Facets.....	126
Table 4.10. Distribution of Content Units (N = 1,483) across Nine Raters and 10 Essays	127
Table 4.11. Writing Features Attended to by Nine Raters When Rating 10 Essays.....	128
Table 5.1. Additional Warrants, Assumptions, and Backing Associated with Each Inference in the IUA for the EPT Writing	171

ACKNOWLEDGEMENTS

This dissertation has been a long, challenging, yet rewarding journey for me. On this dissertation journey, I have many people to thank for their various roles in helping me reach the destination.

First, I would like to express profound gratitude to my Committee Chair, Professor Carol Chapelle, for her guidance and support. She has been the first responder who has talked me through tough challenges I encountered during the journey. She has been very generous and flexible with her time, encouraging in every interaction, and has put me on the top of her priority list as evident by her overnight feedback on my work. Also, she has been the inspiration for my academic career since the moment I read her book in a small library back in Vietnam years ago. It has been a true honor to have her as my advisor and the role model for my academic career.

I am also indebted to my Committee members who have supported me and prepared me with the equipment, the skills, and the knowledge I needed to complete the dissertation. I would like to thank Professor Gary Ockey for equipping me with knowledge about test development and data analysis. He saw the potential in the test that I developed and recommended it to the then English Placement Test (EPT) Director, which gave me the wonderful opportunity to see it go operational. I am thankful to Professor Volker Hegelheimer, who has worn many hats for me – a professor, a boss, a mentor, and a confidant – for his adoption of my test for the EPT and his generous support with my use of the operational test scores for my dissertation. My sincere appreciation also goes to Professor Evgeny Chukharev-Hudilainen, who assisted me with cognition data collection and eye-tracking data analysis. He once or twice had to act as my technical support for the eye tracking data collection. Last but not least, I am grateful to

Professor Amy Froelich for her advice on statistics and her patiently explaining mixed-effects modeling to me.

I would also like to offer my appreciation to those who thought my dissertation journey worthwhile and were willing to financially invest in it. Specifically, I thank Education Testing Services, who provided a TOEFL Small Grant for Doctoral Research in Second Language Assessment and the Applied Linguistics and Technology and TESL Program, who provided an ALT/TESL Small Grant. Their generous support helped me have the data coded and analyzed in a timely manner.

My sincere thanks go to the raters who were willing to participate in my study. Without them, this dissertation would not have been possible.

I would also like to thank colleagues, friends, and staff at Iowa State University and elsewhere for making my journey smoother and more enjoyable:

- Yongkook Won, who has (I believe) voluntarily listened to my ideas and who has been my after-hours stat person,
- Roz Hirsch, Ananda Muhammad, and Idee Edalatishams, who were great companions during the dissertation writing process,
- Kelly Cunningham, who volunteered to be my personal deadline reminder and who has constantly nudged me to the finish line,
- Ahmet Dursun and Cathy Baumann, who have been very generous and supportive to allow me to work on the dissertation during work hours, and
- Teresa Smiley and Lisa Elm, who helped me sort things out when I got lost with paperwork for exams and graduation.

Finally, my special thanks go to my family. I am indebted to my parents, Tho Nguyen and Vang Bui, who gave me every opportunity to grow and who always support me of my choices and endeavors. I would not have been able to finish this dissertation had it not been for my mom, who is (fortunately for me) still stuck here in the U.S. because of the pandemic but on the bright side, will get to see me graduate. I am deeply grateful to my husband, Thien Le, for his patience and support, as well as our daughter Norah May, for the kisses and cuddles and for pushing me to excel at multitasking.

ABSTRACT

Source-based writing, in which writers read or listen to academic content before writing, has been considered to better assess academic writing skills than independent writing tasks (Read, 1990; Weigle, 2004). Because scores resulting from ratings of test takers' source-based writing task responses are treated as indicators of their academic writing ability, researchers have begun to investigate the meaning of scores on source-based academic writing tests in an attempt to define the construct measured on such tests. Although this research has resulted in insights about source-based writing constructs and the rating reliability of such tests, it has been limited in its research perspective, the methods for collecting data about the rating process, and the clarity of the connection between reliability and construct validity.

This study aimed to collect and analyze evidence regarding the reliability and construct validity of a source-based academic English test for placement purposes, called the EPT Writing, and to show the relationship between these two parts of the study by presenting the evidence in a validity argument (Kane, 1992, 2006, 2013). Specifically, important reliability aspects, including the appropriateness of the rating rubric based on raters' opinions and statistical evidence, the performance of the raters in terms of severity, consistency, and bias, as well as test score reliability, were examined. Also, the construct of academic source-based writing assessed by the EPT Writing was explored by analysis of the writing features that raters attended to while rating test takers' responses.

The study employed the mixed-methods multiphase research design (Creswell & Plano Clark, 2012) in which both quantitative and qualitative data were collected and analyzed in two sequential phases to address the research questions. In Phase 1, quantitative data, consisting of

1,300 operational ratings provided by the EPT Office, were analyzed using Many-Facets Rasch Measurement (MFRM) and Generalizability theory to address the research questions related to the rubric's functionality, raters' performance, and score reliability. In Phase 2, 630 experimental ratings, 90 stimulated recalls collected with assistance from records from eye-tracking technology, as well as nine interviews from nine raters were analyzed to address the research questions pertaining to raters' opinions of the rubric and the writing features that attracted raters' attention during rating. The findings were presented in a validity argument to show the connection between the reliability of the ratings and the construct validity, which needs to be taken into account in research on rating processes.

Overall, the raters' interviews and MFRM analysis of the operational ratings showed that the rubric was mostly appropriate for providing evidence of variation in source-based academic writing ability. Regarding raters' performance, MRFM analysis revealed that while most raters maintained their comparability and consistency in terms of severity, and impartiality towards the writing tasks, some of them were significantly more generous, inconsistent, and biased against task types. The score reliability estimate for a 2-task x 2-rater design was found below the desired level, suggesting that more tasks and raters are needed to increase reliability.

Additionally, analysis of the verbal reports indicated that the raters attended to the writing features aligned with the source-based academic writing construct that the test aims to measure. The conclusion presents a partial validity framework for the EPT Writing, in addition to implications for construct definition of source-based academic writing tests, cognition research methods, and language assessment validation research. Recommendations for the EPR Writing include a clearer definition of the test construct, revision of the rubric, and more rigorous rater training. Suggested directions for future research include further research investigating raters'

cognition in source-based writing assessment and additional validation studies for other inferences of the validity framework for the EPT Writing.

CHAPTER 1: INTRODUCTION

Writing is a constant requirement for active participation in all aspects of today's society as we consistently engage in textually mediated communication: We often write in different genres for different audiences and different purposes with an increasing variety of writing tools (UNESCO, 2004; Yi, 2010). In higher education, students' disciplinary knowledge and understanding are mostly exhibited and assessed through the medium of writing. Therefore, they are required to engage in various types of writing in which they need to meet both general academic convention and disciplinary writing requirements in order to be successful.

Among the writing activities that students need to undertake, source-based writing is an integral part of academic writing in tertiary education. In source-based writing, students are expected to read or listen to academic content pertaining to the topic under investigation prior to writing in response to a prompt. In their response, students are supposed to synthesize and select the information presented in the external sources to support their own arguments. Thus, in source-based writing tasks, the input materials include a significant proportion of language and students are required to use and transformed the source materials to complete the writing task. This type of writing is commonly required of students as course assignments in higher education courses (Shi, 2004; Wette, 2010).

Mirroring the prevalent use of source-based writing at post-secondary education level, the field of second language (L2) writing assessment has started to adopt source-based writing tasks in both influential, large-scale assessment programs and local, lower stakes language assessments. This change is reflected by the introduction of the integrated writing task in the Test

of English as a Foreign Language (TOEFL), a large-scale, high-stakes proficiency English test used for admissions to universities in English speaking countries. Similarly, this task type is also used in the Canadian Academic English Language (CAEL), another influential test for admission to an English-medium university. Additionally, source-based writing is currently employed in many university-based assessment programs in the U.S., such as in Georgia State University, Indiana University, the University of Illinois Urbana-Champaign, the University of Iowa, and Washington State University.

Source-based writing is considered to better assess the writing skills required of prospective university students than impromptu writing-only tasks, which require test takers to respond to a prompt on a topic unrelated to previous instruction. Specifically, source-based writing might help reduce the effects of differences in background knowledge among the test takers because it provides test takers with content materials to work with and may better simulate the process of academic study that students are required to undertake in the real world compared to the stand-alone writing test (Read, 1990; Weigle, 2004). Some researchers have argued that such benefits mean that scores from source-based writing tests can be extrapolated beyond performance on the test to test takers' real-world academic writing performance.

Despite the apparent benefits of source-based writing assessments, their use assumes that test takers actually engage their academic writing abilities when they take source-based writing tests and that raters evaluate the relevant aspects of the complex writing in the response. Scores resulting from ratings of test takers' responses are treated as indicators of their academic writing ability even though the meaning of the scores is not entirely clear. Researchers have begun to investigate the meaning of scores on source-based academic writing tests in studies attempting to define the construct measured by such tests. These studies examining source-based writing

constructs have mainly focused on three areas: (1) the relationship between source-based writing scores, independent writing scores, and independent reading scores (2) test takers' cognitive processes during completion of the writing task, and (3) discourse analysis of source-based writing task responses. These lines of research have come to the following conclusions: First, assessment of the source-based writing construct cannot be based solely on independent reading and writing measures. Second, a construct of source-based writing is related but different from a construct measured by independent writing. Third, a construct of source-based writing ability could include reading ability and discourse synthesis ability because the indirect source use, which requires more advanced reading ability, was statistically different among writing proficiency levels. Finally, test takers' indirect source used in their written responses is impacted by their reading ability.

Previous research has been also conducted quantitatively and qualitatively to examine the reliability of scores on source-based writing tests. Some topics that quantitative studies have focused on include the effects of rater differences in terms of education experience and rating experience on the rating criteria, the comparability of source-based writing tasks and independent writing tasks in terms of score generalizability, and the agreement among raters as revealed in quantitative analysis of ratings. Additionally, raters' cognition during their decision-making in the rating process has been investigated to learn about the accuracy of rating guidelines, develop a rating rubric, or assemble scoring validity evidence. These studies had practical implications pertaining to test development, rating scale development, and rater training for the source-based writing test under investigation.

Although the research in these areas has resulted in many informative observations about source-based writing constructs and rating reliability on tests similar to the academic writing test

investigated in this study, it has been limited in three important ways. First, no studies to date have been conducted to investigate source-based writing constructs from expert raters' perspectives. Examining raters' decision-making processes during response evaluation can provide useful information about the construct that a test measures because these decisions result in the scores that are awarded to test takers (Knoch & Chapelle, 2017). Researchers therefore study raters' introspection to reveal the aspects of test takers' responses that raters attend to in making their rating decisions. The raters' judgement about exactly what to attend to and how to evaluate certain features of the response results in the scores they assign the test takers' performance. In other words, it is not the test task or the test takers' actual performance on a writing alone that determines their scores. It is the raters' judgements about their performance on a particular task that results in the score, which is then used to make decisions about the test taker. If raters' verbal protocols show that raters fail to attend to the key aspects of performance underlying the construct that the test is intended to assess or if raters attend to different aspects from those specified in the rating rubric, then one conclusion could be that the test construct is not well defined and communicated to the raters. Therefore, studies of raters' cognitive processes are needed to demonstrate the extent to which raters' focus of attention during rating is aligned with the construct as it was defined by the test developers and communicated to raters through the rubric.

Second, raters' cognition research exploring reliability issues has limitations in terms of methods used for data collection. The few studies investigating raters' decision-making processes when rating examinees' responses (Cumming, Kantor, & Powers, 2001; Green, 1998; Gebril & Plakans, 2014) mainly adopted the concurrent think-aloud method, which requires raters to describe to the researcher what they are thinking as they complete the rating process.

This method is used to obtain data revealing what the raters are attending to in the written responses and how those features are relevant to decisions about the scores they award to the responses. Nevertheless, the use of the concurrent think aloud method may actually change what the raters attend to, failing to reveal the normal rating process of interest (Barkaoui, 2011; Lumley, 2005). Therefore, studies which adopt a combination of different methods of cognition data collection, such as eye-tracking and stimulated recalls, are needed to provide a more complete picture of the processes that raters undertake when making decisions about test takers' performance.

The third limitation lies in the connection between findings in past studies investigating rating processes on source-based writing tests and the construct validity of the resulting scores. These studies have been designed to investigate either the reliability of the rating processes by exploring raters' thought processes and the resulting test scores (Barkaoui, 2010; Gebril & Plakans, 2014; Green, 1998) or the construct validity of the resulting scores by examining the extent source-based test scores are correlated with independent writing test scores (Asención, 2008; Shin & Ewert, 2015; Watanabe, 2001). Each of these aspects of the rating process is important, and so is their connection to one another. The connection between the reliability of the ratings and their construct validity needs to be taken into account in research on rating processes.

Goals of the Study

This study addresses the need for more studies investigating source-based writing test constructs in an investigation of a new writing placement test for incoming students at a large Midwest university. It advances research in this area by providing insights into source-based

writing constructs from raters' perspectives, improving the methodology for data collection of rating processes, and framing the study to take into account both reliability and construct validity. The test under investigation, a new source-based English Placement Test (EPT) Writing, was adopted in Fall 2015. The test was intended to measure test takers' source-based academic writing ability for use in placing students into appropriate English-as-a-second-language (ESL) writing courses. The development entailed review of best practices and previous research on source-based writing tests, and its implementation included investigation of the reliability of the results. The study was, therefore, able to draw upon the operational testing program to undertake the research required to learn more about the construct meaning of the test scores.

The goals of my study were to collect and analyze evidence regarding the reliability and construct validity of the EPT Writing test and to show the relationship between these two parts of the study by presenting the evidence in a validity argument (Kane, 1992, 2006, 2013). In particular, I investigated critical aspects of reliability, including (1) the appropriateness of the rating scale, (2) the performance of the raters in terms of severity, consistency and bias, and (3) test score reliability. I examined the construct of the EPT academic source-based writing test by studying raters' decision-making processes while they rated test takers' responses. To collect reliable data on raters' cognition without interfering with their rating process, I used eye-tracking technology and stimulated recall: raters recalled their thinking during rating by being prompted by the record of their gazes obtained by the eye-tracking. The findings were presented in a validity argument to show the connection between the reliability of the ratings and the construct validity of the scores. The structure for the validity argument is first outlined in an interpretive/use argument (IUA) created by the researcher (Kane, 2006, 2013).

Significance of the Study

The study has important implications for language testing research and the local EPT Writing. In terms of source-based writing assessment research, the study contributes to our understanding of source integration ability because of the findings pertaining to raters' decision-making process. Additionally, this study demonstrated how eye-tracking technology can be used to prompt data collection on raters' cognition, especially in source-based writing assessment. Regarding language test validation research, this study showed how reliability-related findings could be connected to findings about construct validity using the argument-based approach to validation.

This study also has more local implications for the EPT Writing. First, it provides insights into the test construct as it is operationalized by the raters. This information is valuable to the revision of the rating scale because raters reporting on features of test takers' responses that they attended to allows test developers to identify aspects of source-based writing that need to be revised or added in the rating scale. Second, it helps identify areas where raters need additional training. More importantly, it helps build a validity argument for the EPT Writing which is needed to understand the meaning of the test scores before making decision about test takers' placement in the ESL Writing courses.

Dissertation Outline

The first chapter has introduced the issues in construct and reliability research in source-based writing assessment that foreground the study as well as outlined the overall purpose and significance of the dissertation. The remainder of the dissertation is organized as follows.

Chapter 2 consists of five parts. The first two parts provides a review of the research in source-based writing assessment constructs and reliability research in source-based writing. The third part introduces eye-tracking technology, including its technicality and measures as well as its use in L2 assessment research. The fourth part provides an overview of an IUA framework for validating rating processes. The last part presents the study goals and six research questions this study aimed to address.

Chapter 3 focuses on the methodology used for data collection and analyses. It presents the overall design, context, participants, along with the materials and instruments used for data collection. It also describes the procedure for collecting the operational and experimental ratings of test takers' performance, interviews, and stimulated recalls. The chapter ends with a detailed explanation of the types of data analyses conducted to answer the six research questions.

The next chapter reports on and discusses the results of data analyses described in the previous chapter. Specifically, it presents the results of the analyses conducted to examine the appropriateness of the rating scale and raters' performance based on raters' interviews and the EPT operational ratings. Additionally, it reveals evidence related to the rating reliability resulted from the analyses of EPT operational ratings. It also provides results of the qualitative analysis of raters' verbal reports conducted to investigate the construct of source-based academic writing assessed by the test.

The dissertation will conclude in Chapter 5 which summarizes and highlights the main findings of the analyses as well as interprets them for their degree of support for the validity argument for the EPT Writing. This chapter also discusses the implications the study has for source-based writing construct research, cognition research, as well as language testing

validation research. Additionally, it provides recommendations for the local EPT Writing test. The chapter ends with a discussion of the directions of future research.

CHAPTER 2: LITERATURE REVIEW

This study investigated the construct assessed by an English source-based academic writing test and the reliability of the ratings awarded by human raters. Previous studies have examined the constructs in source-based writing tests in three ways: 1) exploring relationships between source-based writing scores and scores on tests of constructs such as independent writing, and independent reading, 2) investigating test takers' cognitive processes during test taking, and 3) analyzing the discourse of test takers' responses. Previous research has also examined the reliability of scores on such tests by investigating the agreement among raters as revealed in quantitative analysis of ratings and qualitative analysis of raters' cognition during their decision making in the rating process. Although these studies have provided useful information on the constructs and reliability of source-based writing assessments, they have three major limitations. First, studies investigating raters' decision-making processes have focused solely on reliability but have not been used to gain insights into source-based writing test constructs even though the rating process is recognized to contribute to the construct meaning of the scores. Second, studies examining raters' cognition during rating source-based written responses have limitations in terms of data collection methods as they have mainly employed think-aloud protocols, which can affect the cognitive processes under investigation. Third, studies of rating reliability in source-based writing tests have failed to connect their findings on reliability to construct validity. This study aimed to address these limitations by examining the construct assessed by an English source-based academic writing placement test using eye-tracking technology and stimulated recall as well as investigating the reliability of the rating process and rating results. The findings related to raters' cognition and rating reliability are presented in a validity argument to provide a clear connection between the reliability of the

ratings and the test construct, which enhances understanding of the test score meaning. Such an investigation requires a review of research investigating source-based writing constructs, reliability research in source-based writing assessment, eye-tracking technology, and the interpretive/use argument for test validation studies for the test under investigation.

This chapter consists of five sections. It starts with a review of source-based writing construct studies investigating the relationship between source-based writing scores and scores from independent reading and independent writing tasks, test taskers' cognitive processes during their performance on a source-based writing test, and discourse analysis of source-based written responses. Next, it explores reliability research in source-based writing with an analysis of existing methodological approaches to investigating raters' cognition in source-based writing research. The next section reviews topics related to eye-tracking technology as a tool for collecting cognition data, including eye-tracking technicality and measures along with the use of eye-tracking technology in L2 assessment research. The chapter then provides an overview of the interpretive/use argument for test validation research with a discussion of its strengths. The last section presents the construct that the test under investigation is intended to measure, the research goals, and the six research questions.

Research on Source-Based Writing Constructs

With the widespread use of source-based writing tasks especially in high-stakes testing, L2 writing assessment researchers are interested in investigating source-based writing constructs measured by these writing tasks (for example, Esmaelii, 2004; Gebril & Plakans, 2014; Plakans, 2008, 2009). Overall, L2 writing assessment studies examining source-based writing constructs have mainly focused on three areas: (1) the relationship between source-based writing score and

independent writing score, (2) test takers' cognitive processes during completion of the source-based writing task, and (3) discourse analysis of source-based writing task responses.

Relationship between Source-based Writing Scores, Independent Writing Scores, and Independent Reading Scores

Researchers have investigated the relationship between source-based writing scores, independent reading scores, and independent writing scores to explore the extent to which the source-based writing construct is the sum of one's reading and writing abilities or an independent construct. The underlying assumption for these studies is that if there is a strong correlation between source-based writing, independent writing, and independent reading skills, then the source-based writing task may tap into both reading and writing abilities.

Several studies on the relationship between general reading scores and source-based writing scores have revealed a strong relationship between overall reading comprehension scores and source-based writing test scores. For instance, Trites and McGroarty (2005), in a study determining the influence of reading ability on source-based writing ability, found that their source-based writing scores had a correlation of between .68 and .70 with the Nelson-Denny reading test, TOEFL reading comprehension, and two other measures of basic reading comprehension scores. This finding suggested a relationship between all reading measures and the source-based writing scores as well as a possible distinction between independent reading ability and source-based writing ability. This relationship was also found in a subsequent study by Sawaki, Quinlan, and Lee (2013) who examined the factor structures of rating from humans and automated scores of the TOEFL iBT integrated (source-based) essay responses, and independent reading and listening comprehension test scores. They found that the comprehension factor, which underlays factors representing the essay content as well as reading and listening

comprehension, correlated with two related yet distinct writing factors, namely Productive Vocabulary and Sentence Conventions, at between .49 and .72. They also detected an interrelationship of written content, reading, and listening, with the correlation indices ranging from .83 to .87. These findings indicated that a skill of identifying appropriate information from a spoken lecture and a written source text to complete the source-based writing task is closely related to abilities measured in reading and listening comprehension tests. More recently, Shin and Ewert (2015) examined the interrelationships among analytic measures of source-based writing task scores and reading comprehension test scores. They found that the composite analytic writing scores moderately correlated with independent reading scores at .68, indicating that performance on the source-based writing task seemed to tap into reading ability.

On the other hand, other studies have shown a weak association between general reading comprehension scores and source-based writing test scores. For example, Asención (2008) reported low positive correlations between reading scores and the scores of two different types of writing tasks based on the same source text, a summary and a response essay, at .28 and .38 respectively. The researcher argued that although source-based writing ability involves some reading for comprehension, it implies more than this kind of reading as the writer needs to read with a writing goal in selecting information from the source that can help them write their texts. Watanabe (2001), based on his finding that reading proficiency only explained 1% or 2% of the total variances in the two different source-based writing task scores, argued that reading scores alone cannot reliably predict scores on a source-based writing task.

While the relationship between reading scores and source-based writing scores is inconclusive, researchers have found high correlation between source-based writing scores and independent writing scores. For example, Watanabe (2001) analyzed test takers' scores on a

source-based writing task, an independent writing test, and independent reading test. He found that the writing scores were a significant predictor of reading-to-write performance and concluded that the source-based writing tasks in his study were a more reliable measure of writing ability than of reading ability. In the same line, Shin & Ewert (2015) reported a moderate correlation between composite analytic scores on a source-based writing test and independent writing scores, at .65. It is evident from these studies that there is at least a moderate relationship between source-based writing and independent writing. However, test takers did not always seem to perform equally on both tasks. In fact, Brown, Hilgers, and Marsella's (1991) repeated ANOVA analyses of test takers' scores on source-based writing tasks independent writing tasks showed that the performance of test takers on the source-based tasks was significantly different from their performance on the independent tasks. Additionally, Gebril (2009) found that his test takers scored higher on organization and idea development in the source-based writing tasks, leading to his speculation that students could model their writing on the source text.

Overall, research on the relationship between independent reading scores and source-based writing test scores has produced mixed findings. However, based on findings from previous studies, it can be concluded that source-based writing and independent writing are two related yet independent constructs.

Test takers' Cognitive Processes During Completion of Source-Based Writing Tasks

In addition to examining source-based writing constructs based on test takers' performance, researchers have also investigated this topic by analyzing test takers' cognitive processes during their completion of a source-based writing task to understand the demands of source-based writing tasks. These studies have found that source-based writing tasks are affected

by independent reading or writing ability in terms of how test takers select and edit information from the source text and combine that information into their own texts. For instance, Esmaelii (2002) explored the role of reading in a source-based writing task by focusing on the test takers' process when completing the task. His analyses of 34 test takers' writing strategies through a post-task questionnaire and interview showed that the participants, while utilizing various writing strategies, relied extensively on strategies for comprehending the source texts prior to writing. Based on this finding, Esmaelii proposed that a construct of source-based writing consists of two interwoven constructs of reading and writing.

Following the work of Esmaelii, Plakans (2008) compared test takers' processes in composing source-based writing and writing-only test tasks to inform issues of her test's construct validity. She identified differences in terms of test completion processes across tasks (source-based versus writing-only) and writers (more experienced versus less experienced with academic writing). Specifically, the source-based writing task elicited a more interactive process for some writers while the writing-only task required more initial and less online planning. She also found that the interactive process was more prevalent among test takers with more experience and interest in writing. Based on this finding, Plakans suggested that compared to independent writing, source-based writing (which requires test takers to read prior to writing) elicited a more authentic process as it allowed test takers with experience and interest to display their abilities interact with source texts and construct their meaning when writing. In her subsequent study, Plakans (2009), aiming to gain more insights into how test takers approached source-based writing tasks, explored the role of reading strategies in test takers' completion of such a task through think-aloud verbal protocols, interviews, and the resulting written products. The researcher found that compared to lower scoring writers, higher performing test takers used

more mining and global strategies, such as metacognitive and goal-setting strategies, suggesting that reading plays an important role in the process and performance of source-based writing tasks.

Discourse Analysis of Source-based Writing Task Responses

Another area of research that can shed light on source-based writing constructs is discourse analysis of responses to source-based writing tasks. Specifically, researchers analyze test takers' responses to examine the difference between high scoring and low scoring test takers in terms of source use. Such studies are conducted based on the assumption that verbatim use of source-based materials is indicative of lower reading ability and that test takers who are more proficient readers will be more skillful in using the source texts in their written response. Such evidence helps conclude that reading comprehension ability is an important factor in source-based writing as it differentiates successful test takers from less successful ones, and thus, reading ability must be included when test developers define their source-based writing test construct.

Analyzing students' summaries based on prior reading, Johns and Mayes (1990) classified the idea units in the summaries produced by high and low proficiency students into two general categories of source text content representation, namely "correct replications" and "distortions". They found that the lower proficiency writers had more instances of direct copying (or incorrect replication); however, the two groups did not differ in "correct paraphrasing". Additionally, the high proficiency writers used more combinations of idea units from the source texts; yet did not differ from the low proficiency writers in terms of distorting source text ideas. Weigle and Parker (2012) analyzed test takers' responses to a source-based writing task and used

one-way ANOVA to compare source-borrowing across various score levels. They found a relationship between score level and textual borrowing, with lower scoring students quoting longer sections of the source texts than higher proficiency students did. Similarly, in their study on source text use as a predictor of test scores, Gebril and Plakans (2013) found that source-based writing scores were highly impacted by writers' use of the source material, with low scoring writers depending heavily on the reading texts for content and direct copying of words and phrases. Researchers have concluded that indirect source use (i.e., use of ideas and paraphrases), which requires more advanced reading ability, differs significantly among test takers from different proficiency levels. This indicates that reading ability plays a role in test performance, and thus, a test construct definition must include it as part of source-based writing ability.

In sum, past studies investigating source-based writing constructs by focusing on the association between source-based writing scores, independent reading, and independent writing scores, examining test takers' cognitive processes, as well as analyzing the discourse of test takers' written responses have arrived at the following conclusions. First, assessment of the source-based writing construct cannot be based solely on independent reading and writing measures. Second, source-based writing and independent writing are two related yet independent constructs. Third, reading ability plays an important role in how test takers process a source-based writing task since compared to more able writers, less proficient writers tend to copy directly more often, quote longer chunks of source texts, and use fewer idea units from the source texts. Finally, reading comprehension ability influences how test takers perform on a source-based writing task as test takers' indirect source use (i.e., paraphrasing and use of ideas) is impacted by their reading skills.

These studies have provided insightful information about source-based writing constructs. However, to date, no studies have examined source-based writing constructs from raters' perspectives. This line of research is essential as investigating raters' cognitive processes during response evaluation can provide useful information about the construct that a test measures because these decisions result in the scores that are awarded to test takers (Knoch & Chapelle, 2017). Studying raters' cognition during evaluation of test takers' responses can reveal the aspects of test takers' responses that raters attend to in making their rating decisions. It is raters' decisions of which aspects of test takers' response to attend to and how to assess these aspects that result in the final score, which is then used to make decisions about the test taker. If raters' verbal reports indicate raters' failure to attend to the key aspects of performance underlying the construct the test is intended to measure or if their focus is on aspects different from those described in the rating rubric, one possible conclusion will be that the test construct is not well defined and communicated to the raters. Therefore, the study of raters' cognitive processes is essential in demonstrating the extent to which raters' focus of attention during rating is aligned with the construct defined by the test developers and communicated to raters through the rubric.

Reliability Research in Source-Based Writing Assessment

Reliability plays an integral role in test score interpretation. It refers to the degree to which errors are absent from the test taking and rating processes as well as the test scores (Chapelle, 2021). Absence of errors in the test taking process can be a result of consistent test administration procedures and conditions, such as test proctor training, test material preparations, and test security measures. In performance assessment, a rating process has little error when raters adhere to the criteria delineated in the rating scale and agree with one another in terms of

scores awarded to test takers. Test scores are deemed reliable when they reflect the true score, or test takers' ability, while the effects of other factors, such as raters, test tasks, test forms, testing occasions, and their combinations, are minimal. Only when the test taking procedure, rating processes, and resulting test scores are affected by little error can valid inferences about test takers' ability be made based on their test performance. In source-based language assessment, both quantitative and qualitative research has been conducted to investigate issues related to the reliability of the scoring processes and final scores.

Quantitative studies investigating the rating process have generally found that raters are able to produce reliable ratings of source-based written responses. In fact, it has been revealed that raters can be at least as reliable when rating source-based writing test responses as when rating independent writing test responses. For example, Gebril (2009) designed a univariate generalizability study to examine how source-based writing tasks were comparable to independent writing tasks in terms of score generalizability. Analyzing holistic scores provided by three raters to 115 test takers' responses to both the source-based writing task and independent writing task, he found that the former tasks yielded as reliable scores as the independent tasks. This conclusion was echoed in his subsequent study (Gebril, 2010). In some cases, raters have been found to be even more reliable in their evaluation of source-based writing test performance than that of independent writing responses, as demonstrated by a study by Weigle (2004). In this study, she compared rater reliability across test prompts in an independent writing test and a source-based writing test. A comparison of the percentages of agreement among raters showed that while the overall agreement rate between two raters was 79% for the independent writing test, this index was remarkably higher for the source-based writing test, at 95%. This finding indicated that the scoring on the source-based writing test yielded higher

consistency and that there was very little score variance that could be attributed to the rating process.

In addition to comparing score reliability in source-based writing tasks and independent tests, researchers are also interested in examining how scoring conditions affects the reliability of the ratings. Gebril (2010) explored score reliability in two different rating schemes—two different groups of raters scoring each task type (source-based and independent task) versus the same raters scoring both task types using generalization theory (G-theory), a measurement model for estimating the effects of a test taker’s ability, and systematic and random sources of measurement errors on test scores via the estimated variance components contributed by the object of measurement (i.e., test takers), the facets (e.g., test tasks and raters), and their combinations. Results from a multivariate generalization analysis of the test scores demonstrated that the two rating schemes yielded similar reliability coefficients, supporting the use of different groups of raters for different task types to speed up the rating process, minimize rater workload, and make rater training more time efficient. More recently, Shin and Ewert (2015) examined raters’ change in terms of severity across different criteria of an analytic rubric when rating source-based written responses. Their analysis of analytic scores of 83 test takers’ performance using G-theory revealed that in general, raters assigned scores neither too harshly nor too leniently across each analytic rating criteria. However, they found that the person and rater facets contributed to score variability differently in certain analytic categories, with raters being either relatively harsher or more lenient in reading-related criteria on the rubric (i.e., viewpoint recognition and engagement) than writing-related criteria (i.e., *organization*, *development*, and *language use*) on the source-based writing task. However, the researchers did not identify any patterns of rating severity among raters.

From a qualitative approach, researchers have also investigated raters' cognition to find evidence for the rating process reliability. In particular, raters' cognitive processes have been analyzed to examine the accuracy of rating guidelines. For instance, Green (1998) reported findings from a study investigating raters' marking strategies while scoring a source-based writing task from Cambridge's Certificate of Advanced English (CAE) that required examinees to write a letter. Using concurrent think-aloud protocols (TAPs) with a group of raters, Green identified four main categories of rater behavior, namely marking behavior, textual features, evaluative responses, and meta-comments. Specifically, she found that raters' behavior was dominated by essay reading and attendance to linguistic appropriateness (in terms of register, grammar, and vocabulary), technical features (such as vocabulary, grammar, and spelling accuracy, and cohesion), and task realization (such as length, task completion, effect on readers, etc.). Additionally, Green also observed a difference in terms of rating strategy use between good raters and less effective ones and this difference was influenced by the quality of test takers' response quality. Specifically, good and poor raters did not differ in their use of strategies when grading well written responses. However, when grading poorly written scripts, good raters tended to use more strategies for assigning/reviewing grades while less effective examiners attended to language appropriateness, such as register, style, and vocabulary appropriateness, and technical features, such as grammar, vocabulary, and spelling. It should be noted that the study did not report the frequency of different strategies used by the raters.

In a more recent study, Gebril and Plakans (2014) investigated two raters' approach to scoring responses to source-based, reading-to-write tasks and the features influencing their scoring decisions in order to assemble "scoring validity evidence for integrated tasks in L2 writing assessment" (p. 67). Note that "scoring validity" encompasses both reliable rating scale

and valid descriptor interpretation (Deygers & Van Gorp, 2015) and thus, essentially refers to reliability defined by Chapelle (2021). Employing an inductive analysis of interviews and think-aloud data, the researchers found that raters used strategies for assessing test takers' responses more frequently than those for comprehending the responses. They also identified three categories of source use-related strategies reported by raters, namely, (a) locating source information, (b) assessing citation mechanics, and (c) assessing quality of source use. However, the researchers also found that raters attended to source use differently depending on writer proficiency levels, as they focused on surface source use features at lower levels but shifted their attention to more sophisticated issues when responses appeared to reflect higher levels of ability. Gebril and Plakans (2014) argued that these findings demonstrate the complex nature of source-based writing tasks and emphasized the need for refining the rubrics to better reflect the source integration issues and training raters more carefully, which could lead to justifiable and well-articulated scoring decisions and improve score reliability.

Overall, the studies investigating raters' decision-making processes in source-based writing have provided important information on the reliability of the rating process. First, scores on source-based writing tests have been found to be at least as reliable as those on independent writing tests. Second, in instances where an analytic rubric is employed to score source-based written responses, there is no evidence suggesting that the ratings systematically vary in overall severity across different criteria of the rubric or that raters show identified patterns of rating severity across all students. Third, rubrics for source-based writing tasks should include source integration issues such as source information, citation mechanics, and source use quality, which raters must be trained on carefully to arrive at enhance score reliability. These findings have important implications for the use of source-based writing tasks in L2 assessment; however, they

had two limitations pertaining to the methods adopted for cognition data collection and the limited scope of each study.

The first limitation is related to the methods for collecting raters' cognition data. These studies adopted concurrent think-aloud protocols (Ericsson & Simon, 1984) to gather data about raters' cognition in source-based writing. This method has been the subject of controversy (Cumming et al., 2001; Lumley, 2005; Stratman & Hamp-Lyons, 1994) due to the issues of veridicality, or participants' thought processes not being fully or accurately expressed, and reactivity, or interference with the process being investigated. Specifically, TAPs suffer from veridicality since participants can only verbalize what they are aware of, or what they decide to report, and consequently fail to provide a complete picture of the rating process (Barkaoui, 2010; Lumley, 2005). Reactivity is another issue with this method because asking raters to verbalize their thoughts during the rating task alters the cognitive processes required to carry out the task (Lumley, 2005; Stratman & Hamp-Lyons, 1994), making rating more cognitively demanding for raters (Cumming et al., 2001). In fact, TAPs seem to affect various aspects of the rating process, including essay comprehension, rating criteria and writing aspects raters attend to, decision-making processes, rater confidence, as well as rater severity and consistency and to influence different raters differently (Barkaoui, 2011). An alternative method for collecting cognition data could be eye-tracking technology, coupled with another method for triangulation such as stimulated recalls, also known as retrospective verbal reports.

The second limitation pertains to the scope of these studies, specifically the transparency of the association between their findings and the validity of score interpretation and use of the test under investigation. Consequently, this has not allowed researchers to investigate both reliability of the rating process and its effect on the construct validity of the test inferences for

their uses. The past studies investigating rating processes on source-based writing tests have failed to connect the reliability of the ratings and their construct validity. For example, the study by Green (1998), conducted on the Cambridge's CAE Writing, focused more on a model of good marking behavior and raters' decision-making processes when rating scripts of different score levels. The author did not discuss the implications of her findings to the validity of the inference made from the test scores. Although the study by Shin and Ewert (2015) unraveled information about test score reliability, it did not provide implications for the validation of the test under investigation. Overall, discussion of the meaning and implication of the findings from these studies were decontextualized relative to the overall test validation, leaving the reader with responsibility of making inferences about the validity of the test score interpretation and use.

To address these limitations, this study employed eye-tracking data, combined with stimulated recall, to examine the rating criteria or aspects of language that raters attend to while rating test takers' performance on a source-based academic writing test. Also, the findings from the study will be presented in the validity argument (Kane, 1992, 2006, 2013) that makes an explicit connection between the findings on reliability and the construct validity. The next two sections discuss eye-tracking technology as well as provide an overview on argument-based validity for validation research.

Strengths and Weaknesses of Eye-Tracking Technology

Compared to TAPs, eye-tracking technology has many advantages. A major advantage of this technology is its ability to capture what a reader/viewer is focusing on without interfering with the reading process (Godfroid & Spino, 2015). Also, when used in combination with

stimulated recall, eye-tracking technology presents participants with detailed recordings of their eye movements, and thus, allows for comprehensive recall of their thought processes.

However, it should be noted that eye-tracking systems can only record explicit information about eye movements and fixation. They do not always provide information about participants' attention or cognitive processes. The use of eye-tracking technology to collect cognition data assumes that participants' eye movements are indicative of their visual attention. This assumption is not always reasonable as research has shown that one can cognitively concentrate on an area of interest while their gaze is directed elsewhere (e.g., Anderson, Bothell, & Douglass, 2004; Deubel, 2008). Also, records of participants' interaction with an area of interest provide no indication about how participants process the information presented on that area as information processing requires complex, higher level mental operations, such as analysis, synthesis, and organization, that cannot be observed (Chubb, 2013). Eye-tracking data do not always reveal what area of the text or visual that has the reader's/viewer's attention or how they process what they see. Thus, they should not be interpreted in isolation, but instead should be triangulated with other types of data such as verbal reports to gain insights into one's attention and cognitive processes.

Eye-Tracking Technology

Eye-tracking technology allows the researcher to record and examine participants' eye-movements. Collecting such data requires an eye-tracker, an electric device capable of recording eye movements, and software for recording and analyzing the data obtained from the eye tracker. Eye-movements can be recorded with invasive eye-tracking systems where the eye-tracker is attached to the participant's head or face, or with non-invasive systems where the participant does

not have any physical contact with the eye-tracker. No matter what type of eye-tracker is employed, the use of eye-tracking data as an indication of cognition is based on the assumption that eye movement is strongly associated with the human mind and that eye movements reflect ongoing cognitive processing (Pollatsek, Reichle, & Rayner, 2006; Reichle, Pollatsek, & Rayner, 2006). Eye-tracking technology has been used as a data collection instrument in such areas of applied linguistics as second language acquisition and learning. These studies investigated various topics, among which are learners' use of video caption (Winke, Gass, & Sydorenko, 2013), learners' noticing of feedback or grammar (Godfroid & Uggen, 2013; Smith, 2012), accidental vocabulary acquisition (Godfroid, Boers, & Housen, 2013), and grammatical processing during reading (Foucart & Frenck-Mestre, 2012; Frenck-Mestre, 2005).

Eye-Tracking Measures

It is essential for researchers utilizing eye-tracking technology in their study to understand eye-tracking measures prior to selecting what measures to analyze to address their research questions. Depending on the type of eye-tracker and the accompanying software for data recording and analyses, eye-tracking measures can be different. However, three main types that are frequently reported are saccades, fixations, and areas of interest (AOIs).

Saccades refer to rapid eye movements from one point to another. Saccades also include backward eye movements, called regressions. In reading for example, these movements can be as short as a few letters within a word, suggesting word-specific processing difficulties, or as long as words or sentences in a larger text, indicating that processing difficulties and comprehension failures with respect to a larger sentential context (Roberts & Sivanova-Chanturia, 2013).

Fixations are measures of time in between saccades where the eyes remain stationary. Usually,

the number and durations of fixations are reported to offer information regarding the features of the text or visual being viewed. Researchers believe that it is during fixations that the reader/viewer actually acquires new information (Rayner, 2009). These measures are considered “indirect measures of cognitive processes that cannot be directly assessed” (Holmqvist et al., 2011, p. 65).

Additionally, researchers, such as Winke and Lim (2015) and Ballard (2017), have also reported AIOs as their eye-tracking measures. AIOs refer to the regions in the text or visual that are of interest to the researcher. For example, AIOs can be used to examine what category of a rating scale that raters attend to the most frequently. This measure is usually used in association with other measures, such as number of fixations or time to first fixation.

In addition to the measures, visual displays of eye movements are also used to demonstrate the reader’s /viewer’s eye movement trajectories or their view time. The first type of visual representation is called *gazeplot* (also known as *scanpath*). A gazeplot shows participants’ paths of eye movements in time and space. It can be either static as an image of eye movements at a particular time point, or dynamic as a video recording which the researcher can replay to observe participants’ eye movements. Eye movements can also be displayed as a *heatmap*, which captures participants’ viewing at a specific time point. Heatmaps can also provide a visual summary of all the data over time.

Eye-Tracking Technology in L2 Assessment Research

Although eye-tracking technology has been employed extensively to investigate reading, its use in L2 assessment research has been less prevalent. Recently however, the use of eye-

trackers has gained attention from language testing researchers to investigate various aspects of language testing in many skill areas such as reading, listening, and writing.

Test takers' cognitive processes and behaviors during task completion

Eye-tracking technology has been utilized to elicit information about test takers' cognitive processes during task completion to gain insights about test constructs. In reading assessment, eye-tracking has been employed to investigate the test taker's cognitive processes while completing reading test items to find evidence that the test items elicit the range and level of cognitive processes relevant to the construct assessed by the test. The first study in this area, conducted by Bax and Weir (2012), examined the cognitive processes employed by test takers when completing a computer-based Cambridge CAE Reading test to decide if test items actually measure the cognitive processes pertaining to academic study in English. The researchers collected eye-tracking data using Tobii T60 eye-tracker and retrospective questionnaire responses from 35 participants. Analyses of the visual data, such as video recordings, gazeplots, and heatmaps, and the statistical data, such as *fixation duration* (i.e., duration of all eye fixations within an area of interest), *time to first fixation* (i.e., time before a participant fixates his or her eyes on an area of interest), and *visit count* (i.e., the number of eye visits within an area of interest), for five test items showed that the items were effective in eliciting both the range of cognitive processing (i.e., skimming, scanning, and search reading) and also the different levels of processing from lower areas to more complex levels (i.e. word-level to whole-text comprehension).

In a subsequent study, Bax (2013) investigated the difference in test takers' cognitive processing while they completed onscreen IELTS reading test items with the goal of to evaluate

the cognitive validity of the reading test items. Seventy-one Malaysian students completed an onscreen test consisting of two IELTS reading passages with 11 test items. A Tobii T60 tracked eye movements of a random sample of 38 participants and stimulated recall data was collected from 20 participants to assist in interpretation of the eye-tracking data. Comparing reading behaviors of successful and unsuccessful test takers using eye-tracking measures, such as total fixation duration, fixation count, visit duration, and visit count, Bax identified five test items where there was a statistically significant difference between the two groups. Detailed analysis of the test takers' gazeplots and heatmaps for these items, in combination with stimulated recall data, revealed that unsuccessful test takers were not as efficient in expeditious reading (i.e., quick, selective reading to obtain desired information) as more proficient ones. Bax concluded that the use of eye-tracking succeeded in demonstrating clear differentiation between able and less able readers at three different levels of cognitive processing and allowed for precise identification of which elements of a text and test items may be most significant in this differentiation. More importantly, he concluded that the eye-tracking data showed evidence that IELTS items elicited the cognitive operations they were targeting. More recently, Brunfaut and McCray (2015), aiming to find "cognitive validity evidence" for the Aptis Reading, investigated the cognitive processing of 25 test takers of different proficiency levels while completing test. Using a combination of eye-tracking metrics provided by Tobii 300 and retrospective interviews with eye-tracking traces as stimuli, the researchers found that test takers engaged in a wide range of cognitive processes, from lower- (e.g. lexical access, syntactic parsing, propositional meaning building) to higher-level processing (e.g. inferencing, building a mental model, creating paragraph/text level representations). They also found that some gap-fill tasks may elicit different cognitive reading processes than those intended to be assessed. Based on their findings,

the researchers concluded that the test, as a whole, elicited a wide range of cognitive processes as intended and thus, sampled extensively from the construct of reading in terms of cognitive processing.

In listening assessment, eye-tracking technology has been employed to examine the effects of visual cues on listening comprehension, which has implication on the definition of a construct of listening comprehension ability. To date, the only study of this kind is one by Suvorov (2015) who investigated 33 test takers' viewing of context videos and content videos in a video-based academic listening test as well as the relationship between their viewing behavior and test performance. Comparing test takers' eye-tracking measures extracted from EyeTech Vision Tracker 2, such as fixation rate, dwell rate, and the total dwell time, for context and content videos and correlating each measure with test scores, Suvorov (2015) found statistically significant differences between fixation rates and between total dwell time values but insignificant difference between the dwell rates for context and content videos. Additionally, he could not identify any statistically significant relationship between the eye-tracking measures and the test scores. From the findings, Suvorov supports the expansion of the construct of L2 listening to also include ability to comprehend visual information.

Similar to reading assessment research, writing assessment research has also seen the use of eye-tracking technology for investigating the comparability of the actual writing processes test takers undertook and the processes to be measured by the test. For instance, Révész, Michel, and Lee (2017) employed this technology to find evidence supporting the warrant about the cognitive processes defined as part of the construct assessed by the online IELTS Academic Writing Test. Specifically, they examined the cognitive processes and behaviors of L2 writers while performing the test as well as the relationship between test takers' behaviors and the quality of

the text produced. Thirty Chinese participants completed an online version of the test task which required them to present their viewpoint on a controversial topic. Data collected were test scores, participants' eye-movements and logs of their keystrokes captured by Tobii TX60 mobile eye-tracking system and a keystroke logging software, and stimulated recalls from a subset of 12 participants. The researchers identified a wide range of cognitive processes and writing behaviors elicited by the task which were in accordance with the construct that the task was intended to measure. From these findings, they suggested that “the IELTS Academic Writing Task 2 has cognitive validity in the sense that the cognitive processes in which L2 writers engaged while completing the task reflected the processes which L1 writers typically employ” (p. 35).

It can be seen from these studies that eye-tracking data, used in combination with other types of data such as questionnaires or stimulated recalls, can provide information about test takers' cognition and behaviors during task completion. This, therefore, can help substantiate or reject the claim that the test tasks functioned as intended and assessed the targeted construct. For example, if a video-based listening test does not include visual comprehension ability as part of its construct, yet eye-tracking data reveal a long visual viewing duration among test takers and a significant association between this duration and their test scores, then the test scores may reflect construct irrelevant variance due to the influence on the visuals on test performance. This type of information will ultimately help test users arrive at a valid interpretation and uses of the test scores.

Raters' cognitive processes and behaviors

The second aspect in language testing that eye-tracking research has focused on is raters' attention to the rating rubric while rating the written responses. Such studies can provide information us about whether raters can consistently apply the rubric to arrive at reliable scores. For example, Winke and Lim (2015) examined raters' attention on different categories of an analytic rubric and identified whether different patterns in processing the rubric are related to unreliability in test scores. In their study, nine raters were asked to rate 40 essays using a five-category analytic rubric while their eye movements were recorded by a Tobii TX300 eye-tracker. Quantitative analysis of test scores and eye-tracking measures, such as time to first fixation, total fixation duration, and visit count, revealed an association between raters' attention and inter-rater reliability. Specifically, the more time a rater spent on one category, the higher the interrater coefficient for that category was. Also, detailed analysis of the gazeplots and heatmaps showed that raters who agreed attended to all rubric categories while those that disagreed had their attention focused on different locations of the rubric. The researchers concluded from the findings that an analytic rubric had a primacy effect, with the categories on the left (i.e., presented first) being more salient or considered more heavily than those on the right (i.e., presented last). In a subsequent study on ordering effects in rubric format on raters' cognition and behavior, Ballard (2017) collected eye-tracking data and recalls of the descriptors in each rubric category as well as the test scores from 31 raters. Using the same eye-tracking measures as Winke and Lim (2015), she found that categories at the outer-most positions (e.g., left-most and right-most) on the rubric are more likely to influence raters' attention to a category, their beliefs on what criteria are more important, and their ability to recall from a category. Based on the findings, the researcher suggested that rubric designers and test developers consider the ordering

effects for rater training and rubric design. Both studies above are concerned with how raters applied the rubric when evaluating test takers' performances. They have demonstrated the use of eye-tracking technology to investigate how raters attended to the rating rubric and its connection with interrater reliability and thus, provided implications in terms of rubric design and rater training.

Overall, though limited in number, the studies described above have demonstrated that eye-tracking technology, when coupled with another data collection method such as questionnaires, interviews, or stimulated recalls, can be employed to investigate various aspects in language testing. As shown above, it can be used to investigate topics such as test takers' cognitive processes and behaviors during task completion and raters' cognitive processes when rating test takers' performance. Therefore, eye-tracking has proved to be an innovative and useful methodology and could be extremely valuable for test validation research. For example, it could be applied investigate topics pertaining to the reliability of the evaluation process, i.e., whether raters consistently apply the rating criteria as they are supposed to. Additionally, it can be used to investigate test takers' cognitive processing and thus, provide evidence that the test tasks measure the construct as intended by the test developers.

To date, however, no eye-tracking studies have been conducted to investigate raters' cognitive processes to provide evidence for the test construct. Eye-tracking data, coupled with stimulated recalls, can also be employed to examine the rating criteria or aspects of language that raters attend to while rating test takers' performance. If raters consistently attend to features extraneous to the rating rubric, which reflect the operational definition of the construct being measured (Fulcher, 1996; Weigle, 2002), test developers might need to modify their rubric to better communicate their test construct to the raters. In contrast, if raters ignore certain rating

criteria, revision of the of the rating rubric or test construct redefinition is also needed. This type of research could help test developers deploy the test tasks or items in operational tests which more accurately reflect the test construct and help mitigate two threats to validity, namely construct-underrepresentation and construct-irrelevance in L2 assessment research.

An Argument-based Validity for Validation Research

The past studies investigating rating reliability on source-based writing tests does not show a clear connection between their findings on reliability issues and construct validity. The validity argument (Kane, 1992, 2006, 2013) is a beneficial framework to make this link because it shows the role of both reliability and construct validity in making inferences about the validity of the test score interpretation and use. This section provides an overview of argument-based validity for validation research, including its overall structure and advantages.

Overview

Developing an argument for interpretation and use of a test is the first of two processes in the argument-based validation approach conceptualized by Kane (2006). It refers to specifying the proposed interpretations and uses of test scores and offers a structure for a validity argument (Chapelle, Enright, & Jamieson, 2008; Kane, 2006). The Interpretive/Use Argument (IUA) provides a framework for validation by (1) clearly outlining the claims to be evaluated and (2) making the reasoning inherent in the proposed interpretations and uses explicit so that it can be better understood and evaluated (Kane, 2006, 2013). By outlining the claims to be evaluated, it provides a framework for validation and specifies the types of evidence needed to support test score interpretations and uses.

An IUA provides the structure for a validity argument. That is, if all inferences and assumptions entailed in the IUA are supported to the extent possible and all challenges that pose potential threats to these inferences and assumptions are weakened or disproved with reasoning or evidence, then the IUA can be a plausible argument for the proposed score interpretation and use (Kane, 2013), or the validity argument. In other words, the validity argument is the IUA with supporting evidence for the assumptions (Chapelle, 2012).

Structure of an Argument-based Validity Inference

An argument-based validity states the inferences underlying the interpretation and use of test scores. Structure of an inference in argument-based validity is illustrated in Figure 2.1.

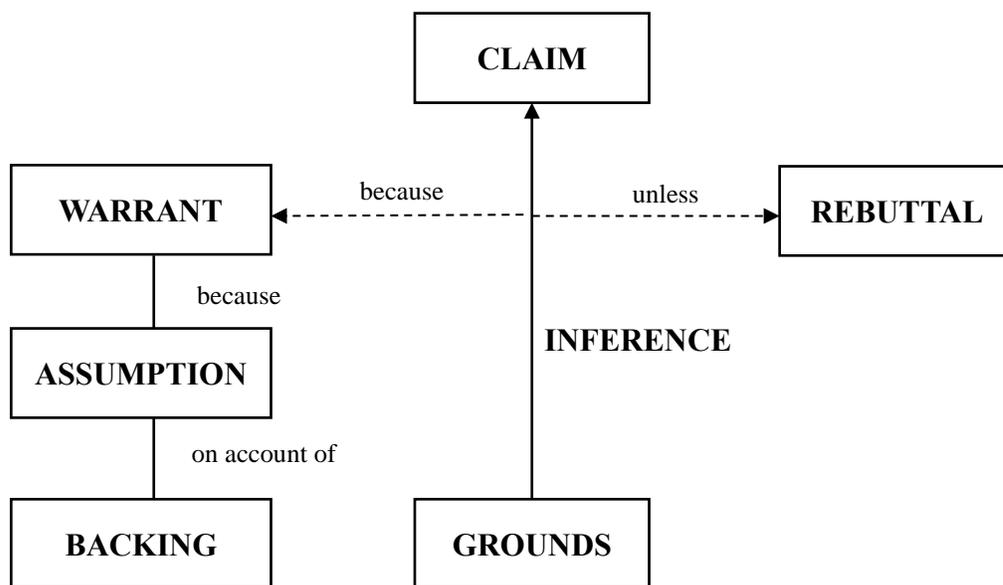


Figure 2.1. Structure of an inference in argument-based validity

Following Toulmin's (2003) argument model, each inference leads to a claim. Claims refer to conclusions about test score interpretations and uses based on various observations and data, called grounds. The linking process from grounds to a claim is labeled as a specific

inference depending on the nature of the claim and is authorized by one or more warrants. Each warrant has underlying assumptions, or statements specifying which type of backing is needed. Rebuttals are statements indicating conditions where in inference cannot be authorized in a particular context. The claim in an inference serves as the grounds for the next inference, making the reasoning inherent in the proposed interpretations and uses explicit so that it can be better understood and evaluated (Kane, 2006, 2013). By outlining the claims to be evaluated, argument-based validity provides a framework for validation and specifies the types of evidence needed to be collected in each inference to support test score interpretations and uses.

Strengths of Argument-based Validity for Test Validation

Argument-based validity has many advantages. First, it helps identify priorities in validation research and “provides guidance as to the types of research needed” (Chapelle, 2008), producing the type of backing required to support the inferences. As Kane (2006) argued, “[t]he main advantage of the argument-based approach to validation is the guidance it provides in allocating research effort in gauging progress in the validation effort” (p. 23). This approach defines essential validation research by systematically examining the inferences in the interpretive argument, offering explicit “guidance and conceptual infrastructure” for developing validity arguments (i.e. inferences, assumptions, and claims) which helps researchers to see where research is needed (Chapelle, Enright, & Jamieson, 2010). By adopting this method, researchers can focus their efforts on collecting the most relevant and problematic validity evidence to support the inferences and assumptions in the IUA. In addition, at the end of the project, this method also allowed me to identify possible gaps in my validity argument, leading to further research to seek additional backing evidence to strengthen my validity argument in the future.

Second, argument-based validity creates the explicit logic that links elements in a validity argument such as inferences, warrants, and their assumptions, making explicit the inferences about test takers' ability that score users make when they use the test scores (Chapelle, 2021). It allows the researcher to present the proposed interpretations and uses underlying test scores, define essential validation research, and structure the connection between validation research results and the claims they wish to make about the test in a coherent, systematic way to their readers. Going through the bridges of different inferences outlined in the validity argument, readers can easily learn about the claims for each inference, evaluate the appropriateness of assumptions underlying the inference, and evaluate whether the evidence provided to support the assumptions are plausible. In this way, researchers' arguments will be more easily followed and more convincing to the audience.

In addition to the two general advantages, argument-based validity also has other strengths more specific to the validation of the EPT Writing test. It distinguishes reliability-related issues in the *evaluation* inference from those in the *generalization* inference. Specifically, the *evaluation* inference is made when evidence related to reliability at the test task level (i.e., score reliability and accuracy of ratings provided by each rater to each task response) is gathered. Meanwhile, the *generalization* inference is drawn when evidence related to reliability beyond the test tasks (i.e., score consistency across tasks and raters) is collected. In other words, the *evaluation* inference arrives at a conclusion about the quality of the performance sample at the task level whereas the *generalization* inference is concerned with the total score consistency (Chapelle, 2021). This distinction provided ample opportunities for me as a researcher to state assumptions about the reliability and generalizability of the EPT Writing test scores so that all issues related to reliability in source-based academic writing were thoroughly investigated.

The second advantage of argument-based validity specific to the current study is the argument's ability to connect reliability with test construct validity. The fact that the conclusion about the test score reliability made in the *generalization* inference serves as the grounds for the *explanation* inference pertaining to the test construct allows for a logical connection between reliability and construct validity. In this way, test users can easily make inferences about the validity of the test score interpretation and use.

Overall, by laying out in detail the interpretations and uses intended for the test scores, the IUA helped me identify the types of evidence needed to support the assumptions underlying the warrants in the three inferences that I focused on in this project, namely the *evaluation*, *generalization*, and *explanation*. They helped me visualize the steps I would need to take to collect the relevant evidence to validate the interpretations and uses of the EPT Writing test scores. Also as important, this approach was useful for presenting my arguments, assumptions, and validity evidence in a logical, coherent way to allow others to evaluate for themselves the clarity and coherence of my argument and the plausibility of in inferences and assumptions for the test scores.

Research Goals and Research Questions

My study aimed to examine the construct of the EPT source-based academic writing test by studying raters' decision-making processes using eye-tracking technology and stimulated recall. As the construct can only be as meaningful as the rating process is reliable, I also investigated critical aspects of reliability, including the appropriateness of the rating scale, the performance of the raters in terms of severity, consistency and bias, and test score reliability. The findings were presented in a validity argument to show the connection between the reliability of

the ratings and their construct validity that needs to be taken into account in research on rating processes. This section starts with an introduction of the EPT Writing construct. It then presents an IUA for the EPT Writing with a focus on the three inferences related to the reliability of the ratings and the construct validity, namely, the *evaluation*, *generalization*, and *explanation*. It concludes with a list of research questions that this study aimed to address.

The EPT Writing Construct

The EPT Writing is one of the two components of an English placement test designed for first year nonnative English-speaking students arriving at a large Midwest university to start their study program. The purpose of the EPT Writing is to determine if these students need additional writing classes to meet the English language requirement for academic success at the university. Based on their test scores, students can be placed into one of the following ESL writing courses:

1. 101B, a lower level writing course with an emphasis on grammar and paragraph level writing;
2. 101C, a more advanced writing course for undergraduate students with an emphasis on writing beyond the paragraph level; or
3. 101D, a more advanced writing course for graduate students with a focus on research writing.

Those who pass the test are allowed to take first-year composition courses if they are undergraduate students or meet the English language requirement if they are graduate students. This section explains the theoretical perspective adopted for the test construct definition and provides a description of the test construct.

Theoretical perspective adopted for the definition of the EPT Writing construct

I developed the EPT Writing using the interaction-focused approach (Bachman, 2007) or interactionalist perspective (Chapelle, 1998) to test construct definition. Interactionalists see constructs as both language ability and contextual features, along with their interaction (Chapelle, 1998). This perspective is mainly based on literature in psycholinguistics and sociolinguistics along with research in SLA, social interaction, and discourse analysis. According to some proponents of this perspective who draw largely on sociolinguistics and discourse analysis, *interactional competence* is not limited to individual test takers but an interactive process by which cultural meanings are created (He & Young, 1998). Therefore, for those researchers, the construct assessed by a test is not “an attribute of either the individual language users or of the context, but as jointly co-constructed and residing in the interactions that constitute language use” (Bachman, 2007, p. 42) or in other words, *interaction* is the language construct. From a slightly different angle, Chalhoub-Deville (2003), drawing on the literature in learning and cognition, proposed *ability-in-language user-in context* as the language construct. As Bachman (2007) commented, this perspective distinguishes ability from context, but claims that interaction is a factor resulting in changes in ability. Meanwhile, other interactionalists (Chapelle, 1998) posit that the ability interacts with the context, with metacognitive strategies responsible for mediating between ability and context. To sum up, although the proponents of the interactionalist perspective might draw on different types of theories and practices, they all agree that the interaction between ability and context is an important component in defining language constructs.

Compared to the other approaches to construct definition, namely, the trait theorist approach and the behaviorist approach, the interactionalist perspective seems to be the most

logical and thorough. The trait theorist perspective, defining the construct as test takers' characteristics, can potentially provide information about the nature of the ability itself (Upshur, 1979, quoted in Bachman, 2007). However, this perspective ignores the effects of non-linguistic variables on performance while previous research has shown that test takers' ability alone does not explained test score variability. On the other end of the continuum, the behaviorist perspective, while accounting for the role of contextual factors, limits our interpretations of test scores to predictions about future performance on that same task, and subsequently, results in the problem of lack of generalization from one observed instance of behavior to other unobserved instances (McNamara, 1996). In the middle of these two extremes is the interactionist perspective. Considering performance as the results of traits, contextual features, and their interaction, it views performance as "a sign of underlying traits, and is influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts" (Chapelle, 1998, p. 43). This perspective takes into account the role of contextual factors while allowing for the generalization of test scores beyond the immediate testing instance.

Description of the EPT Writing construct

The test aims to assess test takers' source-based academic writing ability. The test construct includes three groups of factors, namely the learner factors, contextual factors, and task characteristics. The construct assessed by the test is illustrated in Figure 2.2.

As seen in Figure 2.2, the construct of source-based academic writing ability that the EPT Writing is intended to measure consists of various factors related to the context, the test task, and the learner. Regarding the contextual factors, the terms discussed in Chapelle (1998), i.e., field, tenor, and mode, were also used for this construct definition. For field, it is obvious that the

topics are academic. In the tenor component, the participants in this context are the academic writer, who needs to explain the main points discussed in the stimuli and then state and support their viewpoints, and their educated audience. The mode component is identified as computer-mediated written language for summaries and argumentative essays. All of these contextual factors influence the language test takers produce in their responses.

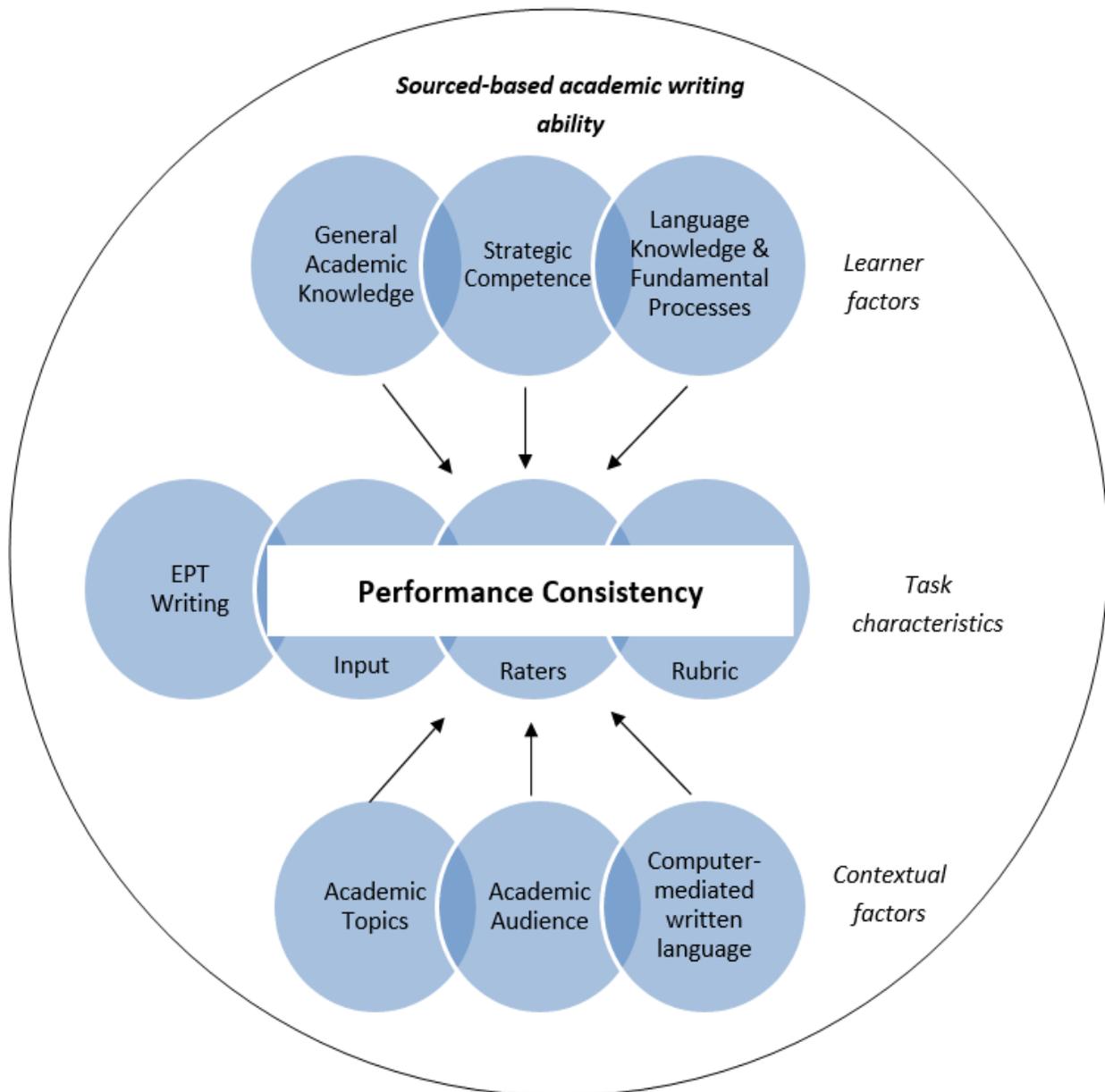


Figure 2.2. The EPT Writing construct of source-based academic writing ability

From the interactionalist perspective, performance consistency, attributed to both test takers' characteristics and contextual factors, is observed within a setting described as task characteristics (Chapelle, 1998). Therefore, factors of task characteristics also affect performance. For example, studies have shown that the characteristics of the input can result in test takers' performance (Cho, Rijmen, & Novak, 2013; Plakans & Gebril, 2013). Additionally, raters and rating scales have been found to influence score variability (Barkaoui, 2010; Knoch, 2009; Shin & Ewert, 2015). These factors must be considered in test score interpretations and use, and thus, are included in the test construct.

More importantly, the construct of the EPT Writing test must include the learner factors such as general academic knowledge, strategic competence, and language knowledge and fundamental processes. Because this is an academic writing test, test takers' knowledge on general academic topics is expected to influence their performance. Additionally, an interactionalist construct definition must include strategic competence, defined as "a set of metacognitive components, or strategies, which can be thought of as higher order executive processes that provide cognitive activities" (Bachman & Palmer, 1996, p. 40). These strategies allow test takers to assess the situation, decide how to respond to the writing topic (i.e., goal setting), and decide the types of language knowledge and background knowledge to use to achieve that goal (i.e., planning). In her study on the strategies used by writers in a reading-to-write test, Plakans (2009) found that the writers employed strategies such as organizing (such as arranging essay content, identifying rhetorical structures, and summarizing source text), selecting ideas from the readings, connecting ideas from the reading to their own in the essay, and monitoring their own writing and style issues. Other studies have also shown that writers in source-based writing test engage in construct-relevant strategies during task completion

(Barkaoui, 2015; Yang & Plakans, 2012). Strategic competence, therefore, is considered a component inextricable from the construct measured by the test.

The last yet most important learner factor that must be included in the test construct is test takers' language knowledge and fundamental processes that the test aims to measure. The types of these knowledge and processes are summarized in Figure 2.3.

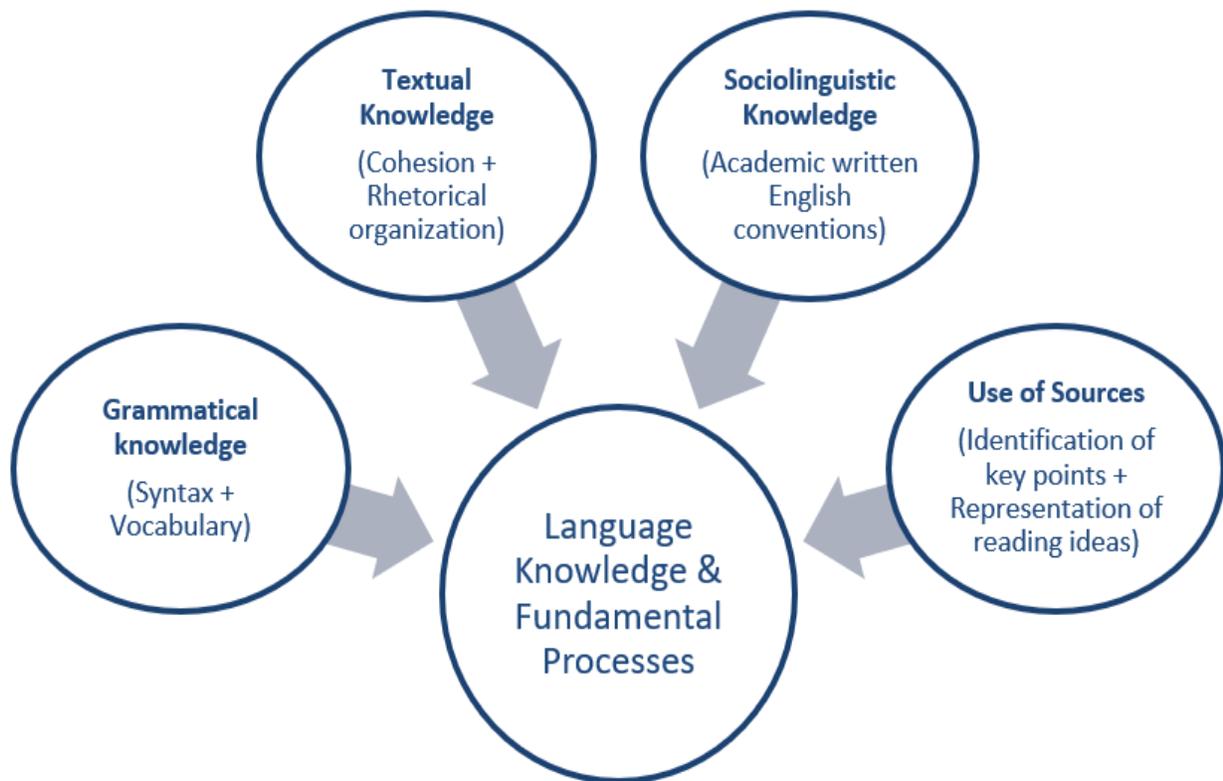


Figure 2.3. Components of the language knowledge and fundamental processes measured by the EPT Writing

The language knowledge that the test attempts to measure is specified in previous applied linguistics research on components of language ability. Specifically, the first three components of the language knowledge measured in this test, namely grammatical knowledge, textual knowledge, and sociolinguistic knowledge, are based on the framework of language knowledge

proposed by Bachman and Palmer (1996). However, since the context dimension must be included in an interactionist construct definition, these components of the language knowledge are more peculiar to source-based academic writing. First, grammatical knowledge involves knowledge of syntax and vocabulary to produce formally accurate sentences. Second, textual knowledge refers to ability to produce explicitly marked relationships among written sentences (knowledge of cohesion) and to produce organizational development in written texts (knowledge of rhetorical organization). In the context of this test for university level students whose writing is almost based on some prior reading, test takers are expected to provide appropriate and support for ideas using information in the reading stimuli (Weigle, 2002). Third, sociolinguistic knowledge allows learners to create language appropriate to a particular language use setting. In this case, test takers' responses to the reading stimuli on general academic topics are intended for the academic audience, which means that they have to follow the conventions of written academic English in their response. In addition, because this is a source-based, reading-to-write test, an additional dimension of the construct should include ability to identify and present appropriate source text material and use this material as a means to develop a larger argument (Bereiter & Scardamalia, 1987, quoted in Weigle, 2002).

In general, these factors from the learner and context, together with task characteristics, interact with one another and influence performance consistency and are therefore explicitly delineated in the construct definition of the EPT Writing. Overall, the EPT Writing aims to assess test takers' ability to summarize and synthesize information presented from different sources, state and support their arguments with sufficient details and examples in a well-structured, coherent way with a good command of English vocabulary, grammar and writing conventions.

An Interpretive/Use Argument for a Source-based Academic English Writing Test for Placement Purposes

In order to build a validity argument for presenting the findings from the study, an IUA, presented in Figure 2.4, was first created to direct research aiming to support the interpretation and use of the EPT Writing test scores. A complete IUA framework is useful as it reminds the researcher to use a conclusion in one inference as the grounds for the next inference. This IUA includes seven inferences, namely *domain definition*, *evaluation*, *generalization*, *explanation*, *extrapolation*, *utilization*, and *consequence*. Each inference is presented by an arrow connecting the grounds, which is also the claim for the previous inference, to its claim, which then serves as the grounds for the next inference. Each inference is authorized by one or more warrants and the associated assumptions, which need to be supported by backing.

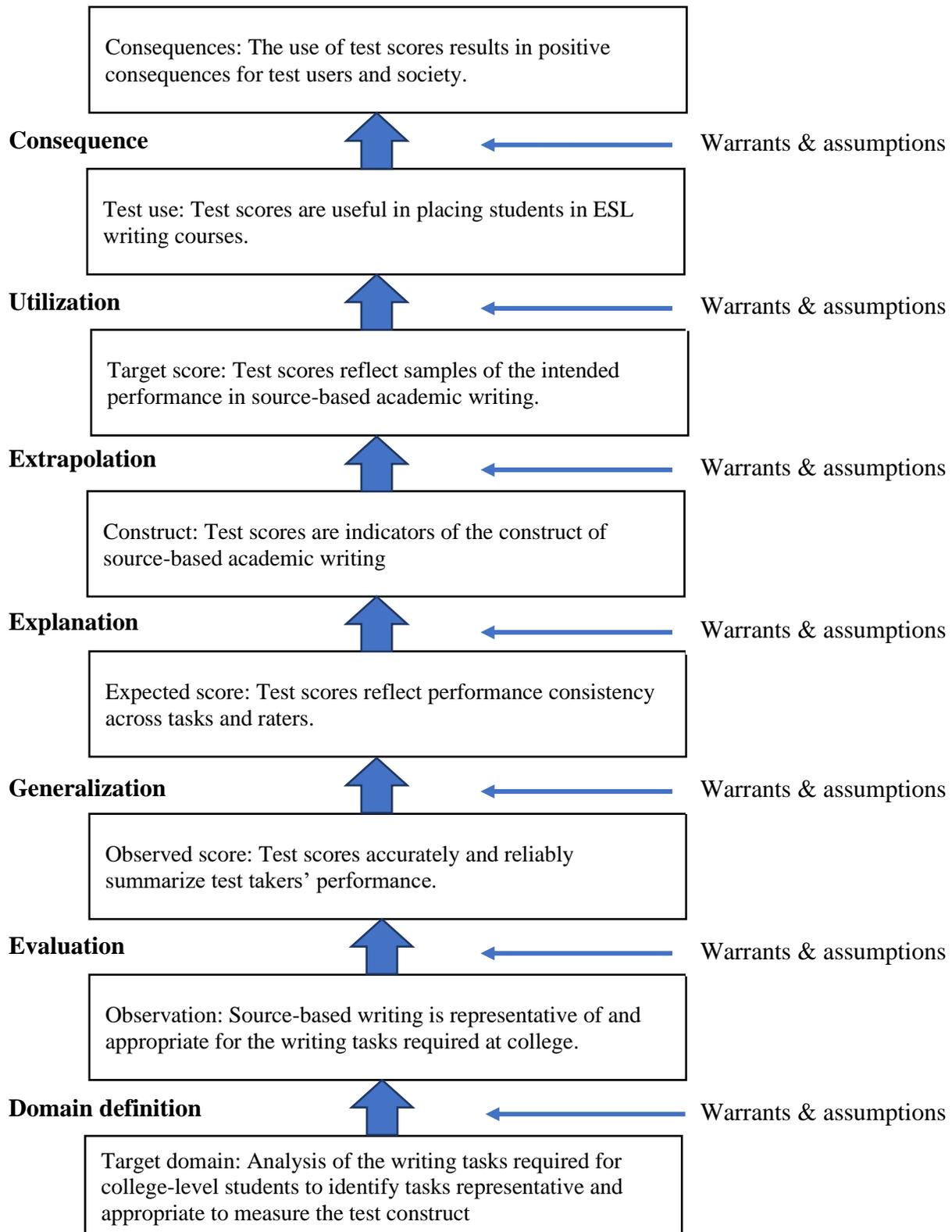


Figure 2.4. Structure of the IUA of the EPT Writing

The Interpretive/Use Argument for the EPT Writing

The *domain definition* inference in the IUA for the EPT Writing is made when a claim is stated about the quality of the test development process of obtaining test takers' written responses appropriate for making inferences about test takers' source-based academic writing ability and placing them into appropriate ESL writing classes based on their test scores. This inference is authorized by warrants about the relevance and representation of the source-based writing tasks (i.e., summary and argumentative essay) in academic writing at the university. Backing for the warrants' associated assumptions comes from findings of an analysis of the writing tasks required for college-level students to identify tasks that are representative and appropriate to measure the test construct. The claim for the domain definition inference, once its warrants and their associated assumptions are supported with backing, is that the test tasks are representative of and appropriate for the writing tasks required at college. This claim serves as the grounds for the *evaluation* inference, which links observation of test takers' performance to test scores. This inference is built on the warrants about test administration conditions and scoring quality including rater performance, and rating scale quality. If the warrants and their associated assumptions are backed by sufficient evidence, the claim resulting from this inference is that test scores accurately and reliably summarize test takers' performance,

The next inference, *generalization*, which links the observed score to the expected score (i.e., the estimated score expected of test takers across parallel tasks and raters), uses the claim in the *evaluation* inference (that test scores accurately and reliably summarize test takers' performance) as the grounds for its claim. The *generalization* inference is permitted if the warrants about performance consistency across raters and test tasks and their underlying

assumptions are supported by evidence. In that case, the claim for this inference that test scores reflect performance consistency across tasks and raters can be made.

The *explanation* inference links the claim about the expected score and the one about test construct. The claim in the *generalization* inference regarding test score consistency serves as the grounds for the claim that test scores are indicators of the test construct (i.e., source-based writing). The *extrapolation* inference, which takes the claim in the *explanation* as its grounds and which links the test construct to the target score (i.e., “the examinee’s expected score overall possible performances in the target domain”, Kane et al., 1999, p. 7), is based on the support of warrants about performance in the target domain of academic writing. When the warrants and their underlying assumptions are supported, it can be concluded that the test scores reflect samples of the intended performance in academic writing.

The claim that test scores reflect samples of the intended performance in the target domain serves as the grounds for the next inference, *utilization*, which connects target scores with the use of the test scores. This inference is supported by warrants about utility of the intended test scores to place students into ESL writing courses and the appropriateness of the cut scores. If the warrants and their underlying assumptions are bolstered by backing, the claim that the test scores are useful in placing students in ESL writing courses is justified. This claim then becomes the grounds for the next inference, *consequence*, which links test use to consequences. This inference is licensed by warrants about the test benefits to various users such as students and instructors at the university and to the academic community of practice. Having sufficient evidence to support the warrants and their associated assumptions allows for the claim that the use of test scores results in positive consequences for test users and the society.

The three inferences for reliability and construct validity

This study investigated issues related to reliability and test construct of the EPT Writing. The main reliability issues are captured by two types of claims, one about the absence of errors in the test taking and scoring processes, and the other about the absence of errors in the test scores (Chapelle, 2021). In a validity argument, these claims about reliability appear as the conclusions for two inferences: *evaluation* and *generalization*, respectively. More importantly, it also aimed to examine raters' cognition to investigate support for one of the inferences about the test construct. Claims about the test construct in a validity argument are conclusions from the *explanation* and *extrapolation* inferences. Thus, this section focuses on the inferences related to the rating reliability, (i.e., the *evaluation* and *generalization* inference), and the test construct validity, (i.e., the *explanation* inference). Findings regarding the rating scale appropriateness and raters' performance were evidence backing for assumptions in the *evaluation* inference. Findings pertaining to test score reliability were used as evidence backing for assumption in the *generalization* inference. Information about raters' decision-making processes served as evidence backing for assumptions in the *explanation* inference.

The *evaluation* inference leads to a claim pertaining to the reliability of the ratings: that test scores accurately and reliably summarize test takers' performance. This inference is authorized by two warrants: (1) the rating rubric is appropriate for providing evidence of variation in source-based academic writing ability and (2) raters' performance is reliable. The first warrant related to rubric appropriateness is based on the assumptions that experts believe that the rubric is appropriate and that there is statistical evidence supporting the usefulness of the rubric. These assumptions are supported by raters' interviews and Many-Facets Rasch Measurement (MFRM). The second warrant related to raters' performance is supported by the

assumptions that the raters are comparable and consistent in their ratings and that the raters do not exhibit significant bias against one or the other task type. These assumptions are backed by MFRM analysis showing that raters are consistent in their rating and that the scores they provide are analogous to those given by other raters as well as the results from bias analysis. Figure 2.5 illustrates the structure of the warrants and assumptions with the evaluation inference in the IUA for the EPT Writing.

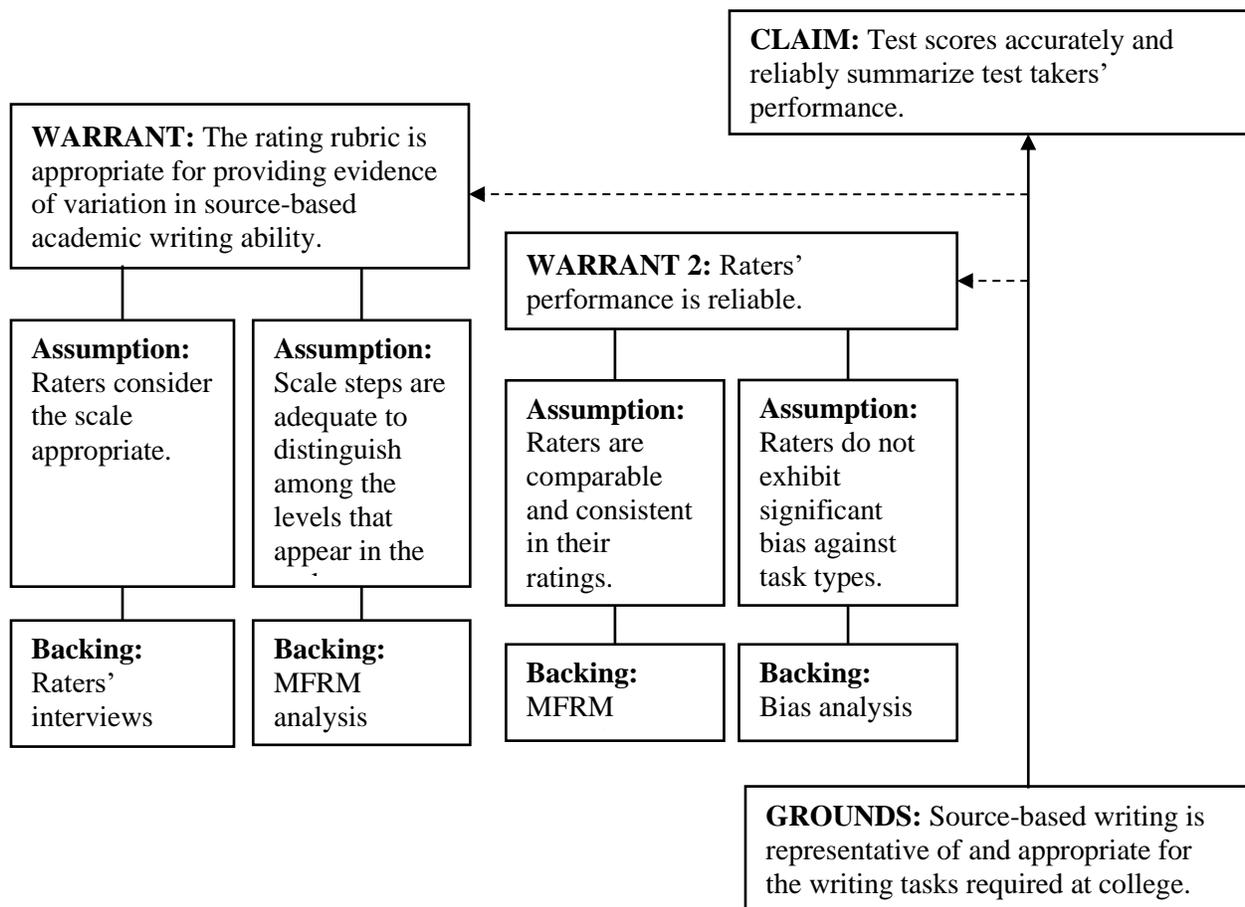


Figure 2.5. Structure of the evaluation inference in the IUA for the EPT Writing

The next inference, *generalization*, is made when the test takers' scores on EPT Writing test tasks are accepted as generalizable to performance on all writing tasks in the domain of

college writing. Using the conclusion of the *evaluation* inference that the ratings accurately and reliably summarize test takers' performance, the *generalization* inference leads to the claim that test scores reflect the desired level of performance consistency across tasks and raters. Thus, it is the link between the score assigned to student performance on the test and their expected score in similar test forms under similar testing conditions (i.e., the universe of relevant observations). The warrant supporting this inference is that the score reliability is adequate for ESL writing course placement. This warrant rests on the assumption that the number of raters and tasks is adequate to result in consistent scores. To support the assumption, generalizability studies (G-studies) and decision studies (D-studies) should be conducted to examine the impact of raters on individual test takers' performance, which subsequently helps decide on the number of raters needed for the test. Figure 2.6 illustrates the structure of the generalization inference in the IUA for the EPT Writing.

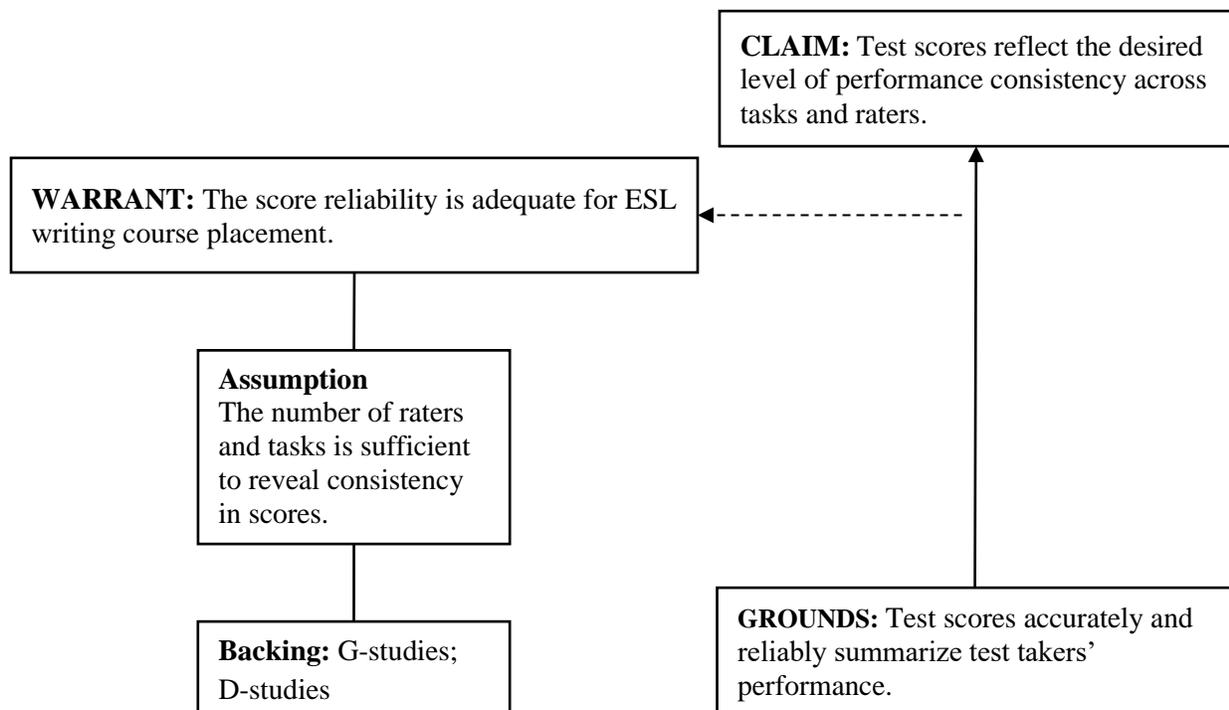


Figure 2.6. Structure of the generalization inference in the IUA for the EPT Writing

The conclusion from the *generalization* inference regarding rating reliability serves as the grounds for the subsequent *explanation* inference about the test construct. The *explanation* leads to the claim that the ratings are indicators of the theoretical construct the test aims to measure, which is test takers' ability to summarize and synthesize information presented from different sources, state and support their arguments with sufficient details and examples in a well-structured, coherent way with a good command of English vocabulary, grammar and writing conventions. The *explanation* inference is authorized by the warrant that raters' attention is aligned with the construct of source-based academic writing ability. This warrant is supported by the assumption that the writing features that raters attend to are appropriate in view of the construct of source-based academic writing ability defined for the test. Backing for this assumption is collected from qualitative analysis of raters' verbal reports. Figure 2.7 illustrates the structure of the explanation inference in the IUA for the EPT Writing.

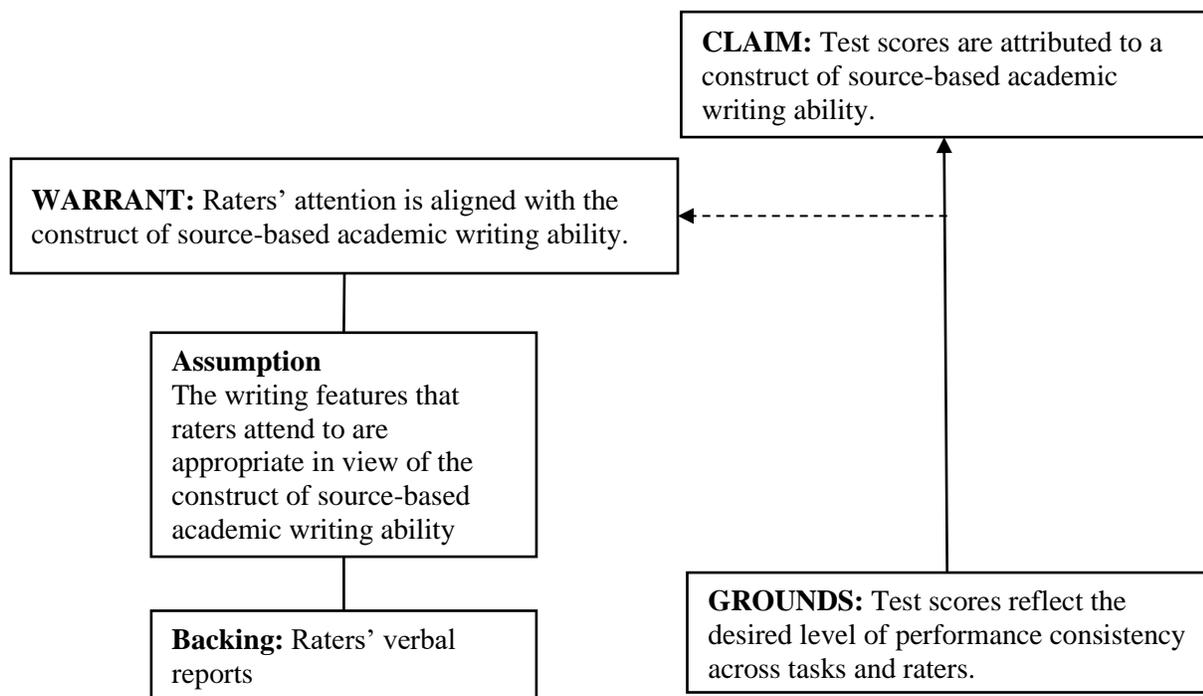


Figure 2.7. Structure of the explanation inference in the IUA for the EPT Writing

As can be seen from the three inferences, the conclusion for the *evaluation* inference becomes the grounds for the *generalization* inference. Similarly, the conclusion for the *generalization* inference serves as the grounds for the *explanation* inference. This argument illustrates the need to have support for claims about reliability in order to make an inference about the construct that the test measures. The research questions were framed within these three inferences and will be presented in detail in the next section.

Research Questions

My dissertation investigated the appropriateness of the rating rubric, raters' performance, rating reliability, and raters' decision-making processes during their rating of source-based written responses. Framed under the validity argument, the study examined the extent to which assumptions underlying *evaluation*, *generalization*, and *evaluation* inferences in the IUA for the EPT Writing test were supported by the rating processes used to assign scores to test takers' responses. Table 2.1 presents the research questions formulated to address each assumption presented in the three inferences of the IUA for the EPT Writing introduced in the previous section.

Table 2.1. *Warrants, Assumptions and Research Questions Associated with the Evaluation, Generalization, and Explanation Inference*

Warrant	Assumption	Research Questions
Evaluation inference		
Claim: Test scores accurately and reliably summarize test takers' performance.		
The rubric is appropriate for providing evidence of variation in test takers' source-based academic writing ability.	Raters consider the scale appropriate.	1. What did experts think about the appropriateness of the rating scale for evaluating test takers' performance?
	Scale steps are adequate to distinguish among the levels that appear in the scale.	2. Was there statistical evidence supporting the number of scale steps?
Raters' performance is reliable.	Raters are comparable and consistent in their ratings.	3. Were raters comparable and consistent in their ratings?
	Raters do not exhibit significant bias against certain task types.	4. To what extent did raters exhibit bias against certain test tasks?
Generalization inference		
Claim: Test scores reflect performance consistency across tasks and raters.		
The score consistency is adequate for ESL writing course placement.	The number of raters and tasks is sufficient to reveal consistency in scores.	5. What was the score reliability with two tasks are included, and each response double-rated?
Explanation inference		
Claim: Test scores are attributed to a construct of source-based academic writing ability.		
Raters' attention is aligned with the construct of source-based academic writing ability.	The writing features that raters attend to are appropriate in view of the construct of source-based academic writing ability defined for the test.	6. What writing features did raters attend to when rating source-based writing task responses?

Chapter Summary

This chapter reviewed research on four main areas relevant to the study: research on source-based writing assessment constructs, reliability research in source-based writing assessment, eye-tracking technology, and argument-based validity for test validation research, as well as presented the research goals and questions. The review of the previous literature has reached the following conclusions: First, an investigation of raters' cognitive processes could shed lights on source-based writing test constructs. Second, a combination of think-aloud protocols and eye-tracking technology could give insightful, reliable information about rater's cognitive processes. Third, a validity argument is useful in framing studies on rating reliability and test constructs in source-based writing assessment. The chapter concluded with an introduction of the EPT Writing construct, the IUA for the EPT Writing, and the research goals and questions that guided the study. The next chapter will detail the methodology adopted to address the research questions.

CHAPTER 3: METHODOLOGY

This chapter delineates the research methodology of the study. It begins with a description of the overall research design adopted in the study. Next, detailed information about the EPT Writing tasks and the nine raters participating in the study is provided, followed by a description of materials and instruments used for data collection: the Essay task (Task 2) responses, the EPT Writing rubric, the web-based training materials, the eye-tracking equipment and software, and the interview protocol. In addition, the chapter explains the procedures used to collect four types of data: the operational EPT ratings of the Summary task (Task 1) and Essay task (Task 2) responses, the experimental ratings of the Essay task (Task 2) responses, the stimulated recall data from the raters, and the interview data from the raters. The last section provides information about the data preparation and analyses carried out to address the research questions (RQs) in the study.

Research Design

This study employed the mixed-methods multiphase research design (Creswell & Plano Clark, 2012) in which both quantitative and qualitative data were collected and analyzed in two sequential phases to address the research questions. Quantitative data consisted of the operational ratings of 283 summaries (Task 1 responses) and 283 essays (Task 2 responses) given by 10 trained EPT raters from the official Fall 2018 EPT administration and the experimental ratings given by nine participant raters who rated 60 essays (Task 2 responses) randomly selected from the EPT pool of essays on the same topic. Qualitative data included raters' responses from interviews and stimulated recalls collected from raters during their experimental rating of the task responses. The data were analyzed qualitatively and quantitatively to arrive at an overall understanding of reliability and construct validity of the EPT Writing.

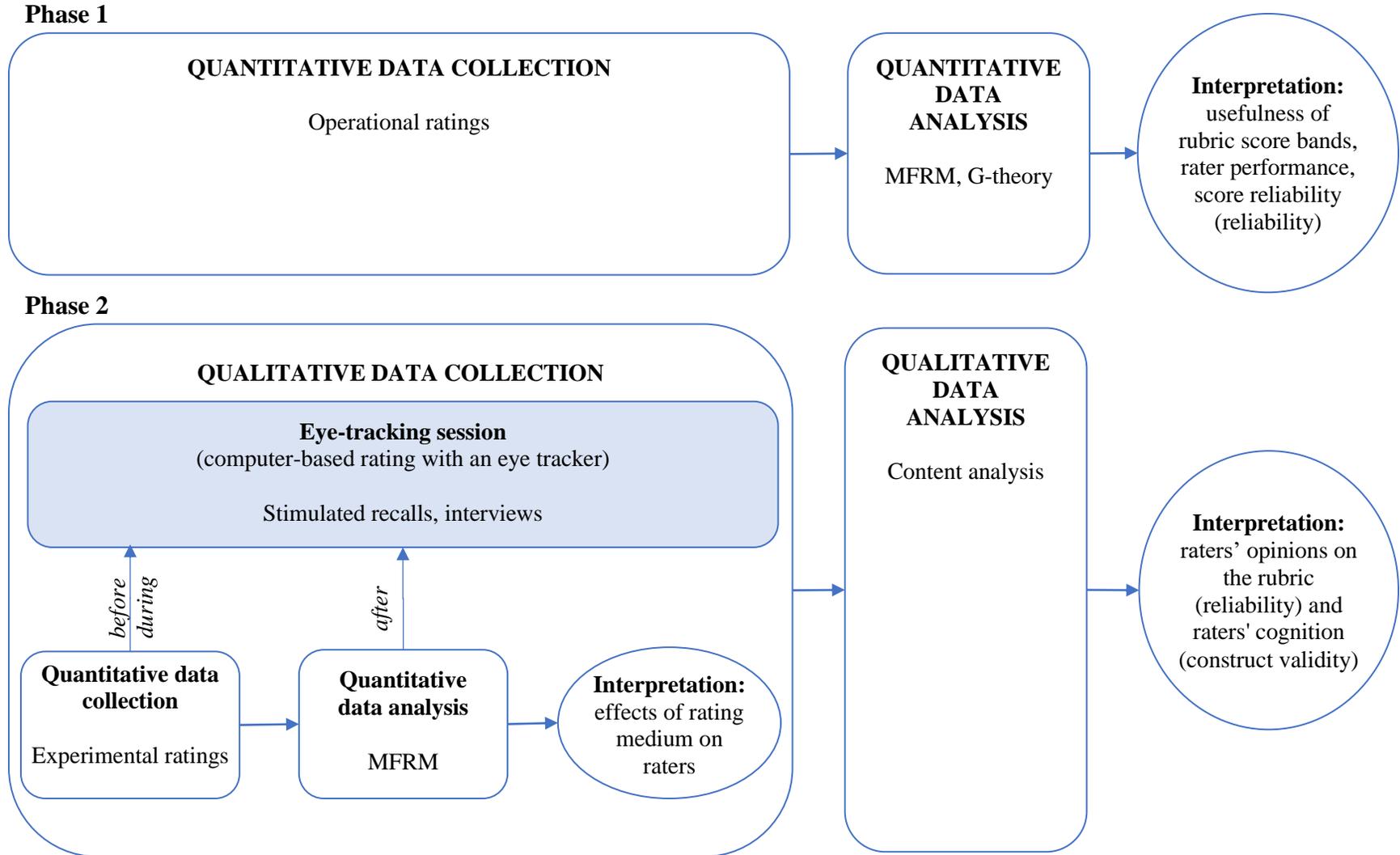


Figure 3.1. Mixed-methods multiphase research design

Figure 3.1 illustrates the design of the study. The study consisted of two phases. In Phase 1, operational ratings were collected from the EPT Office and analyzed using MFRM and G-theory to make an interpretation about the reliability of the operational test scores based on the usefulness of the rubric score bands (RQ2), rater performance (RQs 3 – 4), and score reliability (RQ5). The Phase 1 analysis used existing deidentified ratings, and therefore, the operational raters are not considered participants in the study.

Phase 2 followed the mixed methods embedded designed where quantitative data were collected and analyzed within a larger, overarching qualitative research design (Creswell & Plano Clark, 2012). Specifically, the quantitative strand, where experimental ratings were gathered, was embedded within the larger qualitative design in which stimulated recalls and interviews were collected to address the research questions on raters' opinions about the rubric (RQ1) and their decision-making processes (RQ6). The collection of the experimental ratings occurred before and during the eye-tracking session. During this session, the raters ratered 10 essays from the pool of 60 essays used to obtain the experimental ratings and performed recall protocols after each essay. The experimental ratings were analyzed with MFRM after the eye-tracking session to make an interpretation of the effect of rating medium (i.e., paper-based versus computer-based) on the raters, and thus provided a supportive, secondary role for the qualitative strand in which stimulated recalls and interviews were collected and analyzed. Content analysis was conducted on the interviews and stimulated recalls to arrive at an interpretation about the reliability of the operational test scores based on raters' opinion about the appropriateness of the rubric for distinguishing test takers' source-based academic writing ability and about the EPT Writing construct validity based on raters' decision making processes.

The EPT Writing Tasks

The EPT Writing is a computer-based English test aiming to assess source-based academic writing ability, including ability to summarize, synthesize information, and present opinions with supporting details from external sources. Test takers are presented with two reading texts of approximately 300 words each expressing a contrasting perspective from the other on a controversial general, academic topic. Test takers are required to summarize the two readings including a comparison and contrast of the perspectives (Task 1) in 100 – 120 words within 15 minutes (referred to as “Summary” hereinafter). After that, they are asked to give their viewpoints on the same general academic topic with supporting details from the texts and their own experience (Task 2) in 300 – 350 words within 30 minutes (referred to as “Essay” hereafter). The test lasts for 50 minutes, including test takers’ 5-minute reading of the sources before writing. Test takers can access the readings and task instructions during task completion.

Participants

Nine raters, who were graduate students in TESOL/applied linguistics, were recruited to serve as raters to provide experimental ratings of the responses while being monitored using eye-tracking. These raters, aged 26 – 40, came from different L1 backgrounds and were highly proficient in English. Raters 1 – 4 had experience teaching first-year composition while raters 5 – 9 had taught the ESL writing courses at the institution. Five raters, Raters 5 – 9, had rated for the EPT Writing for at least one semester. Their EPT Writing ratings were included as part of the operational ratings (described in greater detail later) for this study. However, Raters 1 – 4 never worked as official rater for the EPT Writing. Table 3.1 presents background information about the raters.

Table 3.1. *Raters' Background Information (N = 9)*

Rater	Sex	Rating experience (in semesters)	Teaching experience
1	Female	0	First-year composition
2	Male	0	First-year composition
3	Female	0	First-year composition
4	Male	0	First-year composition
5	Male	1	Undergraduate ESL writing
6	Female	3	Undergraduate ESL writing
7	Male	5	Undergraduate ESL writing)
8	Male	6	Undergraduate/Graduate ESL writing
9	Male	8	Undergraduate/Graduate ESL writing

Materials and Instruments

The materials and instruments used to collect the data include (1) the Essay task (Task 2) responses, (2) the EPT official rating rubric, (3) the web-based rater training materials, (4) eye-tracking equipment and software, and (5) the interview protocol. Details about the materials are provided below.

Essay Task (Task 2) Responses

In order to collect the experimental ratings before and during the eye-tracking session, a total of 60 essays on the topic of genetically modified (GM) food written by EPT test takers in Spring 2018 were randomly selected from a pool of 115 EPT essays. As mentioned in the earlier section about the context of the EPT Writing, the test takers were asked to read two texts of

different views on GM food and respond to the question of whether GM food should be encouraged (Appendix A) in 300 – 350 words within 30 minutes. Among the 60 essays, 10 essays, used for eye-tracking and stimulated recalls, were selected based on their infit values produced by MFRM in *Facets*. Specifically, essays with an infit value larger than 1.5 logit were considered problematic (Huhta et al., 2014; Linacre, 2014) and chosen for eye-tracking and stimulated recalls. The selection of the misfitting essays was motivated by the expectation that because they were hard to classify, they have the greatest potential for eliciting insightful data from raters illustrating the complexity of task of rating. Identifying characteristics of examinees were removed from the 60 essays before the nine participating raters started rating.

Rating Rubric

The rubric used for the study was a holistic scale consisting of five levels: B (lowest level), B+, C (for undergraduate students) or D (for graduate students), C/D+, and Pass (highest level). It should be noted that although the scale is holistic, raters consider four categories, namely *Organization*, *Arguments and Details*, *Grammar and Lexis*, and *Conventions* when rating. The weight of each category on the total score is as follows: *Organization* 25%, *Arguments and Details* 30%, *Grammar and Lexis* 30%, and *Conventions* 15%. A level B test taker can:

- somewhat organize their response despite of a lack of focus in paragraphs and rare or inappropriate use of transitional devices;
- support their arguments although more relevant examples and more elaboration are needed;

- control simple grammatical structures despite a lack of range in vocabulary and errors that interfere with comprehensibility; and
- mostly abide by writing convention in academic writing despite spelling errors that interfere with comprehensibility and failure to paragraph source-based text or acknowledge sources.

As a result, test takers placed into Level B are required to take an ESL intermediate writing course which focus on grammar and writing at paragraph levels (101B). Meanwhile, a level C/D is intended to mean that this test taker can:

- adequately organize their response using simple transitional devices despite of some lack of focus in paragraphs and some inappropriate use of transitional devices;
- mostly support their arguments with relevant details from external sources although some elaboration is needed;
- display good control of simple and some complex grammatical structures and a range of vocabulary despite some inappropriate vocabulary use and errors that might occasionally interfere with comprehensibility; and
- mostly abide by writing convention in academic writing such as correct spelling and appropriate paraphrasing, despite some spelling errors and failure to acknowledge sources.

So, those receiving a C or D must take an ESL advanced writing course for undergraduate students, which focuses on writing at the discourse level and source use (101C),

or an ESL advanced writing course for graduate students, designed to mainly develop academic and professional writing skills at graduate level (101D).

In addition, the plus levels (i.e., B+ and C/D+) were added to allow for a more fine-grained classification of test takers. For example, level B+ was used to place test takers whose writing displayed characteristics of both level B and level C/D. Similarly, level C/D+ was used for writers who showed characteristics of both a level B and Pass test taker. There were no descriptors for the B+ or C/D+ score band. The rubric is presented in Appendix B.

Web-based Training Materials

The rater training materials developed by the EPT Office were used to train the raters participating in this study. The training materials, delivered on Canvas, the University's learning management system, consists of four sections. In the first section, raters familiarize themselves with information about the EPT Writing and the ESL writing courses. The second section presents anchor essays with explanation of their score to raters. In the last two sections, raters practice rating the responses and receive feedback on their rating. The essays used for the online training are responses to the prompt on the topic of homeschooling. Additionally, three responses, representative of the three main score bands (B, C/D, and Pass) on the topic of GM food, the same topic as the essays they would be rating on, were used in the face-to-face norming section at the beginning of the rating session.

Eye-Tracking Equipment and Software

A remote eye-tracker, Gazepoint GP3 eye-tracker (0.5 – 1 degree of visual angle accuracy, 60 Hz) was employed to collect eye-tracking data. The eye-tracker was mounted at the bottom edge of a computer 24-inch display and ran on a PC using Windows 10 64-bit OS. A

second display was used by the researcher to monitor the data collection progress. The eye-tracking data were recorded and processed by Gazepoint Analysis Professional, an add-on software product required to capture the data generated by the eye-tracker. This software provides heatmaps, which capture raters' viewing at a specific time point, and static and dynamic gazeplots, which show raters' eye movement trajectories in time and space. The software also allows for screen video, web multiple user data aggregation, and exports for images, videos, and statistics.

Interview Protocol

The interview protocol employed for the study consisted of two sections. The first section sought information regarding raters' background, such as education, years in ESL instruction, and experience with rating EPT written responses, to learn about the raters' profile. The second section asked for the raters' opinions about the rubric to address RQ1 regarding raters' opinions about the appropriateness of the rubric for evaluating test takers' performances. The main question in this part is "*what do you think about the rating scale?*" Depending on what the raters said, I asked follow-up questions, such as "*How is it useful or not useful for you to grade the essays?*", and "*Are there any things that need to be changed (in terms of wording, categories on the scale, descriptors of each category, number of score levels)? Please explain why*". A copy of the interview protocol is presented in Appendix C.

Procedures

This section describes the procedures for collecting the operational and experimental ratings, interviews, and stimulated recalls. An IRB approval was obtained to conduct the study.

The de-identified, operational ratings for all writing responses in Fall 2018 were collected from the EPT Office with an approval from the EPT Office Director.

Operational EPT Writing Ratings

To address RQs 2 – 5, the official, operational ratings given to 283 test takers, each of whom submitted two task responses, in Fall 2018 were used for the study. These ratings were given by a total of 10 trained EPT Writing raters using a 5-point rubric described earlier. The raters were graduate students and faculty members in the Applied Linguistics and Technology program, most of whom had worked as ESL instructors.

During the EPT Writing rating sections, the 10 raters were paired to grade both responses to the Summary task (Task 1) and the Essay task (Task 2) produced by groups of 10 test takers. One rater rated the summaries while the other rated the essays. After that, they switched the responses and continued rating. The raters recorded their grades on a scoring sheet for each task and were discouraged from discussing their grades. Therefore, the ratings provided by the raters were considered independent. After raters finished with a group of test takers, they were paired with a different partner to grade responses produced by another group. Although the ratings were not fully crossed, the rater pairing was done in a way that allows for linking between ratings provided different raters.

Table 3.2. Rating Design for Fall 2018 EPT Writing

TOTAL NUMBER OF...		RATER									
Test takers (N = 283)	Responses (N = 566)	6	8	11	13	9	7	12	14	15	5
42	84		x	x			x		x		
20	40	x				x					
19	38			x	x						
19	38					x		x			
13	26	x		x							
10	20	x	x								
10	20	x								x	
10	20	x			x						
10	20	x									x
10	20			x					x		
10	20				x	x					
10	20				x						x
10	20					x	x				
10	20					x					x
10	20						x			x	
10	20							x	x		
10	20							x			x
10	20								x	x	
10	20							x			x
8	16		x			x					
7	14		x	x							
5	10	x					x				
5	10				x					x	
5	10						x		x		
Total number of ratings per rater		156	134	182	108	154	144	98	134	90	100
Total number of ratings		1,300									

It should be noted that responses from 42 randomly selected test takers were rated by four raters, who were also selected randomly, to address RQ5 regarding the score reliability (i.e., the G-theory studies). This was to ensure that the G-study design would be fully crossed. That is, both responses from each test takers were rated by all four raters, resulting in a total of 336 ratings for the G-study (42 test taker x 2 tasks x 4 raters). This sample size is considered adequate for a robust estimation of G and Phi coefficients in the G-study and D-study (Atilgan, 2013). The rating design for the EPT Writing in Fall 2018 is presented in Table 3.2.

Overall, a total of 1,300 ratings from Fall 2018 were used for MFRM analysis to address RQs 2 – 4. Among these, 336 ratings were used for G-theory study to address RQ5.

Experimental Ratings, Stimulated Recalls, and Interviews

The experimental ratings, stimulated recalls, and interviews were collected from nine raters who volunteered to participate in this part of the study. To recruit the raters, I emailed information about the study and electric copies of an informed consent form to graduate students in the Applied Linguistics and TESL program at a large Midwest institution. The procedure for collecting the experimental ratings, stimulated recalls, and interviews from the rater participants is summarized in Figure 3.2.

After signing the consent form showing their voluntary participation in the study, the nine raters underwent the online rater training individually and the face-to-face norming session with the whole group. They then rated 60 essays on paper, which was similar to the regular rating session for the EPT Writing. After that, they met with me individually for the eye-tracking rating session.

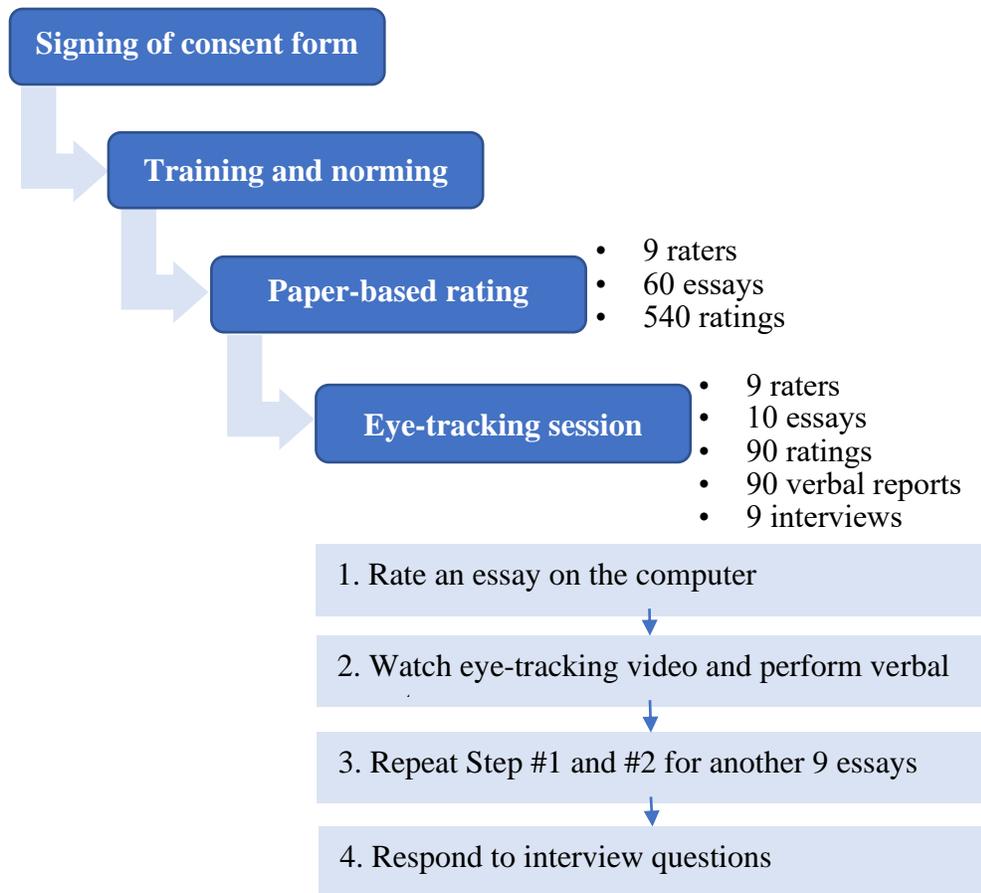


Figure 3.2. Procedure for collecting Phase 2 data: experimental ratings, stimulated recalls, and interviews

In the eye-tracking rating session, the nine raters individually met with me in a computer lab. They sat in front of a PC computer where the eye-tracker was mounted and adjusted so that raters felt comfortable in their sitting position. The distance between raters and the computer screen and the eye-tracker was approximately 25 inches (60 centimeters). The setup for eye-tracking and essay rating is illustrated in Figure 3.3.

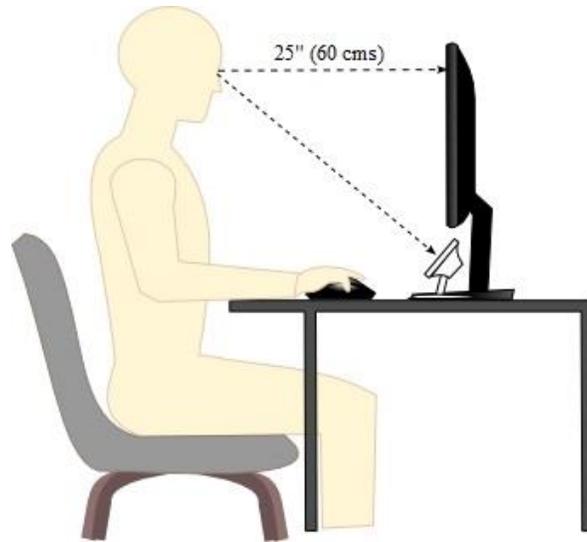


Figure 3.3. Illustration of the eye-tracking and essay-rating setup

The 10 essays randomly selected from the 60 essays, which the raters had already rated on paper, the rubric, and the two source texts, referred to as rating materials, were stored as separate images (1920 x 1080 pixels). To eschew variation in the raters' cognitive processes resulting from different font sizes and to ensure that each essay or source text fits within an image only, the font size for the essays and the source texts were formatted at 18 point font, the largest font size that allowed each essay to be stored as a single image. Each essay was presented with the rubric and source texts in the following order: rubric – essay – source text 1 – source text 2. Instructions about how to navigate around the images (i.e., how to move back and forth among the four images) was given at the beginning of the rating session. After that, calibration of the raters' eye movement was conducted and repeated after every break at hourly intervals. Every time a rater looked at each of the rating materials, their eye movement trajectories on that rating material were stored as a separate recording. For example, if a rater read the rating materials following the following sequence (1) rubric – (2) essay – (3) source text 1 – (4) essay, their eye movement trajectories were stored in four separate recordings. In the end, this eye-

tracking data collection process resulted in a total of 90 sets of records (9 raters x 10 essays), each of which varied in the number of recordings ranging from one to 10, depending on how many times the raters moved back and forth between the rating materials.

After finishing rating each essay, the raters were asked to watch the eye-tracking recordings, which showed their eye movement trajectories and gaze points when they observed each of the rating materials (i.e., the essay, rubric, and two source texts), and recall the processes they had undergone when rating the essays. It should be noted that since the raters' eye movement trajectories on rating materials were stored as separate recordings every time they moved back and forth between two rating materials, the raters had a different number of recordings depending on the number of times they visited a rating material. The use of their eye movement trajectories, along with their gaze points, was intended to facilitate the recall of their thought processes during their rating of the essay. Figure 3.4 below shows a screen capture of a video recording of a rater's eye movement trajectory when this rater was reading part of an essay.

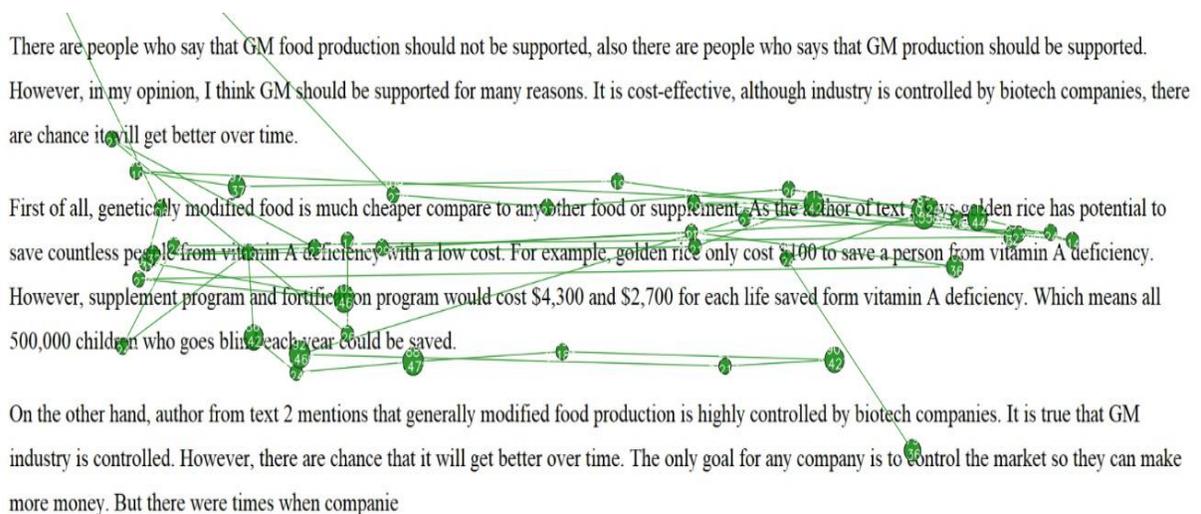


Figure 3.4. Screen capture of a video recording from Gazepoint GP3 showing a rater's eye movement trajectories

In Figure 3.4, the green lines show raters' eye movement trajectories while the green dots indicate their gaze points. The size of the green dots is proportional with the time raters spent gazing at a certain point. It should be noted Figure 3.4 shows all trajectories and gaze points for paragraph 2. However, raters only saw the trajectories and gaze points in chunks of three seconds at a time. Raters could pause the recordings whenever they wished to verbalize their thoughts. When necessary, I asked probing questions, such as "*what were you thinking at this point?*" "*what was going on?*" or "*what were you having in your mind?*" I also asked clarifying questions, such as "*what do you mean by that?*"

After each of the nine raters finished all 10 essays, I interviewed each one for their background information, and their opinions about the rating rubric using the interview protocol. Raters' verbal reports and interviews were audio-recorded. The estimated average total participation time for each rater was approximately 6 hours. The raters were compensated with 12 US dollars per hour for their participation.

In summary, the data analyzed for the study include:

1. Operational ratings provided by the EPT Office (N = 1,300). They were given by 10 raters to 283 test takers on the Summary task (Task 1) and the Essay task (Task 2).
2. Experimental ratings given on 60 Essays task responses (Task 2) by each of the nine raters participating in the eye-tracking study (N = 630). They include 540 ratings from the paper-based rating of the 60 essays and 90 ratings from the eye-tracking session during which raters rerated 10 of the 60 essays.
3. Interviews collected from the nine raters (N = 9)
4. Verbal reports provided the nine raters about the 10 essays each rated (N = 90)

Data Preparation and Analyses

The operational ratings and experimental raters were converted to a scale of 1 – 5 (1 = B, 2 = B+, 3 = C/D, 4 = C/D+, 5 = Pass) in preparation for the MFRM analyses in *Facets*. To prepare for the analyses of the qualitative data, audio recordings of the verbal reports and interviews were transcribed verbatim by *rev.com*, a third-party transcribing service. This resulted in a total of nine interview transcripts (Appendix D) and 90 verbal report transcripts (Appendix E). After that, I checked the accuracy of the transcripts before they were analyzed qualitatively. The average text length is approximately 851 words for the verbal report transcripts and 1,048 words for the interviews. The transcripts were then coded by me and another graduate student in TESOL/applied linguistics. Both of us worked together to create the coding scheme, although I also prepared the data and logistics needed for the coding, such as NVivo, the software for qualitative research.

The analyses of the data were conducted in two main phases following the research design. In Phase 1, quantitative analyses of the operational ratings were conducted to address the research questions about the usefulness of the rubric score bands, (RQ2), rater performance (RQs 3 and 4), and score reliability (RQ5). In Phase 2, qualitative analyses of the interviews and verbal reports were conducted to investigate raters' decision-making processes (RQ6) and their opinions on the rubric (RQ1). Additionally, the experimental ratings were analyzed using MFRM to examine the effects of the rating medium (paper-based vs. computer-based). Although this analysis did not directly address the research questions, it was important because if the raters rated differently on the two media, findings about their decision-making processes, collected from the computer-based rating, might not be generalizable to other rating conditions. A summary of the analyses of the data to address the research questions is presented in Table 3.3.

Table 3.3. *Summary of Analyses Conducted to Address the Research Questions*

	Research question	Data	Analysis
RQ1	What did experts think about the appropriateness of the rating scale for evaluating test takers' performance?	Interviews with nine raters (N = 9)	Content analysis; raw counts of comments about the rubric features
RQ2	Was there statistical evidence supporting the number of scale steps?		MFRM analysis: score band average measures and mean-square outfit statistic, Andrich-Rasch thresholds
RQ3	Were raters comparable and consistent in their rating?	Operational ratings for 283 test takers on two tasks provided by two to four raters (N = 1,300*)	MFRM analysis: Wright map, reliability and strata indices, Chi-square statistic, fit statistics
RQ4	To what extent did raters exhibit bias against certain test tasks?		MFRM analysis: Bias analysis
RQ5	What was the score reliability when two tasks are included, and each response double-rated?	Operational ratings for 42 examinees given by four raters on two tasks (N = 336)	G-study with a fully crossed, two-facet design (42 examinees x 2 tasks x 4 raters), D-study
RQ6	What writing features did raters attend to when rating source-based writing task responses?	Experimental ratings given by nine raters on 60 Essay task (Task 2) responses (N = 630**)	MFRM analysis: Bias analysis
		Verbal reports from nine raters (N = 90)	Content analysis, percentages of instances of writing features

* The total number of operational ratings (1,300) = 964 ratings (241 test takers x 2 tasks x 2 raters) + 366 ratings (42 test takers x 2 tasks x 4 raters)

** The total number of experimental ratings (630) = 540 on-paper ratings (60 essays x 9 raters) + 90 ratings on the computer (10 essays x 9 raters)

RQ1: What Did Experts Think About the Appropriateness of The Rating Scale for Evaluating Test takers' Performance?

To address RQ1 regarding raters' opinions about the rating rubric, the transcribed interviews were examined through an inductive approach to which themes and patterns emerged from the data using the interview questions as categories of analysis (Paltridge & Phakiti, 2010). The comments were segmented according to the content of each unit that would allow for analysis of raters' opinions on features of the rubric such as *Descriptor Clarity*, *Score Band Number*, and *Category Weighting*. If a single idea was repeated without adding new information, it was treated as a single text unit. For example, the excerpt below was segmented into two content units, as indicated by the double slash and the numbers. Note that the underlined part is repetitive of (1), which discussed the clarity of the descriptors, and thus was not treated as a new content unit.

“(1) *It's good, detailed. // (2) Actually, I most like the portion like organization and grammar, I think. It's good because they are important things in student essays... // Actually, once you get used to it, it's very good. When you're teaching those levels, it's understandable. It's very clear. You know what the rubric says.”*

Other themes that emerged from the data included *Criteria Relevance* and *Score Band Labeling*. The content units in each theme were then classified as “positive”, “mixed”, or “negative”. If comment contained adjectives such as “useful”, “beneficial”, “better”, “good”, “clear”, or “straight-forward”, it was coded as positive. A comment was also coded “mixed” if it showed mixed feelings from raters, such as in the following example:

“I have two views about that. From the testing perspective, I think this [having more score bands] is good. From the teaching perspective, I think students in 101C can survive the college life in terms of the writing ability. So, even without taking 101C, I guess they can survive the 4 years at [university] whereas in 101B they definitely need training... I don't know. So, I think my opinion might be a big issue. I personally think the scoring can be binary, pass or fail.”

Meanwhile, a negative comment included words such as “hard”, “difficult”, and “not really useful”. A complete coding scheme for the interviews is presented in Appendix F.

Intercoder reliability coefficients were relatively high, with percentage of agreement between the two coders at 89.4%, Cohen's kappa at .881, and Krippendorff's alpha at .882. We used *NVivo 12*, a qualitative analysis package for coding, retrieving, and reviewing textual data, to code all 10 interview transcripts. The coded data were then exported to Excel and compared. We discussed any discrepancy in coding to arrive at a consensus.

After that, the frequency of each code was calculated. Also, the number of raters who reported negative, neutral, or positive attitudes to different aspects of the rubric (such as wording, number of score levels, and clarity of score descriptors) was also reported. The coders then selected quotes representative of positive and negative attitudes to the rubric to illustrate raters' opinions to these aspects.

RQ2: Was There Statistical Evidence Supporting the Number of Score Bands?

To address the second research questions related to the rubric's ability to distinguish the proficiency levels, MFRM analysis was conducted on the operational ratings for 283 test takers

($N = 1,300$) using *Facets V.3.71.4* (Linacre, 2014). MFRM is an extension of the one-parameter Rasch model in the Item Response Theory (IRT) family. MFRM is a linear model that logistically transforms polytomous scores on a performance-based test into an interval logit scale (Eckes, 2015). In language assessment, it is used to adjust the estimate of the targeted ability while controlling for the effects of other facets. It also provides individual-level diagnostic information to help researchers identify problematic individuals, for instance, raters, tasks, or rating scale categories, so that they can provide additional training to raters, revise the tasks, or revise the rating scale (Lynch & McNamara, 1998). Additionally, MFRM is used for bias analysis to analyze interactions between facets, for example, whether raters' severity is significantly different when rating different tasks. For this study, three facets, namely *examinee*, *rater*, *task*, and the interaction between *rater* and *task* were modeled.

First, MFRM's assumptions of unidimensionality, local independence, and certainty of responses (Ockey, 2012) were examined. The unidimensionality assumption requires that score differences are attributed to one single trait. This assumption of psychometric unidimensionality, implying that test task or rating criteria work together to form a single underlying pattern of empirical observations, was checked using the task fit statistics (Eckes, 2015). The local independence assumption states that responses to each test item and task are independent of one another. Thus, the assumption of local independence with regards to raters and test ratings was examined using Rasch-Kappa (Eckes, 2015) as well as the process of test administration and rating. The third assumption, assumption of certainty of responses, requires that test takers exert efforts to complete tasks.

In addition to the assumptions for MFRM, global model fit, i.e., whether the data fit the model, was also examined using log-likelihood chi-square statistic. If the chi-square statistic,

standardized residuals would be examined because the chi-square statistic is sensitive sample size effects (Eckes, 2015). The data are considered to fit the model if (1) the number of standardized residuals less than $|\pm 2|$ accounts for less than 5% of the observations, and (2) the number of standardized residuals less than $|\pm 3|$ accounts for less than 1% of the observations (Eckes, 2015).

A number of indices were examined to find evidence supporting the effectiveness of the rating scale. First, average measures by score band was examined to see if the measures increased monotonically, which is an indication that test takers receiving higher grades correspond to “more” of the variable being measure. Additionally, the mean-square outfit statistic computed for each score band was also scrutinized. A value under 2 would indicate that the difference between the average and expected examinee proficiency measures are within the acceptable range (Eckes, 2015). Finally, the Andrich-Rasch threshold values were examined. For evidence that the number of scale steps are reasonable, these threshold values should increase in a linear fashion and should be separated well enough before we can argue that the raters, as a group, were able to distinguish all the levels of the scale. Linacre (2002) argued that the minimum distance between thresholds is related to scale length, but for a 5-point scale, a separation of 1.0 logit suffices.

RQ3: Were Raters Comparable and Consistent in Their Rating?

To address the third research question regarding raters’ comparability and consistency in terms of severity, the Wright map, rater separation ratio and strata index along with their reliability (KR-20), Chi-square statistic, and fit statistics generated by the MFRM described in RQ2 above were examined. In terms of rater comparability, the Wright map was first examined

for tight clustering of the raters. Additionally, a nonsignificant Chi-square statistic (i.e., homogeneity index) indicates that raters are similar in severity (Barkaoui, 2014; Eckes, 2015). If the Chi-square statistic is significant, Wald statistics can be computed to detect difference in severity estimates between two raters (Eckes, 2015). The selection of the rater separation ratio or strata index, both of which are measures of the spread of the rater severity estimates relative to their precision, depends on the actual distribution of the rater severity estimates. That is, if the distribution has heavy tails or outliers, strata index should be examined. Meanwhile, reliability of rater separation index provides information about how well the elements within the rater facet are separated in order to define reliably the facet and represents the proportion of the observed variance of rater severity measures that is not due to measurement error (Eckes, 2015). A value of separation ratio, strata index, and reliability closer to zero means that the raters are homogeneous in their severity.

In terms of consistency, fit statistics were examined. These statistics show the degree to which a rater is (1) internally self-consistent across test takers, criteria, and tasks and (2) able to implement the rating scale to make distinctions among test takers' performance (Bond & Fox, 2007). An acceptable range for this index is .7 – 1.3 (McNamara, 1996), showing that raters use the scale consistently and maintain their personal level of severity.

RQ4: To What Extent Did Raters Exhibit Bias Against Certain Test Tasks?

To address RQ4 regarding the level of bias displayed by raters to the tasks (Summary vs. Essay), the interaction between *rater* and *task* in the MFRM analysis conducted above was examined. When the *t*-test comparing a rater' ratings for Task 1 and those for Task 2 was significant (*p*-value lesser than or equal to .05), indicating that this rater rated differently

depending on the tasks, the bias size (i.e., effect size) and its sign was examined to determine which task this rater had favored and which task they had biased against.

RQ5: What Was the Score Reliability When Two Tasks Are Included, and Each Response Double-Rated?

To address RQ5 regarding score reliability when two tasks were included and two raters were employed to rate each response, G-theory studies were conducted. Stemming from Classical Test Theory (CTT), G-theory examines the sources of measurement error or factors that might influence performance and determines the relative impact of these different potential sources of error applying random effects of Analysis of Variance (Eckes, 2015, Lynch & McNamara, 1998; Marcoulides & Ing, 2015). In language testing, G-theory is used to examine the relative effect of the variability of each of the sources (such as examinees, raters, and tasks) on test scores as well as to estimate score dependability for different decision-making purposes so that test developers can make better choices in test design (Lynch & McNamara, 1998). G-theory is based on a sampling framework; so, it requires that an object of measurement and conditions of the facets should be randomly selected from the populations and the universe of observations so that the characteristics found in the sample can be generalized to the population (Kane, 2002).

First, a univariate G-study was conducted using the *gtheory* package in *R* (Moore, 2016) with the fully-crossed, two-facet design (42 examinees x 2 tasks x 4 raters) to specify and estimate the relative effects of variation in test tasks and rater judgments on test scores (Bachman, Lunn, & Mason, 1995; Marcoulides & Ing, 2014). In this G-study design, persons (p) were treated as the objects of measurement. The tasks (t) and raters (r) were defined as the

random facets because both tasks and raters could be regarded as random samples selected from their respective universe of interest.

Results from the G-study provided the basis for the D-study where (1) the relative importance of the effects of different facets and the interactions among the facets, and (2) the dependability index of the test scores could be estimated. The D-study results were then used to suggest the optimal number of the raters and tasks needed to arrive at an acceptable reliability of at least .7.

RQ6: What Writing Features Did Raters Attend to When Rating Source-Based Writing Task Responses?

The main focus was on the qualitative analyses of the verbal reports to investigate the writing features the raters attended to (RQ6). Additionally, the quantitative analysis of the experimental ratings was conducted using MFRM to examine the effects of the rating medium (paper-based vs. computer-based). Although this analysis did not directly address the research questions, it was important because if the raters rated differently on the two media, findings about their decision-making processes, collected from the computer-based rating, might not be generalizable to other rating conditions.

Effects of the Rating Medium (Paper-based vs. Computer-based) on the Raters

Since the experimental rating was done on the computer so that eye-tracking data could be collected, the first step was to examine if the rating medium (i.e., paper-based vs. computer-based) had an effect on the raters' rating. Therefore, bias analysis was conducted in *Facets* to ascertain that rating on the computer with eye-trackers did not affect the raters significantly since this was not the usual operational rating condition for the EPT Writing. Three facets, namely

examinee, medium (paper vs. computer), and raters, with interaction between medium and raters were modeled on the 630 experimental ratings provided by the nine participants on the Essay task (Task 2) responses. These ratings resulted from the raters' scoring of 60 essays on paper (N = 540) and 10 essays, selected from the 60 essays, on the computer (N = 90).

Writing Features Raters Attended to When Rating Source-Based Writing Task Responses

Ninety verbal reports collected from the nine rater participants were analyzed to address RQ6 on the writing features the raters attended to when rating source-based writing task responses. This analysis included four stages: creating the codebook, establishing reliability, independent coding, and data presentation. Figure 3.5 summarizes the steps taken to analyze verbal reports to address RQ6.

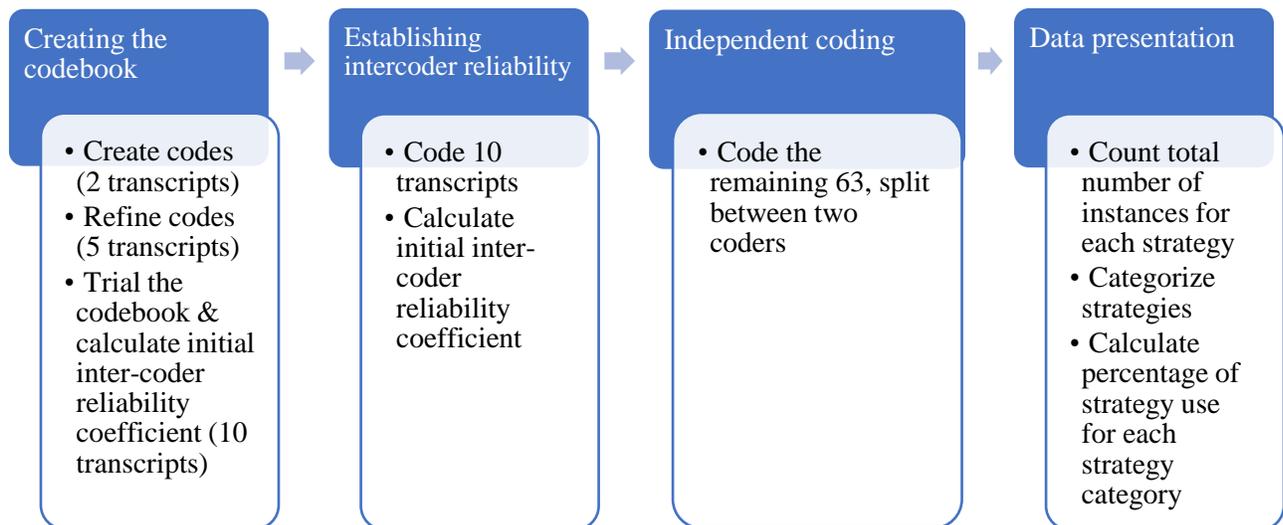


Figure 3.5. Steps for analyzing 90 verbal reports

Creating the Codebook. The process of creating the codebook involved different sub-processes such as initial coding, creating the codes, and trialing the codebook. Each of these sub-processes will be described in detail below.

In the initial coding, which aimed to get an impression of the writing features the raters attended to and provide training to the coders. First, the other coder and I segmented, and then coded two transcripts selected at random using the categories based on language knowledge and fundamental processes included in the EPT Writing construct, which is presented in Chapter 2. The construct includes textual knowledge, grammatical knowledge, sociolinguistic knowledge, and source use. *Textual knowledge*, including knowledge of cohesion and knowledge of rhetorical organization, refers to test takers' ability to produce explicitly marked relationships among written sentences and to produce organizational development in written texts. *Grammatical knowledge* involves test takers' ability to use syntax and vocabulary to produce formally accurate sentences. *Sociolinguistic knowledge* allows test takers to create language appropriate for written academic English using appropriate style and conventions. *Source use* includes test takers' ability to identify and use appropriate information from source text materials for their argument. Thus, the categories included in the coding scheme were identified as *Arguments & Organization* (i.e., textual knowledge), *Grammar & Lexis* (i.e., grammatical knowledge), *Style & Conventions* (i.e., sociolinguistic knowledge), and *Discourse Synthesis* (i.e., source use). The first category, *Arguments & Organization*, includes codes such as topic development, relevance of ideas, coherence, text organization, writers' stance, and ideas or rhetoric. The second category, *Grammar & Lexis*, refers to test takers' use of English, such as word use, syntax or morphology, overall comprehensibility, overall language, and spelling and punctuation. *Style & Conventions* refers to test takers' use of correct style, register, discourse functions, or genre for academic English writing. *Discourse Synthesis* includes accurate understanding of source texts. Features that did not belong to these categories were coded as

“extra” and then classified into themes. Table 3.4 summarizes the aspects of the test construct, their definitions, and coding scheme categories used for the initial coding.

Table 3.4. *Aspects of the Test Construct, Their Definitions, and Coding Scheme Categories Used for the Initial Coding*

Aspect of construct	Definitions	Coding scheme category
Textual knowledge - Knowledge of cohesion - Knowledge of rhetorical organization	ability to produce explicitly marked relationships among written sentences and to produce organizational development in written texts	Arguments & Organization - reasoning, logic, or topic development - relevance of ideas - coherence - text organization - writers' stance - ideas or rhetoric
Grammatical knowledge - Syntax - Lexis	ability to use syntax and vocabulary to produce formally accurate sentences	Grammar & Lexis - word use - syntax or morphology - overall comprehensibility - overall language - spelling and punctuation
Sociolinguistic knowledge	use of correct style, register, discourse functions, or genre for academic English writing	Style & Conventions - style, register, discourse functions, or genre
Source use	accurate understanding of source texts	Discourse synthesis - understanding of source texts

Following Lumley (2005), a separate segment was used for (1) each time rater read or referred to the rating rubric, (2) each separate comment from raters related to the writing features in the code book, and (3) each separate comment from raters about the score for a given

category. However, when a single idea was repeated without adding new information, it was treated as a single text unit.

After the initial coding, we randomly coded another five transcripts to create new codes that fit within existing categories or refine the existing category codes. Specifically, we created new codes that emerged from the data, which were identified as “use of transitions” in *Arguments & Organization*, “source use convention” in *Style & Conventions*, and “citation quality” in *Discourse Synthesis*. Also, we refined “syntax or morphology” into “grammatical complexity and accuracy”. Extraneous features, classified as *Others*, included “task completion” and “text length”. A complete list of codes for rater attention is presented in Appendix F. At this point, any coding discrepancies between the coders’ coding were resolved through discussions to arrive at higher inter-rater reliability in the future.

After the codebook had been created, we trialed it by coding ten randomly selected transcripts. Three reliability coefficients, namely agreement percentage, Cohen’s kappa, and Krippendorff’s alpha, were calculated using the *irr* package (Gamer, Lemon, Fellows, & Singh, 2012) at this point to examine the initial intercoder reliability. While agreement percentage and Cohen’s kappa are common indices used to show intercoder reliability, Krippendorff’s alpha has been used in content analysis, a research technique for making replicable and valid inferences from data (e.g., textual, audio, pictorial, and video data) to their context (Krippendorff, 2004a). This coefficient was calculated instead of the familiar agreement percentage or Cohen’s kappa because of its robustness, including its ability to account for chance agreement, be applied for any number of observers or coders, handle missing data, and correct the issue of small sample sizes (Krippendorff, 2004a, b, c, 2011; Lombard, Snyder-Duch, & Bracken, 2002). At the stage, agreement percentage was reported at 71.3%, and Cohen’s kappa and Krippendorff’s alpha at

.63, lower than .67, the acceptable agreement between coders (Krippendorff, 2004b). Therefore, we discussed instances where we disagreed as the additional training.

Establishing Intercoder Reliability. After the additional training, we independently coded another ten randomly selected transcripts and re-calculated the intercoder reliability coefficients to establish intercoder reliability for the data. These coefficients improved relatively substantially, with agreement percentage at 82.3% and both Cohen's kappa and Krippendorff's alpha at this stage were reported at .75, higher than the acceptable level of .67.

Independent Coding. As Krippendorff's alpha indicated acceptable intercoder reliability, we then split the remaining 63 transcripts and coded them independently in *NVivo 12*.

Data Quantification. Then, the coded texts were exported to MS Excel for analysis. I counted the total of instances when each code was used by each of the nine raters on each of the ten essays. After that, the other coder and I grouped the features into five broader categories, namely *Arguments & Organization*, *Grammar & Lexis*, *Style & Conventions*, *Discourse Synthesis*, and *Others*, based on the construct that the test is intended to measure.

To address the RQ6 pertaining to the writing features the raters attended to when rating the essays, I summed the raw counts of all features belonging to the same broad group (i.e., *Arguments & Organization*, *Grammar & Lexis*, *Style & Conventions*, *Discourse Synthesis*, and *Others*) across the nine raters. Then, I calculated the percentages of feature for each group. After that, another coder and I selected verbal report segments representative of the writing features to demonstrate the writing features attended to by the raters.

Chapter Summary

This chapter explained the research design, the procedures for data collection, as well as the data analysis methods from both the quantitative and qualitative aspects of the research. The first section described the design of the study, the mixed-methods multiphase design (Creswell & Plano Clark, 2012), with detailed information about the context, participants, and the materials and instruments employed for the study. The second section explained the procedures for obtaining the data in the two phases of the study: the quantitative data such as the operational ratings (Phase 1) and experimental ratings (Phase 2), and the qualitative data, including the stimulated recalls and interviews collected from raters during their rating of the Essay task (Task 2) responses. The last section described the analyses conducted to address the research questions. Those analyses include statistical procedures such as MFRM and G-theory, and content analysis of the stimulated recall and interview data. The next chapter presents in detailed results of these analyses along with a discussion of what their meaning.

CHAPTER 4: RESULTS AND DISCUSSION

This chapter provides answers to the research questions using the results of multiple analyses outlined in the Methodology Chapter. Specifically, each answer targets a particular assumption for a certain warrant in the IAU. This chapter covers results revealing (1) raters' opinions about the appropriateness of the rating scales, (2) the functionality of the scale, (3) the performance of raters in terms of comparability, consistency, and bias, (4) the reliability of the ratings with the number of tasks and raters at the time of the study, and (5) the raters' attention to the writing features when rating source-based academic writing task responses. Four types of data were analyzed to address the research questions: the operational EPT ratings of the Summary task (Task 1) and Essay task (Task 2) responses, the experimental ratings of the Essay task responses, the stimulated recall data collected from nine raters, and the interview data with the raters. The first research question about raters' opinions about the appropriateness of the rating scale was answered by a qualitative analysis of the interview data from the nine raters. Research questions two, three and four related to the functionality of the scale and raters' performance were addressed by a quantitative analysis of the operational EPT ratings using MFRM. The fifth research question pertaining to the rating reliability with the number of tasks and raters employed for the EPT Writing at the time of the study was answered by a quantitative analysis of the operational EPT ratings using G-theory before a D-study was conducted. Finally, the last research questions regarding rater cognition while rating the reading-to-write responses was addressed by a qualitative analysis of the stimulated recall data. The presentation and discussion of the results in this chapter are structured by the research questions.

Raters' Opinions about the Appropriateness of the Rating Scale

The first research question investigated the opinions of raters regarding the appropriateness of the rating scale use for grading the source-based academic writing test responses. To address this question, transcribed interview data with nine raters were segmented and coded into themes. The coders identified a total of 58 unique comments (content units) from the transcripts. Table 4.1 below summarizes the number of comments on different aspects of the rubric.

Table 4.1. *Summary of Raters' Comments on The Rubric*

	Positive	Mixed	Negative	Total comments
Overall usefulness	8 (8)*	1 (1)	0 (0)	9
Descriptor clarity	7 (7)	1 (1)	6 (2)	14
Criteria relevance	4 (3)	1 (1)	8 (7)	13
Number of score bands	5 (5)	4 (4)	0 (0)	9
Score band labeling	0 (0)	1 (1)	1 (1)	2
Category weight	6 (6)	2 (2)	3 (3)	11

** The number in parentheses refer to the number of raters who gave the comments.*

Overall, raters had positive opinions towards the rating scale as indicated by the number of positive comments coded as “overall usefulness”. All of them agreed that the scale was useful. For example, Rater 1 stated that the rubric helped her identify criteria for each score level and thus, provided basis for the rating she gave, as she said:

You need to identify like the criteria of a certain band. You can't just assign, like, B randomly. So definitely, when I was grading, I would go and make sure that's ok, just like

some checklist thing. The grading scale itself is very important and very helpful (Rater 1, Interview, 02/02/2018).

Rater 5 also highly valued the rubric, as he admitted, “*given that I do not have a lot of experience in academic writing, I think this is good, like a beacon for me.*” One rater, however, had mixed feelings about the usefulness of the rating scale. According to this rater, although this scale was more useful than the previous 3-point scale, he did not find it very useful after he had internalized the criteria on the scale, as seen below.

I think this one is more useful than the previous one. Before, we had the 3-point scale... When I did rater training, it was useful but once I started grading, I just used my criteria. And only when I was not quite sure, I referred back to the evaluation rubric, but I didn't use that much. (Rater 9, Interview, 02/08/2018)

Additionally, the raters provided useful feedback regarding the clarity of the descriptors, relevance of criteria, number of score bands, score band labeling, and weighting of each rubric category.

Descriptor Clarity

Fourteen comments from the raters pertained to the clarity of the scale descriptors. In general, the raters had mixed feelings about the clarity of the descriptors, with seven positive comments from seven raters, one comment showing raters' mixed feelings about this descriptor clarity (Rater 5), and six negative comments from two raters (Raters 7 and 9). According to the raters who thought positively about the descriptor clarity, the descriptors were easy to follow and

very straight forward. These raters thought the language used in the descriptors was clear and detailed enough for them to distinguish the different score bands, as Rater 3 said below.

So that language to me is very clear, although I can't always necessarily say, "what does somewhat organized mean?". I always want to quantify it in some way, I don't know how you can do that; but at least this rubric really works to differentiate between those different levels. Like "arguments are vague", "mostly developed", "fully elaborated". I mean that clearly gives me sort of a way to think about how I should grade this essay, there's not sort of confusion between what a C, D is or isn't and what a Pass should be, at least based on the language of the rubric. (Rater 3, Interview, 02/03/2018)

This point is echoed by Rater 4, who praised the rubric for being detailed but not so complicated that it would cause confusion.

I think the different descriptions in the different levels are pretty good. Some other rubrics I used in the past are a lot harder to use... Here, they include enough details without including too much. Too much is confusing and not enough is also confusing. (Rater 4, Interview, 10/18/2018).

However, while the three raters (i.e., Raters 1, 7, and 9) who had mixed or negative feelings about the descriptor clarity agreed that the scale was quite clear, they proposed several modifications to make the scale clearer. Two raters (Rater 1 and Rater 7) felt that the quantifiers used in the descriptors, such as "somewhat" or "adequately", could be vague. For example, Rater 1 said:

I have a hard time with the “many” and “some”, “many”, “some”, “mostly”. But if I try to visualize it as a certain number, it can help sometimes. But that’s totally on my part how I interpret “many”. (Rater 1, Interview, 02/02/2018)

Similar to Rater 1, Rater 7 also found that the use of quantifiers hindered his understanding of the rating scale, as quoted below.

It is difficult to figure out what is the difference between “somewhat” and “adequately”. ... It is difficult to understand “somewhat organizing” and “adequately organizing”. The root level, that is a perfect definition, but it is difficult to reflect to the grading process. (Rater 7, Interview, 02/01/2018)

In addition to the use of quantifiers, Rater 5 and Rater 9 thought that the rubric should describe more clearly how source integration should be expected for each score level, as indicated by Rater 5 below.

But source integration is the one thing that I’m not sure about. I usually looked at if the students cite the sources correctly. In this case I only found one text which misinterpreted the source texts. In that case, I usually penalize these people. But as long as test takers say something about the source texts, I will consider it integration about the source texts. But I don’t know how integrative it is; so, that’s my issue as a rater. (Rater 5, Interview, 09/06/2018)

Similarly, Rater 9 also reported having difficulties assessing source use and indicated that he needed clearer instruction on what should be considered “proper” in source text integration.

... I don't know where the raters are formatting the wording "sources cited properly". Does that mean that we have to put the year, the names or is it ok to use titles? It's not really clear, this part. (Rater 9, Interview, 02/08/2018)

Other weaknesses mentioned by Rater 7 include the fact the descriptors used phrases instead of full sentences and that there were no descriptors for the two new score bands, B+ and C/D+. In fact, this rater mentioned that since he liked to pinpoint the differences between score bands, he would like to see some descriptions for score bands B+ and C/D+ so that he could highlight any possible differences between these bands and their adjacent levels.

Criteria Relevance

Relevance of the rating criteria received 13 comments, most of which showed raters' unfavorable opinions about this category. The only three positive comments came from Rater 6, Rater 8, and Rater 9, who praised the scale as *"very well constructed"*, *"reflect[ing] the distinct components of the writing ability of the examinees"*, and *"follow[ing] the previous studies in the multifaceted facets of writing ability."* On the other hand, this feature received eight negative comments from seven raters and one mixed comment from Rater 7.

Most raters commented on the role of reading comprehension, citing, and paraphrasing in students' responses and suggested that these three criteria should be included in the rubric. For example, Rater 2 thought that test takers' ability to comprehend the source texts, indicated by the accuracy of the source-based information used in their response, should be considered an important factor in the rubric.

I'm not sure how much the writer's comprehension of the text factored in. And I feel like that should play a role because if for whatever reason... and I'm debating because it might be that they're not expressing themselves properly, which might be a thing of itself anyway. Just my thought process is if someone is reading a text but doesn't understand it and then gives an argument and then we ignore the accuracy of the idea or the argument, but the language itself was good, so we'll give them a Pass. When they come to say me when I'm teaching 150, they're in trouble because then when I'm grading their paper it will be "no this isn't what the material is saying". And so that was something I was thinking about when I was reading and I'm not sure how you would do it but that was my thought was some way of measuring or evaluating their comprehension of the text. (Rater 2, Interview, 02/03/2018)

Rater 2's comment on the importance of reading comprehension ability was echoed by Rater 6, who affirmed that reading comprehension was crucial for test takers' ability to integrate ideas from the source texts into their response.

... So, they could find the main idea or a good example from the texts and react to them. That means their reading ability is good, right? Or finding the overall idea of the text. Then reading ability is important because they should know what the texts are talking about, the main ideas or the general ideas. (Rater 6, Interview, 10/26/2018)

Rater 3 remarked on the position of citing and paraphrasing in the rubric, stating that this criterion should be part of the *Arguments and Details* category to highlight its importance instead of the *Conventions* category.

I think the citing external sources gets a little bit lost. I wonder if this would put up in Arguments and Details... because for me paraphrasing and citing external sources is important... Even if it's just at the bottom of the page, I feel like paraphrasing and citing almost gets forgotten. And if you want that to be more important, I'm wondering if that was actually put in with Arguments and Details. If that would help because it's hard to separate our supporting details from citing and paraphrasing. (Rater 3, Interview, 02/03/2018)

Similar to Rater 3, Rater 8, believing in the importance of test takers' ability to cite and paraphrase, agreed that citing and paraphrasing did not fit in the "Conventions" category in the rubric. Instead, this criterion qualified to have its own category, as shown in his comment below.

But somehow, I think the first three categories, Organization, Grammar and Lexis, and Arguments should be separated from Conventions because Conventions is a distinct quality... By conventions, I mean source use. It should be a different category. (Rater 8, Interview, 09/29/2018)

However, the raters also believed some criteria should not be included in the rubric. While most raters believed that ability to integrate source texts was essential in sourced-based writing, two raters had opposite opinions about the role of this ability in rating test takers' responses, as Rater 4 commented below.

Maybe I shouldn't do this, but I do think about this whether I like it or not. I subconsciously compare with the students in my 150 classes and I think "how good are they at integrating sources?" Even when they are a native speaker and a good writer, they might not be good. I don't think we should penalize students for that necessarily

because they don't need to be great at this to take 150. They will improve when they take it. But in order to get out of the 101s, I don't think it's necessary. (Rater 4, Interview, 10/18/2018)

In this rater's opinion, students could always learn how to integrate sources into their writing when they take English 150, a first-year composition course. Meanwhile, Rater 9 believed that ability to cite external sources should be a criterion for graduate students instead of for undergraduate ones.

... And then source citation, I don't think this is good criterion for undergraduate students... here in sense that the sources might not be cited not cited a properly. (Rater 9, Interview, 02/08/2018)

Additionally, spelling was also considered inconsequential for two raters, Rater 7 and Rater 8. For example, Rater 7 thought that spelling should be considered irrelevant, as shown below.

... The spelling errors. Should I consider spelling as a criterion?... To me, it is becoming a little irrelevant because in real life, spell check is always on, so why are we doing it during a test? And if that is the case, then why is spelling on the rubric? Because it's not going to be an issue anymore thanks to technology. I don't know.... In the typing skills you can put it in the rubric but maybe in the training we can mention that it is a computer based and some people might not have any typing keyboard skills so that it might affect your overall paper quality. (Rater 7, Interview, 02/01/2018)

To sum up, most of the 13 comments on the relevance of the rubric criteria pertained to how to improve the rubric in terms of the role of reading comprehension, integrating source texts, and spelling. While most raters agreed that reading comprehension and integrating source texts were important criteria, two raters (Raters 4 and 9) believed the latter to be irrelevant for the target test takers. Spelling was also deemed an unimportant criterion for evaluating test takers' responses by two raters (Raters 7 and 9).

Number of Score Bands

The raters had positive to mixed feelings about the number of score bands in the rubric, which received five positive comments and four mixed comments from the raters. Five raters, namely Rater 2, Rater 3, Rater 4, Rater 8, and Rater 9, were content with the 5-band rubric with the levels B, B+, C/D, C/D+, and Pass. In fact, the raters found the plus band, B+ and C/D+ to be very useful as they allowed them to have some flexibility when grading. For example, Rater 2 indicated below:

I actually did find the categories (B+/C+) very useful. One thing that I found when I was doing the grading, I felt like I had more options. So those essays I was like, 'I don't know if this is a C or if it's a Pass, it's a C+', and similar with 'I don't know if this is a B or if it's a C, it's a B+'. And so, in that case it was very nice.... because they were also essay where I was like, 'Nope this is a B. Nope this is a C'. But for others where I was like, 'I don't know, it was in the in-between. It was nice having that category to utilize. (Rater 2, Interview, 02/03/2018)

This comment was supported by Rater 3 who confirmed that the in-between bands (B+ and C/D+) helped her make decisions about test takers' writing more easily.

I would say it's hard as a grader to be limited to 3 categories because sometimes I feel like an essay has elements of maybe of a C and elements of a B. By understanding what that means, like placing the student in a B level. I mean that is a whole year of extra English courses and I think maybe grading an essay is more complicated than just having 3 levels. It'd be nice if there could be more flexibility, which is, why it was interesting for me to include like a B+ or a C+. (Rater 3, Interview, 02/03/2018)

Similarly, Rater 4 liked having the B+ and C/D+ score band, which in his opinion, did not complicate rating process.

I think having the in-between levels is good. I don't think the change is bad... I think it is useful because it is not overly complex. Five is still not a lot. But if we have too many choices, then it is more difficult. (Rater 4, Interview, 10/18/2018)

Rater 9 also thought the 5-band rubric was useful. In his opinion, this helped reduce rater errors and thus, was more beneficial in terms of rater reliability.

When we have some students on the bottom line and then the decision is mostly done by which raters they have, depending on which rater they have. I think it's increasing the rater errors. So, I think it's better to have five-band scales... But with the three-band scale, I think we have many raters' errors. (Rater 9, Interview, 02/08/2018)

However, the other four raters, i.e., Rater 1, Rater 5, Rater 6, and Rater 7, had mixed feelings about the number of score bands on the rubric. While the raters appreciated having another score band between C/D and Pass, they believed that B+, the score band between B and C/D, was not necessary, as Rater 1 said below.

I think the distinction between a B and a B+ isn't that useful. The position that I find myself in is the C/D and the C/D+. It's not quite a Pass, but it's not really a C/D. That's the situation that I find myself in most of the time. B and B+... I don't really know how that would be useful, especially if they're going to go into the same B class. A C/D and a C/D+ might be a little but more important because sometimes you just can't give a person a Pass... like, no, no, no. But they're not also that bad. (Rater 1, Interview, 02/02/2018)

Rater 1's remark is further echoed by Rater 7 who said that the big gap between C/D and Pass was relatively large, which justified the need to add C/D+ score band. However, this Rater 7 was also skeptical of the value of B+, as indicated below.

For C and Pass, there's a big difference... So, I was always in between deciding. So that's why I used to refer to my own teaching experience in saying that paper could be a good student in my class or not. So, I felt that it's not really efficient... I would definitely add those extra categories, but I don't know if a student gets a B+, what happens. I'm not sure about that... Other than that, I think that the 4 categories are good. (Rater 7, Interview, 02/01/2018)

This rater admitted that without the C/D+ score band, he had to refer to his teaching experience to make decisions about students' performance. However, this rater, similar to Rater 5 and Rater 6, had doubts about the implications of placing a student in the B+ score band for instruction, and thus, did not believe that this score band was useful. Rater 6 had the same concerns about ESL writing classes available for students placed in the B+ or C/D+ score band, as showed below.

But there's no other classes for C/D+, right?... I think we only have 101B and C... As I say, B+ can go to C/D. Because B is very clear. But B+ is something in the air; so, they can join C/D+ classes if we have C/D+ classes. (Rater 6, Interview, 10/26/2018)

Similar to Rater 6, Rater 5 also thought about the reality of available ESL writing classes when rating students' responses. Rater 5, while admitting that having the extra score bands, B+ and C/D+, were beneficial from the testing perspective, believed that a binary rating scale which distinguished pass and fail students should be sufficient, as he felt some students placed in ENGL101C could be successful without having to take the course.

From the testing perspective, I think this [having B+ and C/D+] is good. From the teaching perspective, I think students in 101C can survive the college life in terms of the writing ability... I wouldn't say that 101C is not necessary, but based on my limited teaching experience, 101C students can survive the college, but 101B definitely they need to take the class. When I classify an essay into C, I usually compare C and Pass. So, if students cannot achieve Pass, then they go to C. So, I think even without Pass, I think some students can survive. I don't know. So, I think my opinion might be a big issue. I personally think the scoring can be binary, pass or fail. (Rater 5, Interview, 09/06/2018)

In sum, all nine raters thought positively about the number of score bands on the rubric because the 5-band scale allowed them to give a more precise rating for test takers' responses and enabled more reliable statistical analyses of the ratings. However, some raters believed the B+ band score might not be very useful and could be removed, especially when thinking about the implications of having the plusses (i.e., B+ and C/D+) for ESL writing courses at the university.

Score Band Labeling

Another feature discussed by the raters was the labeling of the score bands, which received a mixed comment and a negative comment from the raters. Specifically, Rater 1 and Rater 4 thought that labeling the score bands as B, B+, C/D, C/D+, and Pass, which to some extent reflected the ESL writing classes offered at the University, could be confusing to raters. For instance, Rater 1 said:

Probably that [the label] would make knowing what course they're going to put into irrelevant right. Make it more like... less burden thinking process and constantly second guessing your choices because if I feel like I always do that. Am I being too lenient or am I being too harsh? So, I think it would be... probably a number would be better. Like the EPT speaking they use a number. (Rater 1, Interview, 02/02/2018)

For this rater, seeing the score band labels placed more cognitive load on her as a rater. Rater 4 seemed to have a more neutral opinion about the labels, commenting that they might be useful for raters who were familiar with the ESL writing courses but that might not be the case for raters who do know about these courses.

... For somebody who is familiar with the classes that students will take potentially based on these ratings, maybe I think too much about the classes they are going to take. Maybe it's good, maybe it's bad. I'm not sure. But somebody who's more impartial may be thinking the class placement is not helpful. (Rater 4, Interview, 10/18/2018)

In brief, the two raters who commented on score band label thought that changing the labels to numeral ones could help raters familiar with those courses divorce themselves from

their knowledge of the ESL writing courses and thus, make rating less cognitively demanding for them.

Weighting of Rubric Categories

In addition to descriptor clarity, criteria relevance, number of score bands, and score band labeling, the raters also commented on weighting of rubric categories. Specifically, they commented on (1) the appropriateness of the value assigned to each category, which was 30%, 25%, 25%, and 15% for *Organization*, *Arguments and Details*, *Grammar and Lexis*, and *Conventions* respectively, and (2) usefulness of these numbers for rating. Overall, the 11 comments pertaining to weighting of rubric categories showed that the raters had mixed feelings about the value of each rubric category, with six positive comments, two mixed comments, and three native comments from the raters.

Regarding the appropriateness of the value assigned to each category, five raters, i.e., Raters 1, 2, 3, 4, and 6, agreed that the weighting of each rubric category was relatively reasonable, as Rater 1 said below.

I agree that Argument and Organization should have more weight. I feel like that the two components I place a heavy emphasis on are Organization and Arguments and Details, because Grammar and Lexis and Conventions, I think, are pretty easy to teach. (Rater 1, Interview, 02/02/2018)

These raters placed higher emphasis on *Organization* and *Arguments and Details*, and thus, maintained that these two categories should weigh more *Grammar and Lexis* and *Conventions*, which according to them, were easier for students to improve. However, two raters

recommended a higher weighting for ability to source-based information integration, as seen below.

I think the ability to integrate source texts is very important, especially in task 2 because I feel it's very clear in the prompt that it wants you to summarize and incorporate your experience or any background knowledge not just rely on the summarizing part... So, perhaps we should have more weight for it or maybe we should put it in a separate category. (Rater 1, Interview, 02/02/2018)

Similar to Rater 1, Rater 8 also suggested that more emphasis should be placed on source text integration.

... Somehow I feel that Conventions should have more attention. Because this is source-based writing, whether the examinee could paraphrase the original texts or not is a very important criterion to be considered... (Rater 8, Interview, 09/29/2018)

In terms of the usefulness of the numeral values for rating, two raters, i.e., Rater 3 and Rater 8, believed the numeral values attached to each category guided them when making decisions about students' responses.

I think that the numbers are useful because without numbers, I think it suggests that all of these should be weight equally but clearly, they're not... So, it is telling me that I need to consider that, now whether I actually do consider that or not is another story, but I do think that having the percentages there is helpful to sort of guiding us. (Rater 3, Interview, 02/03/2018)

Rater 8, agreeing with Rater 3 in this regard, stated that: “[I looked at the numbers] a little bit, especially when I have conflicting decisions. For example, when I think they did well in convention, and grammar but not very well in organization.” (Rater 8, Interview, 09/29/2018)

Contrary to Rater 3’s and Rater 8’s opinion, Rater 7 reported that having the numeral values attached to the categories did not make any difference in his rating behavior, as he commented below.

I don’t look at that [the percentage] because in my mind, I know Conventions are not really big issue and to me, Arguments and Detail are more important than Organization. So that’s why I’m not really looking at the rates much... They don’t affect my decision. I just ignore them because I have in my mind that to me a paper should have a nice argument there other than just an essay format paper. So, I just ignore them. They don’t bother me. (Rater 7, Interview, 02/01/2018)

It is obvious that for this rater, seeing the percentages was not necessary since he already knew what to expect in a strong response. Rater 5 and Rater 9 also agreed with Rater 7 that the presence of the percentages was not really beneficial to their rating. For example, Rater 5 believed that as his task was to categorize test takers’ responses, the percentages were not very useful to him.

They’re not really useful in the sense that I don’t focus on these numbers. So, this (the score band) is categorical, right? So, we do not score 60, 70... If we were to score them, I think this is beneficial. But because we just classify the essays into these categories, I personally do not care about these numbers. (Rater 5, Interview, 09/06/2018)

To recapitulate, the raters reported that overall, the rubric was useful. They mostly found the rubric appropriate for the EPT Writing test and reported that it boosted their confidence in their rating. However, they suggested improvements in different aspects such as

- clearer quantifiers in the descriptors (Raters 1 and 7),
- clearer the description of source integration (Raters 5 and 9),
- exclusion of spelling (Raters 4 and 9) and source use as criteria (Raters 7 and 9),
- omission of the B+ score band (Raters 1, 6, and 7),
- re-labelling of the score bands (Raters 1 and 4), and
- increased weight for source integration (Raters 1 and 8).

Statistical Evidence Supporting the Number of Score Bands

To answer the second research question related to the usefulness of the score bands in distinguishing the proficiency levels, 1,300 operational EPT writing ratings were analyzed in *Facets* to obtain the score band average measures, the mean-square outfit statistic for each score band, and the ordering of the Andrich-Rasch thresholds.

First, MFRM's assumptions of unidimensionality, local independence, and certainty of responses (Ockey, 2012) were examined. Regarding the unidimensionality assumption, the fit statistics for the two tasks ranged from .92 to 1.05, between the expected range of .5 and 1.5 (Linarce, 2014), indicating that this assumption was met. In terms of the local independence assumption, the Rasch-Kappa index for the current study data was .03, close to the perfect index of zero, indicating that this assumption of local independence in raters at a group level was met. The assumption of the local independence in test ratings was also met because of two reasons: (1) the test takers were not allowed to work together, and (2) the raters rated the test responses

independently. The assumption of certainty of responses was met because of the stakes of the test – that students would need to spend one or two semesters in ESL writing courses, and thus, would not be able to move on to composition courses or meet their college’s language requirement. Additionally, the log-likelihood chi-square value in this analysis was 1151.86 ($df = 1,122, p > .05$), showing that the data fit the model.

Table 4.2 provides information about the functionality of the rating rubric – that is, whether the number of score bands were useful in differentiating writers of different proficiency levels.

Table 4.2. *Score Band Statistics*

Score band	Absolute frequency	Relative frequency	Average measure (in logit)	Outfit	Thresholds	Standard error
B	195	16%	-1.13	1.0		
B+	140	12%	-.54	.9	-.43	.10
C/D	372	31%	.21	.9	-1.11	.08
C/D+	171	14%	.67	1.0	1.17	.08
Pass	386	26%	1.21	1.1	.37	.08

As can be seen from Table 4.2, there was a clear progression of the average measures from -1.13 logits for the lowest score band of B to 1.21 logits for the highest score band of Pass. This result indicates that the higher ratings (with Pass being the highest) corresponded to higher logit scores. That is, test takers who had lower writing proficiency were associated with lower score bands, which is an indication that the five band scores were in the right order, as expected from a good rating scale. The mean-square outfit statistic for each score band was between .9 to 1.1, lower than the acceptable threshold 2.0 (Eckes, 2015), showing that the difference between

the average examinee proficiency measures and the examinee proficiency measures predicted by the model was acceptable.

The thresholds, however, did not advance monotonically with the score bands. Specifically, the threshold between B and B+ was $-.43$ logits (τ_1) while that between B+ and C/D was -1.11 logits (τ_2). Similarly, the threshold between C/D and C/D+ was 1.17 logits (τ_3), higher than that between C/D+ and Pass, at $.37$ logits (τ_4). Figure 4.1 provides a graphic illustration of the threshold ordering. The horizontal axis shows the examinee proficiency scale in logits while the vertical axis denotes the probability of being rated in each score band. The vertical dashed lines indicate the thresholds between the adjacent score bands. The locations of the thresholds on the examinee proficiency scale are estimated primarily from the category frequencies (Linarce, n.d.). In this case, score band B+ and score band C/D+ received fewer ratings, with relative frequency at 12% and 14% respectively, compared to the remaining score bands, which explained the reversed order of the thresholds. This result indicated that the raters did not use the two additional score bands, i.e., B+ and C/D+, very often.

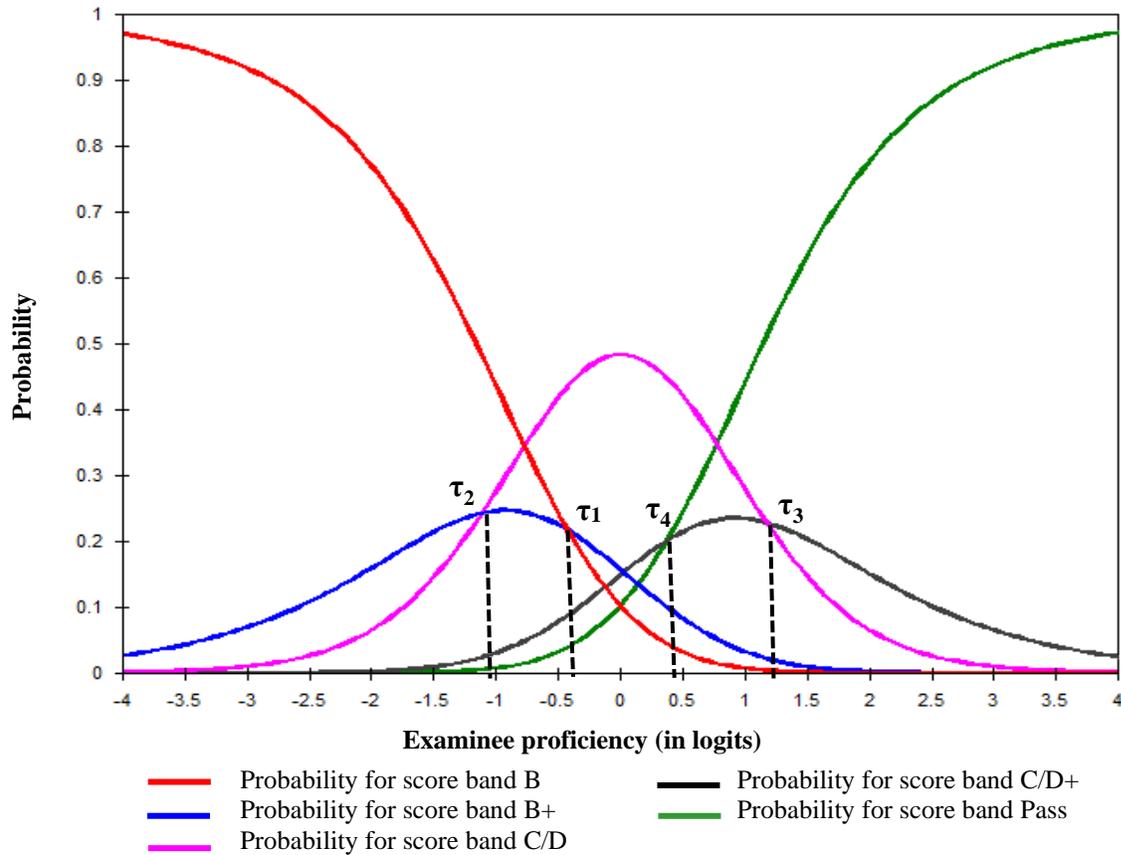


Figure 4.1. Probability curves for individual score bands from MFRM analysis of 1,300 operational ratings in Facets

Overall, the results from the MFRM analysis showed that the score band estimates increased monotonically from the lowest score band (i.e., level B) to the highest score band (i.e., Pass); however, score band B+ and score band C/D+ received relatively fewer ratings compared to the other score bands. These results indicate that the rubric was relatively effective for distinguishing writers at different proficiency levels. However, the fact there were much fewer ratings for the additional score bands, B+ and C/D+, suggesting more rigorous training for raters.

Raters' Comparability and Consistency in Their Rating

To address the third research question related to raters' comparability and consistency, various results provided by the aforementioned *Facets* analysis of 1,300 operational ratings were examined. These results include (1) the Wright map, (2) reliability and separation indices, (3) a Chi-square statistic, and (4) fit statistics for raters.

Rater Comparability

The Wright map, reliability and separation indices, as well as a Chi-square statistic were examined to evaluate raters' comparability in terms of their severity. Figure 4.2 shows the Wright map from *Facets* which places all variables (i.e., examinees, raters, tasks, and rating scale) on the same logit scale. The third column presents the raters' severity or leniency relative to one another when they rated the responses. Raters placed higher in the logit scale are considered more severe. As shown in the Wright map, Rater 12 seemed to be the most lenient rater and different from the remaining raters. Also, Rater 8 and Rater 14 appeared more lenient compared to the remaining raters.

The mean rater severity estimate was $-.25$ ($SE = .1$) with a standard deviation of $.36$ ($SE = .02$). Figure 4.3 provides a visual presentation of the rater severity estimate for individual raters provided by *Facets* along with its 95% confidential intervals. The rater severity estimates in this figure are consistent with the information presented in the Wright map. That is, Rater 11 was the most severe with a logit of $.18$ while Rater 12 was the most lenient with a logit of $-.97$.

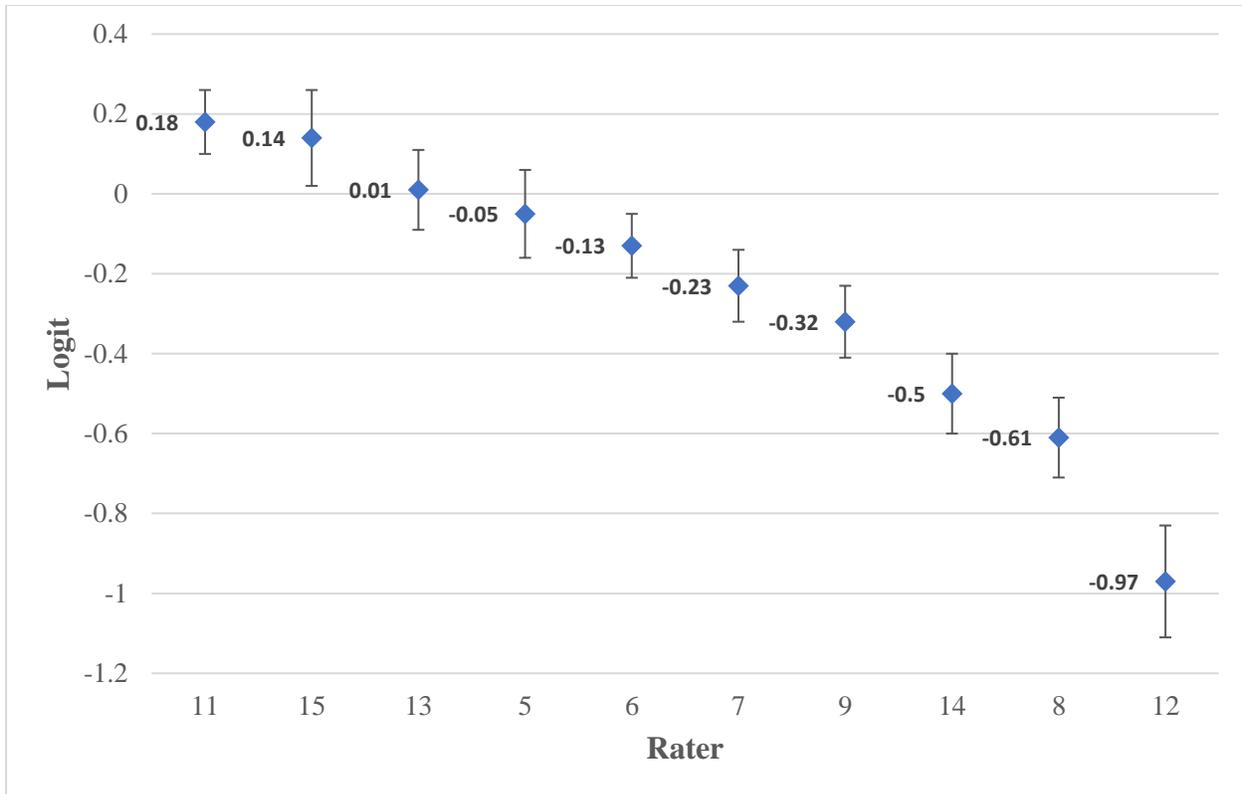


Figure 4.3. Rater severity estimates and 95% confidence intervals from MFRM analysis of 1,300 operational ratings

To decide if the difference in terms of severity among the raters was statistically significant, the Chi-square statistics, rater separation ratio, and reliability of rater separation index were examined. The Chi-square statistic for rater severity (or rater homogeneity index) was significant ($\chi^2_{(9)} = 100.3, p < .001$). It should be noted that the Chi-square statistic is highly sensitive to sample size (Myford & Wolfe, 2004). Thus, other statistics were also considered to examine rater severity variability. Specifically, rater separation ratio was 3.43, indicating that the variability of the rater severity measures was more than three times larger than the precision of those measures. These results indicate that the raters were statistically different in terms of their severity. These results were supported by the fact that the rater separation index is relatively high, at .92. Overall, the results show that the raters exercised a highly dissimilar degree of severity.

The above results show that overall, the raters were significantly different in their severity level. However, they did not specify where this difference was; that is, which rater or raters held responsibility for this difference. Thus, it is important to test for any significant differences between raters in terms of severity in rating using Wald statistics for each pair of raters. Table 4.3 summarizes the results from the statistical testing of differences between pairs of raters using Wald statistics. Complete results of the Wald statistics results are in Appendix G.

Table 4.3. *Wald Statistics for Pairs of Raters*

Rater	11	15	13	5	6	7	9	14	8	12
11 (n = 182)		n.s	n.s	n.s	2.74** (.3)	3.41*** (.38)	4.15*** (.45)	5.31*** (.61)	6.17*** (.70)	7.13*** (.92)
15 (n = 90)			n.s	n.s	n.s	2.47* (.33)	3.07** (.41)	4.10*** (.56)	4.80*** (.65)	6.02*** (.87)
13 (n = 110)				n.s	n.s	n.s	2.45* (.30)	3.61*** (.46)	4.38*** (.56)	5.70*** (.80)
5 (n = 100)					n.s	n.s	n.s	3.03** (.40)	3.77*** (.50)	5.17*** (.73)
6 (n = 156)						n.s	n.s	2.89** (.34)	3.75*** (.44)	5.21*** (.69)
7 (n = 144)							n.s	2.01* (.24)	2.83** (.34)	4.45*** (.59)
9 (n = 154)								n.s	2.16* (.25)	3.91*** (.51)
14 (n = 132)									n.s	2.73* (.37)
8 (n = 134)										2.09* (.28)

* significant at the .05 level
n.s: not significant

** significant at the .01 level
Effect size (Cohen's d) in parentheses.

*** significant at the .001 level

Similar to the Wright map and Figure 4.3, Table 4.3 shows that Rater 12 was significantly different from all other raters in terms of severity. The results show that this rater was significantly less lenient compared to Rater 11, Rater 15, and Rater 13 by the very large effect size of the difference between the severity of this rater and that of Rater 11, Rater 15, and Rater 13 (Cohen's $d \geq .80$). Rater 12 was also significantly more lenient than the remaining raters, although the difference size ranged from medium (Cohen's $d \geq .50$) to small (Cohen's $d \geq .20$). Raters 11, 15, 13, 5, 6, 7, and 9 were relatively comparable in terms of severity as indicated by the fact that the Wald statistics comparing their severity estimates were either non-significant or significant with a small effect size (i.e., Cohen's $d < .5$). Therefore, these raters could be grouped together in terms of their severity. Rater 14 and Rater 8, who were not different from each other in their severity level, significantly differed from the remaining of the raters. These two raters were significantly more generous than Raters 11, 15, 13, 5, 6, 7, and 9 and significantly less generous than Rater 12.

In sum, detailed comparison of pairs of raters revealed that while most raters were relatively similar in their rating severity, Raters 14, 8, and 12 were significantly more generous than the remaining raters. This difference was unacceptable as it was statistically significant with mostly medium to large effect sizes. A closer look of the profiles of these raters showed that they had different teaching and rating backgrounds. For example, Rater 8 had been rating for the EPT Writing for five test administrations and had been teaching ESL writing courses for three years. Rater 12 had rated for the EPT Writing for two test administrations but had never taught the ESL writing courses. Rater 14 was a new rater as this was his first time rating for the EPT although he had experience teaching the ESL Writing courses. Therefore, it seems that there was no clear pattern in these raters' profile that explained the difference in raters' severity. That is, no

conclusion could be made about the relationship between rating experience, teaching experience and raters' comparability.

Rater Consistency

Mean square infits and mean square outfit fits were examined for evidence of the rater's consistency in their severity. Table 4.4 summarizes the measurement results for the rater facet provided by *Facets*.

Table 4.4. *Measurement Results for the Rater Facet*

Rater	Total ratings	Severity estimates	SE	Mean-square infit	Standardized infit	Mean-square outfit	Standardized outfit
11	182	.18	.08	1.27	2.30	1.12	.90
15	90	.14	.12	.68	-2.30	.65	-2.30
13	110	.01	.10	.90	-.70	.97	-.10
5	100	-.05	.11	1.60	3.70	1.60	3.40
6	156	-.13	.08	.81	-1.70	.84	-1.30
7	144	-.23	.09	1.09	.80	1.08	.60
9	154	-.32	.09	.71	-2.60	.71	-2.50
14	132	-.50	.10	1.02	.20	1.04	.30
8	134	-.61	.10	.86	-1.10	.83	-1.20
12	98	-.97	.14	1.05	.30	.86	-.40

It can be seen from Table 4.4 that the fit mean square statistics for the majority of the raters fell within the acceptable range of between .7 and 1.3 (McNamara, 1996), indicating that with the exception of Rater 5, all raters maintained their personal level of severity. The infit

mean squares for Rater 11 and Rater 15 and the outfit mean squares for Rater 15 were at 1.27, .68, and .65 respectively, just outside, or with rounding, at the acceptable range of .7 – 1.3. Thus, these two raters were considered consistent in their severity. The infit mean square and outfit mean square statistics for Rater 5 were reported at 1.6, higher than the upper end of the acceptable range. This means that Rater 5 showed more variation than expected in their ratings, which could be a result of this rater's overuse of the extreme score bands (i.e., B and Pass). In fact, a closer examination of the operational ratings revealed that most of Rater 5's ratings were B or Pass, which accounted for 25% and 31% of the total ratings provided by this rater.

Overall, nine raters were found to maintain their consistency in terms of severity while one rater showed more variations in their ratings than expected. This result was not ideal; however, given the number of raters involved in the rating process and the stakes of the test, raters' severity consistency was considered acceptable. As Rater 5 was a new rater and had experience teaching 101B, the intermediate level writing course that focused on grammar and paragraph level writing, it is important to provide more training to this rater so that this rater can be more consistent.

Rater' Bias

To address research question 4 related to the potential bias displayed by raters towards the tasks (Summary vs. Essay), the interaction between two facets, *raters* and *tasks*, provided by the aforementioned MFRM analysis of the 1,300 operational ratings in *Facets*, was examined. Figure 4.4 shows the results of the Facets analysis displaying the interaction between raters and tasks.

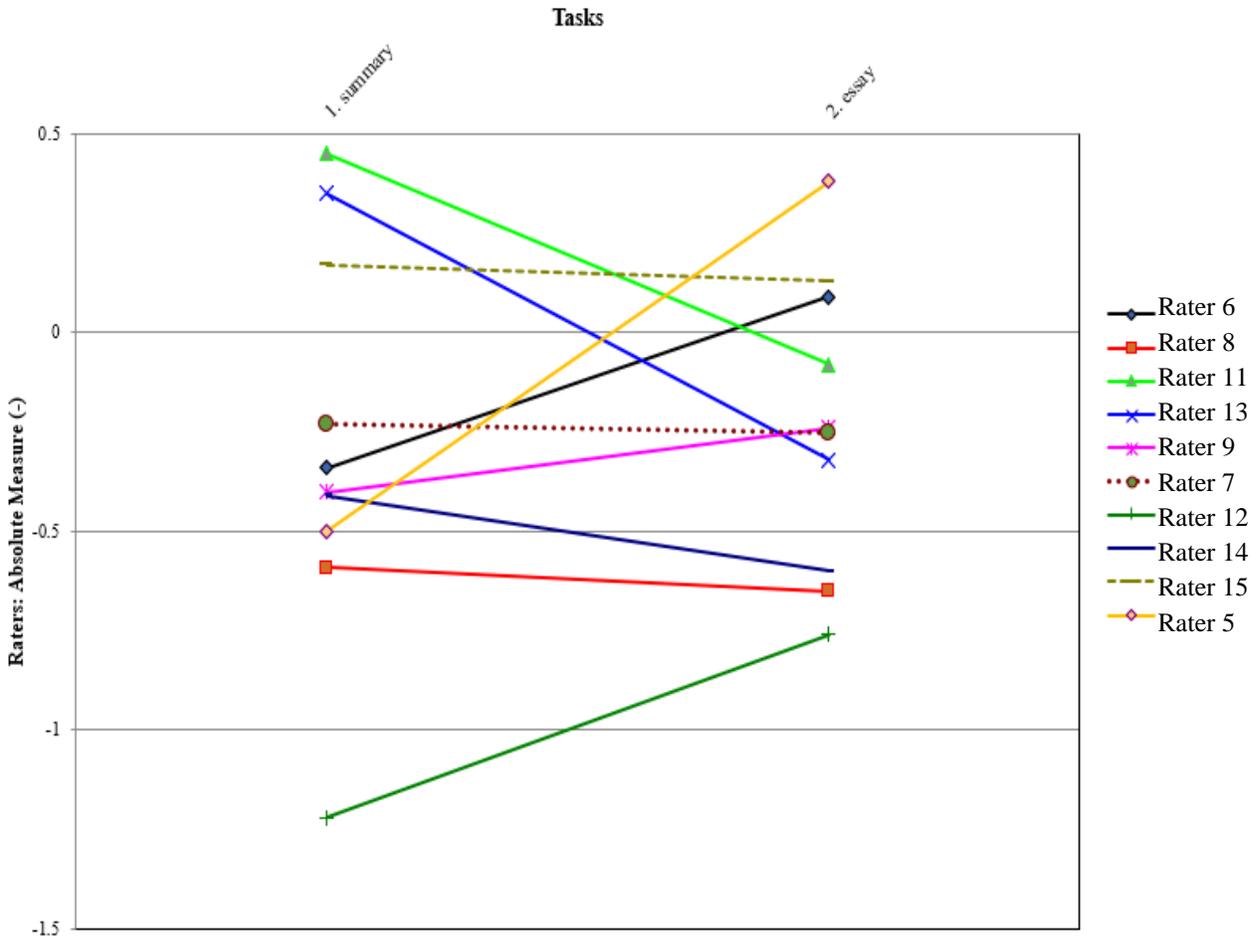


Figure 4.4. Interaction between raters and tasks from MFRM analysis of 1,300 operational ratings in Facets

While three raters, Rater 8, Rater 7, and Rater 15 remained relatively constant in their severity when rating the two tasks' responses, other raters seemed more or less generous depending on which task they were rating. Rater 11, Rater 13, and Rater 14 seemed more generous when rating essays, as indicated by their downward lines from "summary" to "essay". However, the opposite trend was observed in Rater 6, Rater 9, Rater 12, and Rater 5 as shown by their upward lines from "summary" to "essay".

Table 4.5 provides more details about the results of the bias analysis of the two facets, Raters and Tasks. It displays the raters' leniency measures for the two tasks, the bias size (i.e.,

effect size), and the paired t-test results to identify if the difference between the leniency measures across the two tasks was statistically significant.

Table 4.5. *Analysis of Raters' Bias towards Tasks from Facets*

Rater	Summary		Essay		Difference (Bias size)	Joint SE	t	d.f.	p- value
	Measure (logit)	SE	Measure (logit)	SE					
6	-.34	.12	.09	.12	-.43	.17	-2.53	147	.012
8	-.59	.13	-.65	.13	.06	.19	.29	125	>.05
11	.45	.12	-.08	.11	.53	.16	3.27	171	.001
13	.35	.14	-.32	.14	.67	.2	3.29	99	.001
9	-.40	.13	-.24	.13	-.16	.18	-.88	121	>.05
7	-.23	.13	-.25	.13	.02	.19	.12	129	>.05
12	-1.22	.21	-.76	.19	-.46	.28	-1.66	73	>.05
14	-.41	.14	-.6	.14	.19	.19	.96	123	>.05
15	.17	.16	.13	.16	.05	.23	.20	77	>.05
5	-.50	.16	.38	.15	-.88	.22	-4.03	89	<.001

Table 4.5 shows that Rater 6, Rater 11, Rater 13, and Rater 5 were statistically different in terms of their leniency when they rated the summaries and the essays. Specifically, Rater 11 and Rater 13 were significantly more generous with the essays, with significant t statistics at 3.37 ($p = .001$) and 3.29 ($p = .001$) respectively. The bias size for those raters were medium, at .53 and .67, respectively, indicating that they were .53 and .67 logits more lenient when they rated the essays. However, Rater 6 and Rater 5 followed the opposite pattern. The paired t-test for Rater 6 showed that the change in her leniency was statistically significant ($t_{(147)} = 2.53$, p

=.012). The bias size for this change was medium, at .43, indicating that Rater 6 was .43 logits more lenient with the summaries than with the essays. The most dramatic change in leniency level was displayed by Rater 5, whose change in leniency across the two tasks was statistically significant ($t_{(89)} = 4.03, p < .001$). The bias size for this rater was also the largest, at .88, showing that this rater was 0.88 logits more lenient with the summaries.

Overall, the results from the bias analysis showed that while the majority of the raters did not display any bias against a specific task, some of them did behave differently based on the task type. Rater 11 and Rater 13 were significantly more generous with the essays (Task 2) than with the summaries (Task 1). On the opposite trend of Rater 11 and Rater 13 were Rater 5 and Rater 6, who were significantly more generous with the summaries (Task 1). The fact that the bias sizes for Raters 5, 11, and 13 were above the medium level of .5 is concerning because it negatively affects the reliability of the ratings provided by these raters. An examination of the profiles of the four problematic raters did not reveal any common characteristics in terms of their rating and teaching experience. Rater 11 and Rater 13, who were more generous with the essays (Task 2), had different EPT Writing rating and ESL course teaching experience. Specifically, Rater 11 was an experienced EPT Writing rater and ESL writing instructor. Rater 13, however, was new to the EPT rating and ESL writing courses. Similar to raters 11 and 13, Rater 5 and Rater 6, who were more generous with the summaries (Task 1), also had different profiles. In fact, Rater 6 was an experienced EPT Writing rater and ESL writing instructor while Rater 5 was considered a novice rater although he had taught ESL writing courses prior to the time to rating. This finding confirmed the conclusion made earlier that there was no clear pattern in these raters' background that explained their behaviors. Nevertheless, more rater training is needed to ensure that raters were not biased for or against a particular test task.

Score Reliability with Two Tasks x Two Raters Design

To address RQ5 regarding score reliability when two tasks were included and two raters were employed to rate each response, a univariate G-study was conducted using the *gtheory* package in *R* (Moore, 2016) with the fully-crossed, two-facet design ($p \times t \times r$) with 42 examinees, two tasks, and four raters. It should be noted that the 366 ratings used for this analysis were part of the 1,300 operational ratings in the MFRM analysis presented earlier. After the G-study was conducted, different combinations of tasks and raters (including the design of two tasks and two raters of particular interest) were examined to find an optimal measurement design for achieving a desirable level of score dependability of .7.

Table 4.6 displays the descriptive statistics for the ratings provided by the four raters to Task 1 and Task 2 responses from 42 test takers. The four raters assigned similar ratings on Task 2, with their mean ratings ranging from 3.26 to 3.48. Their ratings for Task 1 were less similar, with the mean ratings between 2.93 and 3.45. The average mean of the ratings for both tasks were relatively similar for the four raters, ranging from 3.10 to 3.46.

Table 4.6. *Descriptive Statistics for the Ratings Used in the G-study Provided by the Four Raters to 42 Test Takers*

Rater	Task 1		Task 2		Both Task
	Mean	SD	Mean	SD	Mean
7	2.98	1.51	3.38	1.29	3.18
8	3.45	1.37	3.48	1.21	3.46
11	2.93	1.66	3.26	1.62	3.10
14	3.45	1.27	3.43	1.33	3.44

Table 4.7 presents the variance component estimates from the G-study for the baseline situation where only one task and one rater are involved.

Table 4.7. *Univariate G-Study Results for One Task and One Rater*

Source of variation	df	Sums of squares	Mean of squares	Variance estimate	% of total variance
Person (p)	41	339.2	8.27	.720	35.5
Task (t)	1	2.9	2.86	.005	.2
Rater (r)	3	8.7	2.89	.012	.6
pt	41	67.5	1.65	.265	13.0
pr	123	178.4	1.45	.432	21.3
rt	3	2.9	.98	.009	.5
ptr, error(e)	123	72.2	.59	.587	28.9

As displayed in Table 4.7, the examinee (person) main effect variance explained a much larger proportion of the dimensional score variance (35.5%) than the variances of other main

effects or interactions. This suggests that the test takers' source-based academic writing ability accounted for over a third of the variance in the composite scores. This examinee variance becomes a universe (true) score variance later in the D-study. The main effect for task was very small, at .2%, indicating that there was little difference in task difficulty across the writing tasks. Similarly, the rater main effect variances accounted for a very small proportion of the dimensional score variance, at .6%, showing that there was little difference in terms of severity among raters between the two tasks. The effect of the interaction between task and rater (rt) also explained a small proportion of the composite score variance, at 5%, suggesting that the relative rater severity changed very minimally across different tasks. However, errors account for a large proportion of the dimensional score variance (28.9%). This indicates that the effects of three-way interaction between examinee, rater, and task compounded with other unmodeled errors on the observed scores were larger compared to the effects of other variance components, with an exception for the person variance. This result suggests that the relative ranking of examinees changed considerably across different task-by-rater interaction. Additionally, the interaction between examinee and rater (pr) accounted for a third highest proportion of the dimensional score variance (21.3%). This result shows that the raters did not maintain the same severity level for every test taker, and this interaction had a large effect on the composite score. Thus, more rigorous rater training is necessary. Also, the interaction between examinee and task (pt) explained 13% of the total variance, which is a relatively large effect. This means that the test takers' performance was affected by the tasks administered. To cancel out the person-by-item effects, one solution could be to increase the number of tasks in the test. However, this remedy should take into account other practicality issues for test administration, scoring, and score reporting.

Based on the variance component estimates in Table 4.7, a series of D-studies were conducted. First, the D-study for two tasks and two raters was carried out as this was the design adopted by the EPT Office at the time of the study. It should be noted that the number of tasks was not increased beyond three because it would be impractical to administer the EPT Writing with more than three tasks. Table 4.8 shows the variance component estimates and dependability indices (Φ) in various conditions where the number of tasks (N_{task}) and the number of raters (N_{rater}) varied.

Table 4.8. *D-Study Results for Various Numbers of Tasks and Raters*

N_{task}	1	1	1	1	2	2	2	2	3	3	3	3	
N_{rater}	1	2	3	4	1	2	3	4	1	2	3	4	
Variance estimate (% of total variance)	p	.720 (35.5)	.720 (47.5)	.720 (53.5)	.720 (57.2)	.720 (45.0)	.720 (58.3)	.720 (64.7)	.720 (68.5)	.720 (49.5)	.720 (63.1)	.720 (69.6)	.720 (73.2)
	t	.005 (.2)	.005 (.3)	.005 (.4)	.005 (.4)	.005 (.3)	.005 (.4)	.005 (.4)	.005 (.5)	.005 (.3)	.005 (.4)	.005 (.5)	.005 (.5)
	r	.012 (.6)	.012 (.8)	.012 (.9)	.012 (1.0)	.012 (.8)	.012 (1.0)	.012 (1.1)	.012 (1.1)	.012 (.8)	.012 (1.1)	.012 (1.2)	.012 (1.2)
	pt	.265 (13.1)	.265 (17.5)	.265 (19.7)	.265 (21.0)	.132 (8.3)	.132 (1.7)	.132 (11.9)	.132 (12.6)	.088 (6.0)	.088 (7.7)	.088 (8.5)	.088 (9.0)
	pr	.432 (21.3)	.216 (14.2)	.144 (1.7)	.108 (8.6)	.432 (27.0)	.216 (17.5)	.144 (12.9)	.108 (1.3)	.432 (29.7)	.216 (18.9)	.144 (13.9)	.108 (11.0)
	rt	.009 (.4)	.005 (.3)	.003 (.2)	.002 (.2)	.005 (.3)	.002 (.2)	.002 (.2)	.001 (.1)	.003 (.2)	.002 (.2)	.001 (.1)	.001 (.1)
	prt	.587 (28.9)	.293 (19.3)	.196 (14.6)	.147 (11.7)	.293 (18.3)	.147 (11.9)	.098 (8.8)	.073 (6.9)	.196 (13.5)	.098 (8.6)	.065 (6.3)	.049 (5.0)
	Φ	.35	.47	.54	.57	.45	.58	.65	.68	.49	.63	.7	.73

Note: The numbers in parentheses refer to the % of total variance explained.

It is unsurprising to see from Table 4.8 that the index of dependability increased as more tasks and raters were included in the rating design. Dependability was the lowest for the one task – one rater scenario (.35) and highest when three tasks and four raters were employed (.73). Of particular interest to this study is the index of dependability in the design with two tasks and two raters. In this scenario, dependability was relatively low, at .58. If one more rater was recruited to rate the responses, the dependability index would increase to .65. In cases where three tasks were used to elicit test takers' written performance, dependability would increase from .63 to the desirable level of .70 when responses were rated by two and three raters, respectively. Figure 4.5 provides a graphical demonstration of the changes in dependability resulted from varying the numbers of tasks and raters. It is important to acknowledge that dependability indices displayed in this figure were all estimates based on the particular sample used in this study and thus, each estimate represents one point within a range of possible results.

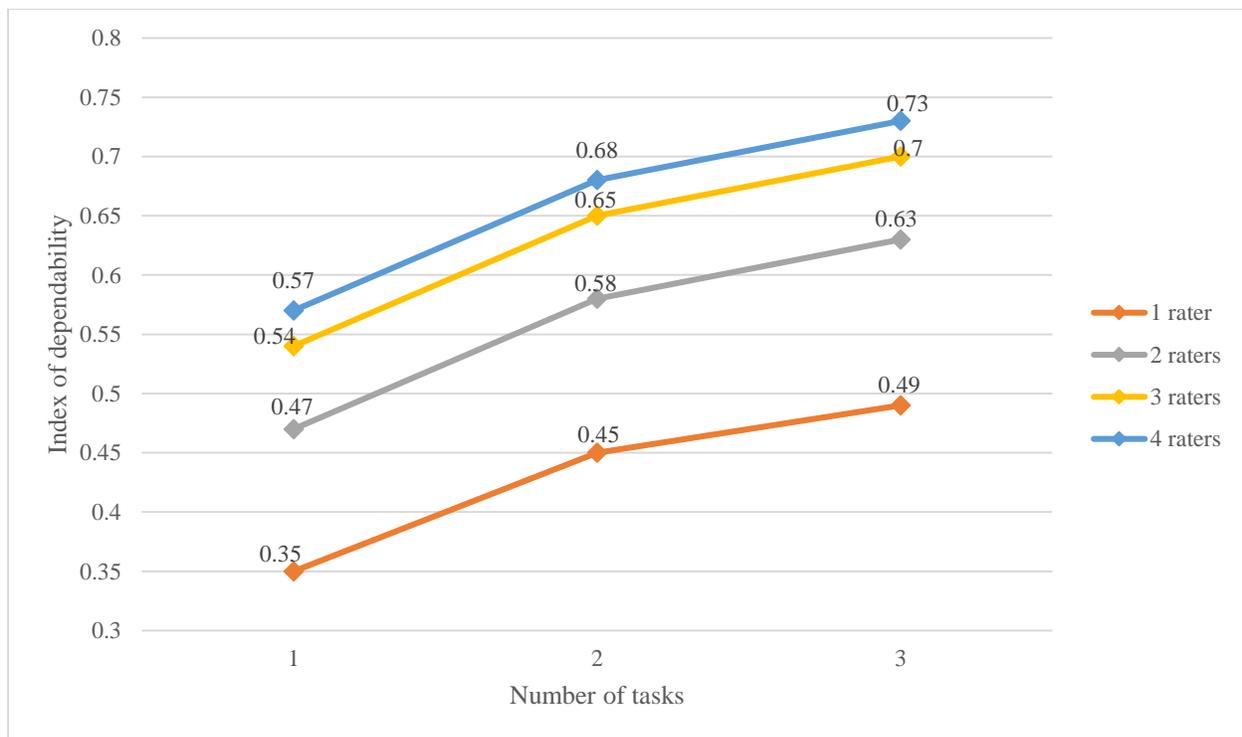


Figure 4.5. Dependability change relative to the number of tasks and raters

In sum, the design of two tasks x two raters at the time of the study did not result in an acceptable level of score reliability, as the dependability index for this condition was at .58. In order to reach the desirable reliability of at least .7, at least three tasks would need to be used and three raters would need to be employed to grade test takers' responses.

Raters' Attention to Writing Features

Research question 6 regarding decision-making processes during the rating of integrated writing task responses was addressed by an analysis of ninety verbal reports collected from the nine raters who participated in the eye-tracking rating session. However, since this rating session was done on the computer, which was different from the usual paper-based rating, MFRM was conducted in *Facets* on 630 experimental ratings of the Essay task (Task 2) responses to identify if the rating medium (computer-based vs. paper-based) was a source of rater bias. This section starts with the relationship between rating medium and rater severity by examining the bias analysis results from *Facets*. After that, it focuses mainly on the results of the qualitative analysis of the raters' verbal reports.

Rating Medium and Rater Severity

The 630 ratings given by the nine raters to 60 essays were analyzed in *Facets* with the three facets (i.e., Examinees, Raters, Medium) and the interaction terms between Raters and Medium. The assumptions of unidimensionality, local independence, and certainty of responses were first examined. Since the task facet only included one task, the unidimensionality assumption was reasonably met. The local independence assumption was also met because of two reasons: the test takers were not allowed to work together, and the raters rated the test responses independently. The assumption of certainty of responses was met because these responses had

been written by real EPT Writing test takers, who were motivated to perform well on the test to avoid the ESL Writing courses. Additionally, the log-likelihood chi-square statistic was reported at 615.12 ($df = 578, p > .05$), showing that the data fit the model.

The interaction between the raters and rating medium are illustrated in Figure 4.6. As can be seen from this figure, the change in rater severity was most pronounced in Rater 1 and Rater 9. While Rater 1 was likely to be more severe when rating on the computer, as shown by the upward line from “paper” to “computer”, Rater 9 seemed more generous when rating on this medium, as displayed by the downward line from “paper” to “computer”.

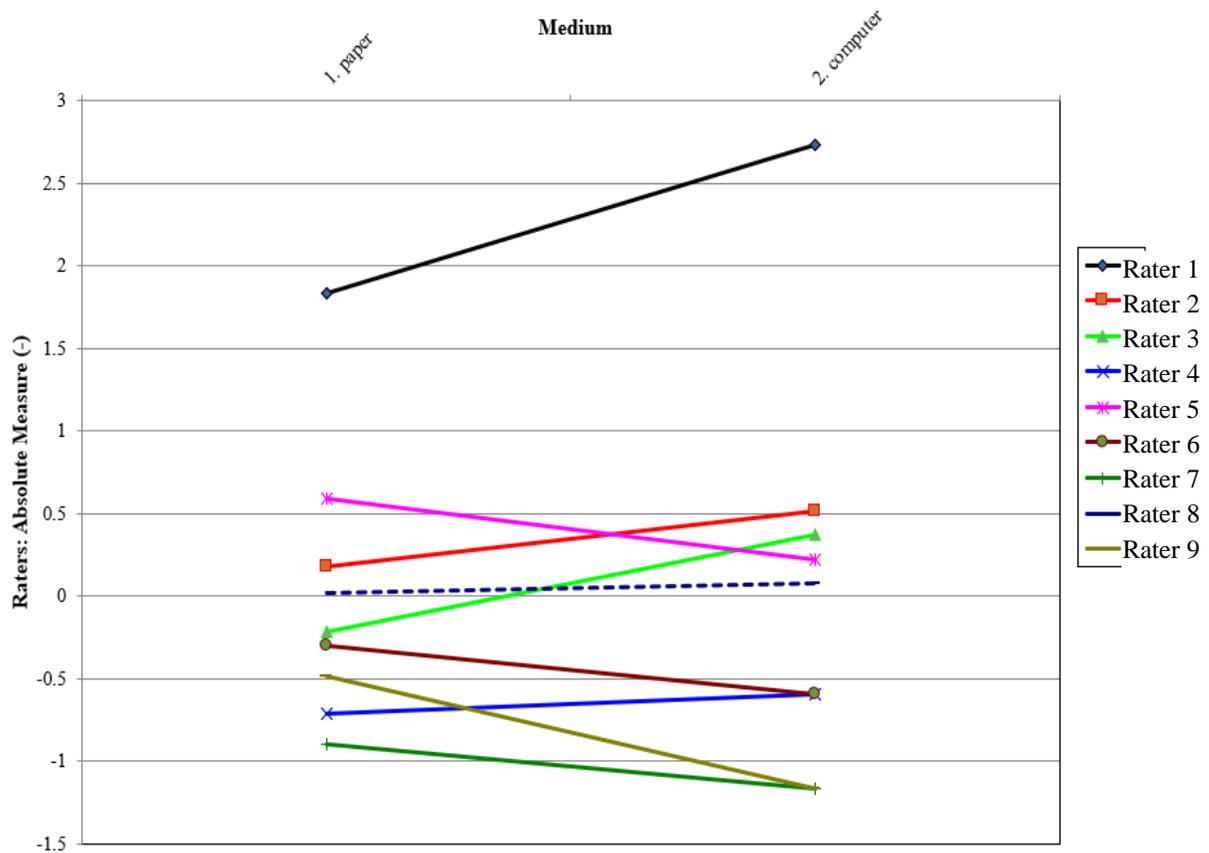


Figure 4.6. Interaction between raters and rating medium from MFRM analysis of 630 experimental ratings in Facets

However, a closer examination of the bias analysis summary from *Facets* (Table 4.9) revealed that the paired t-tests comparing rater severity across the two rating modes were not statistically significant. This result indicated that the rater severity did not depend on whether they rated on paper or on the computer. This means that ratings resulting from the eye-tracking sessions were similar to rating on paper in operational testing.

Table 4.9. *Analysis of Raters' Bias against Rating Medium from Facets*

Rater	Paper		Computer		Difference (Bias size)	Joint SE	t	d.f.
	Measure	SE	Measure	SE				
1	1.83	.19	2.73	.59	-.90	.62	-1.46	10
2	.18	.16	.51	.39	-.34	.42	-.8	12
3	-.22	.16	.37	.38	-.59	.42	-1.41	12
4	-.71	.17	-.60	.37	-.11	.40	-.27	12
5	.59	.17	.22	.38	.36	.41	.88	12
6	-.30	.16	-.60	.37	.30	.40	.74	12
7	-.90	.17	-1.17	.39	.26	.42	.62	12
8	.02	.16	.08	.37	-.06	.41	-.15	12
9	-.49	.16	-1.17	.39	.68	.42	1.61	12

* All *p*-values are greater than .05.

Writing Features Attended to by Raters

Ninety transcripts of verbal reports collected from the nine rater participants on the ten responses were analyzed to address research question 6 on the writing aspects attended by raters during their evaluation of the source-based task responses. The 90 transcripts were segmented,

resulting in a total of 3,140 content units. However, only 1,483 of these content units were relevant to address the research question. Table 4.10 shows the distribution of the content units across the nine raters and the ten essays.

Table 4.10. *Distribution of Content Units (N = 1,483) across Nine Raters and 10 Essays*

Rater \ Essay	1	2	3	4	5	6	7	8	9	10	Total
1	18	20	21	25	20	18	23	28	18	27	218
2	15	8	23	15	10	11	7	13	17	9	128
3	14	8	8	12	7	16	16	10	16	7	114
4	17	32	28	25	23	22	18	17	32	25	239
5	9	10	15	16	22	11	17	15	11	10	136
6	8	21	16	18	18	18	20	19	16	20	174
7	5	14	17	19	13	21	14	13	10	14	140
8	26	20	32	18	20	27	18	25	15	17	218
9	7	15	15	15	10	11	13	12	5	13	116

These units were then coded using the coding scheme (Appendix F) that was developed as described in the previous chapter. Therefore, each content unit was coded as a writing feature that the participant reported attending to during rating. Table 4.11 provides information on the frequency, proportion, along with the descriptive statistics of the writing features attended to by the raters, which were categorized into five groups based on my judgment, namely, *Arguments and Organization*, *Grammar and Lexis*, *Style and Conventions*, *Discourse Synthesis*, and *Others*.

Table 4.11. *Writing Features Attended to by Nine Raters When Rating 10 Essays*

Category	Essay features	Frequency	%	M	SD
Grammar and Lexis	Comprehensibility	55 (9)	3.7	.61	.80
	Lexis	94 (9)	6.3	1.04	1.41
	Grammatical complexity and accuracy	252 (9)	17.0	2.80	2.88
	Overall language	52 (9)	3.5	.58	.94
	Spelling & punctuation	68 (9)	4.6	.76	1.46
	Total	522	35.5		
Arguments and Organization	Reasoning/topic development	175 (9)	11.8	1.94	1.78
	Relevance	45 (8)	3.0	.50	.99
	Coherence	38 (9)	2.6	.42	.73
	Use of transitions	66 (8)	4.5	.73	1.13
	Text organization	74 (9)	5.0	.82	1.15
	Writer stance	47 (7)	3.2	.52	.95
	Ideas/rhetoric	58 (9)	3.9	.64	1.03
Total	503	33.9			
Discourse Synthesis	Source text understanding	101 (6)	6.8	1.12	1.91
	Citation quality	116 (9)	7.8	1.29	1.83
	Total	217	14.6		
Style and Conventions	Style/register/discourse functions/genre	28 (6)	1.9	.31	.73
	Source use convention	157 (9)	10.6	1.74	1.94
	Total	185	12.5		
Others	Task completion	37 (9)	2.5	.41	.69
	Quantity of written production	20 (6)	1.3	.22	.49
	Total	57	3.8		

Note: The numbers in parentheses indicate the numbers of raters who reported a specific feature.

Overall, the raters reported the most on *Grammar and Lexis*, as they gave a total of 522 (35.5%) comments relevant to this feature. *Arguments and Organization* also attracted a lot of attention from raters, with 503 comments in total, accounting for 33.9% of the total comment units. Also, 217 comments pertained to *Discourse synthesis*, making up 14.6% of the total comments. *Style and Conventions* received some attention from the raters, with a total of 185 comments (12.5%). *Others* attracted the least attention from raters, with 57 comments (3.8%).

Grammar and lexis

A closer examination of the individual essay features shows that the most frequently mentioned feature was assessing *grammatical complexity and accuracy*. In fact, 252 comment units (17%) revealed that raters focused on the complexity and accuracy of grammatical structures when evaluating test takers' responses. All raters reported that they focused on grammar to some extent. For example, when recalling his rating process on an essay, Rater 8 reported the following:

I'm lingering there toward the beginning probably because of a verb issue. So, I remember this sentence. "has been affected" is not the correct structure there... So, it shouldn't be, "GM has been affected". They're using the passive there but that's not what they should be saying. (Rater 8, Verbal Reports, 09/29/2018)

Similar to Rater 8, Rater 6 also recalled assessing test takers' ability to use accurate grammatical structures, as in her comment below.

Then, it starts with, "There are many people nowadays." And "needn't" is not correct here. So, it seems like this student has grammar problems. So, that was the first thing that I noticed. (Rater 6, Verbal Reports, 10/26/2018)

In addition to assessing students' grammatical accuracy, raters also considered the complexity of their grammatical structures, as Rater 4 indicated below.

... something I look for is the subordination, things like that. If their clause structure is more complex because they're using relatively simple syntax, if they're using syntax and they make mistakes, it makes me think they're using simple syntax and its correct, but it means another thing, so I'm kind of comparing, seeing what kind of structure there is.
(Rater 4, Verbal Reports, 10/18/2018)

Other features of grammar and lexis that raters were attuned to included *lexis*, which accounted for 6.3% of the total comment units, as Rater 1 said below.

I do recognize in their first paragraph, they use the same vocabulary over and over again like, "Should not be supported", "should be supported", "should be supported". I think I went to the text ... I found the "controlled by biotech companies", a phrase, and then I was like, "Eh, lack of vocabulary". (Rater 1, Verbal Reports, 02/02/2018)

For this rater, test takers' ability to vary their vocabulary was a factor to consider when making her decision about their performance. This comment was echoed by Rater 3, who also commented the test taker's incorrect word use.

Well, interesting word choice there, "should be highly supported and be widespread on earth." I often find is very helpful when they answer the prompt in their thesis statement very specifically by using similar language of the prompt, or if they had used a synonym that was very similar to "support". (Rater 3, Verbal Reports, 02/03/2018)

Another feature in the essays that the raters concentrated on was *spelling and punctuation*, although this feature received only 68 comments (4.6%). For example, Rater 7 revealed that misspelling, although it did not impede his ability to understand the writer's message, was relatively distracting, as follows.

... and "spelling errors are minor; not interfering" and then I was thinking, "Did they interfere?" Not much to be honest. I could still understand the text and everything. But there were again, a lot of spelling errors, like missing letters. For most of them I could understand the words still, although the letters are like mixed up, I could understand it; So, it didn't really affect my understanding of the text. So, I was like going back, reading again trying to understand. They did distract me. (Rater 7, Verbal Reports, 02/01/2018)

In addition to *grammatical complexity and accuracy, lexis, and spelling and punctuation*, the raters also assessed *comprehensibility* (3.7%), as Rater 2 commented below.

I think I ended up reading this a few times because of lack of clarity. I went back and forth a few times because I found this to be unclear... And as I read this one, I'm like, "this is reading really well." And so, I didn't have any areas where I'm like, "I don't understand this." (Rater 2, Verbal Reports, 02/03/2018)

The raters also commented on test takers' *overall language* (3.5%), as Rater 2 and Rater 9 said below.

And so, I was looking at... Okay. Do we have... are we struggling with command of English? Are we struggling with language? The language is not bad. (Rater 2, Verbal Reports, 02/03/2018)

So, I read through the first sentence in the first paragraph. I felt like it was generally well written. (Rater 9, Verbal Reports, 02/08/2018)

Arguments and organization

The second category that the raters attended to when rating the responses was *Arguments and Organization* (33.9%). In this category, *reasoning and topic development* received the most attention, with a total of 175 comments (11.8%), placing it the second most attended feature overall. For example, Rater 3 assessed test takers' development of their argument while she was reading the essay introduction, as shown below.

When I think of an introductory paragraph, I think of that upside-down triangle. Give us the context, move to the thesis. I could see that the writer was trying to give us that background. But then as a reader, I was thinking, "Okay, so what? What is your point? Where are you going with this?" (Rater 3, Verbal Reports, 02/03/2018)

In fact, one of the first things that the raters did when they read the introduction was to identify the stance point of the writer by locating the thesis statement was, as reported by Rater 7.

In that part, I was like, okay, that's the last part of the first paragraph. Therefore, I think this sentence is not the thesis statement. The last sentence here should be the thesis statement, because that's the last sentence of that first paragraph. I was still trying to identify the thesis statement. So, I felt like, okay, the previous part was kind of providing background; so, this sentence is kind of support to the thesis. So, this should be a kind of good thesis statement of an opinion paragraph. I said, okay, that's the thesis, kind of. (Rater 7, Verbal Reports, 02/01/2018)

The raters also focused on whether a claim that test takers made in their essay was well supported by examples or other details, as illustrated in Rater 5's recall as follows.

I thought that, "You have freedom to make decision your health based on your hand and for your future generations". So, this sentence sounds good, but I think there should be more sentences before this to substantiate this nice claim. (Rater 5, Verbal Reports, 09/06/2018)

Additionally, the raters also considered if the logic of test takers' argument, as indicated in Rater 4's comments below.

As I finished it, I thought, "Does this make sense logically?" Because they're saying, "These crops are resistant to herbicides and so GM crops..." Then they say, "That it will be a problem because they use more herbicide to kill those crops." But why would they want to kill these crops? They're food crops. So, this person is saying that they'll need to use more chemicals so they will hurt the environment which that's a bad thing. They're not trying to kill these crops; so; I thought, "Why? Am I understanding that right? Why would they be saying that?" (Rater 4, Verbal Reports, 10/18/2018)

In addition to *reasoning and topic development*, the raters attended to other features such as *text organization, use of transitions, ideas and rhetoric, writer stance, relevance, and coherence*, although the number of comment units for these features fell drastically. *Text organization* received 74 comments from the raters (5%). In fact, the raters' verbal reports indicated that raters paid attention to how test takers organized paragraphs in their essays, as shown in the quote from Rater 1.

So, I feel like this should still be in the third paragraph here, where they're talking about golden rice. But they're starting their conclusion paragraph with it. So not good in terms of organization. (Rater 1, Verbal Reports, 02/02/2018)

Similar to Rater 1, Rater 9 focused on whether test takers organized their essay in paragraphs, as he reported below.

And then when I went through the third or fifth line of the first paragraph, I noticed that this guy doesn't have any paragraph division. It's all written in one single paragraph.
(Rater 9, Verbal Reports, 02/08/2018)

Use of transitions was another feature that attracted attention from the raters, as it received 66 comments, accounting for 4.5% of the total comment units. For example, Rater 8 reported that he was confused because of the incorrect transitional word, “that is to say”, used in test takers’ response.

This is not a correct logical connector. “... and GM technology is much cheaper because they grow faster or produce more per m2”. What is that? I was really confused over that one. ‘That is to say’? Okay, this is not a logical connector. (Rater 8, Verbal Reports, 09/29/2018)

Additionally, the raters also assessed test takers’ *ideas and rhetoric*. In fact, they made 58 comments (3.9%) on the complexity and quality of the ideas put forward in the essays. For instance, Rater 4 commented that one of the arguments that the test taker made in response about the relationship between price and quality did not seem to be a good one, as indicated below.

So, they did mention twice that cheap does not necessarily mean good or excellent thing. They say that twice. But as I read that I thought, "well, nobody thinks cheap means quality." I don't know if it's that good of a point because nobody thinks cheap means excellent. I don't know if it's a great point to make, and they do refer to it twice. (Rater 4, Verbal Reports, 10/18/2018)

Another feature that the raters kept in mind while assessing test takers' responses was *writer stance*, which received 47 comments (3.2%) from seven of the raters. As shown in Rater 6's quote, she expected test takers to express their opinion in the opening paragraph so that she knew what to expect from the remaining essay.

He did a good job, but I don't think it's clear that he agrees or disagrees. I wanted to see if again he or she clearly stated what points he's going to talk about in body paragraphs. (Rater 6, Verbal Reports, 10/26/2018)

Rater 1 also stated that test takers need to clearly voice their own perspective on the topic of writing instead of summarizing the main points in the source text, as shown below.

And then, in general, like the second paragraph, it's just all summary. What about your opinion? There's a little bit at the end of this. There's again an attempt, but it wasn't enough. (Rater 1, Verbal Reports, 02/02/2018)

Relevance and coherence were another two features that the raters assessed. Specifically, they made 45 comments (3%) on the relevance of test takers' ideas. For example, Rater 8 commented on the relevance of the test taker's personal experience to the whole argument in the essay, as he said, *"It's about the examinee, his experience again. Not really relevant here. I think*

that sounds a little bit redundant.” The raters also made 38 comments (2.6%) on the coherence of test takers’ arguments, assessing whether “*the ideas flew well together*” (Rater 2, Verbal Reports, 02/03/2018) or the ideas were “*well-connected*” (Rater 9, Verbal Reports, 02/08/2018).

Discourse synthesis

Discourse synthesis was the third category of writing features that the raters focused on when making decisions about test takers’ performance. This category consisted of two writing features, namely *citation quality* and *source text understanding*. Specifically, this category received 217 comments, explaining 14.6% of the total comments.

Citation quality

The raters assessed *citation quality*, as this feature received 116 comments (7.8%). Specifically, the raters focused on the role of the source-based information in the whole argument that test takers were making. For example, Rater 1 evaluated the purpose of source citation, as she said.

The citation of sources just didn't really make sense. They're citing them, but I don't see the purpose of it. (Rater 1, Verbal Reports, 02/02/2018)

For this rater, test takers must elucidate their purpose for citing external sources. Similarly, Rater 4 considered whether test takers analyzed or commented on the sourced-based information incorporated in their essays, as shown below.

And when they're given these numbers. When I see a lot of lists of numbers and facts in the essays, that makes me question how much they are synthesizing the information or are they just making a list. So, I think that's why I looked at that carefully to make sure it

was ... They're doing some kind of analysis or synthesis of it. (Rater 4, Verbal Reports, 10/18/2018)

Rater 9 had a similar belief about the role of source-based information in test takers' argument. For this rater, test takers must give their own thoughts and comments as it was insufficient to use the information from the source texts without making it apparent how it contributed to the whole message of the essay, as seen below.

And he's citing from the text, but I didn't see any idea that comes from the student himself or herself. They're all just paraphrased versions of the text. There's no argument of their own there. So, all this information there is coming from texts. So, what is the purpose of putting it, if you don't relate it to anything? So, I didn't like the third paragraph, because he's totally summarizing the source texts. (Rater 9, Verbal Reports, 02/08/2018)

In fact, source citation without comments or explanation from the test takers were considered inefficient source citation, as Rater 8 commented below.

And so, this is about the second and third paragraphs about the description of the two source materials. No arguments from the examinee, okay? So, this is not effective source citation. (Rater 8, Verbal Reports, 09/29/2018)

However, according to the raters, even though test takers analyzed or commented on the source-based information, they must ascertain that the information taken from the source texts fitted the logical flow of the argument. For example, Rater 8 commented on the fact that the source citation contrasted with the test taker's argument, as seen below.

And I followed the materials cited here. I think that's the second one, the newspapers by Chivers. "And this will make many people suffer". I think basically the source materials are somehow incorporated to support the arguments, but somehow, it's really contrary to the thesis statements, right? (Rater 8, Verbal Reports, 09/29/2018)

The raters also focused on the relevance of the source-based information cited in test takers' response and advised test takers to focus on overall ideas instead of details, as Rater 4 reported below.

In the second paragraph, I think I thought, "oh this person is integrating source text". Yeah. I think that's what I thought. But this caught my attention because of the citing the source. I thought this information sounds a bit redundant, like "50 grams of the golden rice or for 60 percent of the daily vitamin A requirement". And maybe we can write something more general instead of getting in this detail. (Rater 4, Verbal Reports, 10/18/2018)

Additionally, the raters paid attention to the balance of the source-based information in test takers' response. That is, they took notes of whether test takers integrated information for both source texts. As the example below shows, Rater 4 expected test takers not to rely heavily on one source text.

... They seem to be analyzing it and that's why I think I stopped there to think about if it was the same original text. And if not, then this is a good sign I think because they're extrapolating from the original information possibilities. It could happen... Because it seems like they're relying more heavily on one than the other. I think they could have said more judging from their other source text. (Rater 4, Verbal Reports, 10/18/2018)

Rater 8 second Rater 4's opinion that test takers must make use of both source texts for their response, as he reported the following:

I just wanted to see whether the examinees selected the right material in the right place because I need evidence. That's why I was looking for the information upon golden rice. I couldn't see it here. So, and I thought to myself, "okay, where is the other source material?" I don't see any evidence of the whether the second source material was incorporated in examinees essays or not. (Rater 8, Verbal Reports, 09/29/2018)

Citation quality also referred to whether the source texts were skillfully integrated into test takers' response. For example, Rater 5 focused on whether test takers used more complex expressions to introduce the source-based information in their response.

As far as I remember, the [essay] just said "the second article said" or something like that in the beginning of the second paragraph. I thought that was not skillful even though the sources are there. (Rater 5, Verbal Reports, 09/06/2018)

Rater 1 made a similar point when she reported the following:

Now this got me thinking, which news reporter? I know that it's talking about the second text. But they're not citing appropriately because both of them are news reporters, basically like journalists. But just saying news reporter is not enough to show like good citation skills. (Rater 1, Verbal Reports, 02/02/2018)

In fact, whether the raters could distinguish identify which source texts the information originated from was one important factor that they considered when making decisions about test takers' grade, as shown in Rater 8's comment below.

In the last paragraph, I see some things actually from the original source are incorporated into the last paragraph. So “based on the text”. We have two texts. So, I’m not sure which one is referred to as in here. (Rater 8, Verbal Reports, 09/29/2018)

Source text understanding

The analysis of the verbal reports also showed that 101 comments (6.8%) from most raters pertained to *source text understanding*, which included whether test takers could understand the information from the external sources and thus, cited it accurately in their essay. For instance, Rater 2 reported that he was trying to compare the information cited in an essay and the source texts and paid attention to the fact that this test taker misunderstood the source texts.

Again, I was trying to bridge the gap of that argument discrepancy that I was reading in the essay. The essay had written that governments should spend money on fortification programs than on modifying crops. That’s where I was like, “okay, but it says here [in the source] that it’s actually more expensive to do the fortification programs than to modify the crops.” From an efficiency standpoint, it would seem that actually maybe governments should spend more money on modifying crops. So, this person actually misunderstood the source text. (Rater 2, Verbal Reports, 02/03/2018)

Similar to Rater 2, Rater 5 reported that he evaluated the accuracy of the source-based information cited in test takers’ essay, as he said:

I pay attention to these particular words “go blind each year”, “vitamin A deficiency” ... I guess I am evaluating the accuracy of information in the first sentence of the third paragraph. (Rater 5, Verbal Reports, 09/06/2018)

This feature was also attended to by Rater 4, who recalled the following:

So I think at this point of reading “reduces the probabilities of having independent citizens”, I think I may have slowed down on that just because we're kind of getting away from the original ideas in the text, but that's not necessarily a problem. It's just noticeable. (Rater 4, Verbal Reports, 10/18/2018)

Style and conventions

Style and Conventions consisted of two writing features, *style, register, discourse functions, and genre* and *source use convention*. *Style, register, discourse functions, and genre* attracted the received the least comments from the raters, at 28 comments (1.9%). In fact, Rater 1, Rater 5, and Rater 7 did not attend to this feature at all. Rater 9, however, attended to this feature when he observed test takers’ inappropriate use of register-related linguistic features as he mentioned the following in his verbal reports.

And I found that the expression at the end of the first sentence, “that teaches but to an extent,” and then I thought that, “okay, he or she's writing the way he speaks, he or she speaks.” So, I thought, “it's not that good. That's not a good sign.” That's what I thought. (Rater 9, Verbal Reports, 02/08/2018)

This feature was also considered by Rater 4, who commented on test takers’ failure to use of academic language, as shown below.

The thing is this is not what I would consider a very academic way to say that. So, the student doesn't have a strong control of academic language... This suddenly sounds like

a personal essay to me, which is not appropriate for academic writing here. (Rater 4, Verbal Reports, 10/18/2018)

Source use convention received a lot of attention from the raters, with 157 comments (10.6%), making this feature the third most attended to writing feature overall. In fact, all raters reported that they considered this feature when rating the essays, all though some raters (i.e., Rater 5, Rater 6, and Rater 7) had remarkably fewer comments compared to the others. The raters reported that they assessed if the source texts were cited appropriately, as shown in the example below.

And he copied this whole sentence without rephrasing or doing anything else to copy the sentence from the original text. So, he's making his sentence by copying the sentence from the source text. (Rater 9, Verbal Reports, 02/08/2018)

This example illustrates the fact that Rater 9 did compare the language in test takers' response with that in the source texts to find evidence that test takers paraphrased what they read in their writing. This was also a feature attended to by Rater 3 when rating, as she said:

And I wanted to look at the language that the original writer used to see how it matched what the student had written. I think again there's an attempt to try and use some of the student's own language, but I mean, "fortification" just really stands out for me. (Rater 3, Verbal Reports, 02/03/2018)

For this rater, test takers' use of some complex vocabulary items could be an indication that the items were taken from the source texts and thus, needs to be cited. Similar to Rater 3 and

Rater 9, Rater 1 required that test takers need to paraphrase language used in the source texts for their response.

...because I was a little confused of how they tried to paraphrase it. They're taking words from the text and using them. Like "deficiency, supplementation, fortification", and in the text, those terms were actually defined. So, I feel like you can actually paraphrase it pretty easily because even if you don't really know the term, that's defined for you in the text. (Rater 1, Verbal Reports, 02/02/2018)

In addition to paraphrasing the source texts' language, the raters also expected test takers to acknowledge the sources if ideas or information in their essay originated from the source texts, as Rater 2 recalled.

And then still on the second paragraph, I'm noting, "okay, we're talking about biotech companies. We're using information from the article." But I also noted that there wasn't any citation in the article or references to the article. It was just hearsay information. (Rater 2, Verbal Reports, 02/03/2018)

Echoing this criterion, Rater 6 required appropriate citation if test takers used source-based information in their essay, as indicated in her comment below.

And with the numbers, and I was thinking if again, the student needs to cite the article, the name of the author, or where he read that, or found this information. Because, there is like big numbers, percentages. So, maybe. Something with plagiarism. (Rater 6, Verbal Reports, 10/26/2018)

Other features

The last category of features that the raters assessed included *task completion* and *text length* although these two features were among those that received the least attention from the raters. Specifically, *task completion* received 37 comments (2.5%) while *text length* 20 comments, accounting for the smallest proportion of 1.3%.

All raters considered if test takers competed the requirement of Task 2, which was whether test takers successfully presented their perspective on a topic using supporting details from external sources. For instance, Rater 5 commented it was insufficient when the test taker only wrote a summary instead of stating their opinions, as evident from the following quote.

I thought, "so, this person hasn't finished the text yet." That's what I thought. I thought even though this person tries to express his or her ideas in here, I thought that this is from the source text. Those numbers. So, I thought even though this is task 2 and then even though this person uses "in my opinion" and "I think", what he is writing in here is about summary of the source text. (Rater 5, Verbal Reports, 09/06/2018)

In fact, test takers were expected to respond to the prompt, as Rater 3 mentioned below.

I remember thinking like, "well, that doesn't answer the prompt. That's a nice opinion but, that doesn't answer the prompt." (Rater 3, Verbal Reports, 02/03/2018)

Additionally, *text length* was also a feature that six raters attended to, although minimally. For example, Rater 9 stated that essay length was a factor indicative of test takers' writing fluency, as shown below.

And then, still, I had something in mind. I had the length of the essay in my mind, “Oh, this student might be good in terms of the fluency.” That's what I assumed. (Rater 9, Verbal Reports, 02/08/2018)

The fact that the raters attended to *task completion* and *text length* in addition to the writing features defined by the test construct and described in the rating rubric indicated that the raters also paid attention to features extraneous to the rubric during their evaluation of test takers' responses.

Summary

Overall, the analysis of the 90 verbal reports provided by the raters indicated that the raters paid attention to a variety of features in the writing responses. Specifically, they attended the most frequently to *Grammar and Lexis*, followed by *Arguments and Organization*, *Discourse Synthesis*, and *Style and Conventions*. The fact that raters focused the most on *Grammar and Lexis* and *Arguments and Organization* is consistent with the findings by previous studies (Barkaoui, 2010; Kim & Lee, 2015). In addition, the raters reported attending to different features of *Discourse Synthesis*, including accuracy of source-based information and quality of citation, as well as source use conventions. This finding is consistent with those reported by Gebril and Plakans (2014) who found that their two raters utilized strategies to locate source information, assess citation mechanics, and assess quality of source use. These features are aligned with the components of the language skills and processes defined in the EPT's source-based academic writing construct.

The raters also attended to features external to the rubric, such as *task completion* and *text length*, although these features received the least attention from them. This finding echoes those

reported by previous research (e.g., Barkaoui, 2010, Lumley, 2005; Smith, 2000; Vaughan, 1991) which found that raters attended to features extraneous to the rubric. However, the number of comments on these two features were very small, and thus, should not be considered a test construct-irrelevant variable.

An interesting finding from the analysis of the verbal reports pertained to the variability in the raters' understanding of the source integration. Thus, they varied in what to look for in the use of external sources. An examination of the rating rubric showed that source use was described in terms of (1) whether they were integrated skillfully, and (2) whether they were cited appropriately or paraphrased appropriately. The varied interpretation of source use among raters suggests that the construct will need to be more clearly defined and the criteria related to source use better described in the rating rubric. These findings suggest that that the test construct should be revised and communicated clearly to raters through the rubric and training.

Chapter Summary

This chapter examined (1) the appropriateness of the EPT Writing rating rubric in distinguishing test takers of different proficiency levels from the raters' opinions and statistical perspective (i.e., RQ1 and RQ2 respectively), (2) raters' comparability and consistency of severity (RQ3) as well as bias against certain test tasks (RQ4), (3) score reliability with the 2-task x 2-rater design (RQ5), and the writing features raters attended to when rating test takers' responses (RQ6). The data analyzed to address the research questions included quantitative data, consisting of 1,300 operational EPT ratings and 630 experimental ratings, and qualitative data, consisting of 90 verbal reports and 9 interviews from the raters.

Regarding RQ1, the results from the analysis of the raters' interviews indicated that the raters thought positively about the rating scale, although they suggested improvements in terms of score band descriptors, relevance criteria, the number of score bands, score band labeling, and weight for source integration. To address RQ2, the results from the MFRM analysis showed that the score band estimates increased monotonically from the lowest score band (i.e., level B) to the highest score band (i.e., Pass); however, score band B+ and score band C/D+ received relatively fewer ratings compared to the other score bands. In terms of RQ3, it was found that most raters were comparable and consistent in their severity. However, three raters were found to be significantly more generous than the others and one rater was not consistent in his severity. Regarding RQ4, it was found that most raters were impartial when rating the responses to the two tasks. Nevertheless, three raters were biased against a particular task type and became significantly more severe when rating its responses. Regarding the issue of test score reliability investigated in RQ5, the G-study results indicated that examinee ability explained a large proportion of the total variance. However, the interactions between examinee and rater, between examinee and task, as well as between examinee, rater, task, and the residuals also accounted for a high percentage of the total variance, indicating that more rater training is needed and that more tasks should be added to the test. The D-study results found that score reliability in the 2-task x 2-rater design was relatively low, with dependability index at .58. To reach the acceptable level of score reliability of at least .7, at least three tasks and three raters must be employed for the test and rating. For RQ6 pertaining to the features of writing the raters attended to while rating source-based writing task responses, an analysis of the raters' verbal reports showed that the raters attended to a variety of writing aspects, which were grouped in four categories, namely *Grammar and Lexis*, *Arguments and Organization*, *Discourse synthesis*, and *Style and*

Conventions. *Grammar and Lexis* being the most attended features, followed by *Arguments and Organization* and *Discourse synthesis*. These aspects were in alignment with the language skills and processes defined in the test construct, with *Grammar and Lexis* being the most attended features, followed by *Arguments and Organization* and *Discourse synthesis*. Source use, including *Discourse Synthesis* and part of *Style and Conventions*, was found to elicit different interpretations from the raters. The raters were also found to focus on other aspects such as task completion and text length although the number of comments pertaining to these features were very small.

The next chapter will discuss these findings and evaluate the degree to which they can be used to provide backing for the IUA for the source-based academic writing test for placement purposes introduced in Chapter 2. It will conclude the dissertation by explaining the implications and recommendations for validation research in language assessment and the EPT Writing and suggesting directions for future research.

CHAPTER 5: CONCLUSION

The previous chapter has presented the results of the analyses to address the six research questions regarding the appropriateness of the rating rubric, raters' performance, test score reliability, and the writing features attended to by the raters. These results are used to evaluate the extent to which empirical evidence provides backing for the IUA for the source-based academic writing test for placement purposes presented in Chapter 2. In this chapter, the IUA and the backing are combined and presented together to show the partial validity argument for the interpretation and use of scores from the EPT Writing test. This chapter also discusses implications of the study for source-based academic writing test development, validation research in language assessment and methods for data collection on cognition as well as recommendations for development of the EPT Writing test. The chapter ends with suggestions for future research.

A Partial Validity Argument for the Source-based Academic Writing Test for Placement Purposes

The goals of this study were to investigate the reliability of the test scores and the construct measured by the source-based academic writing test for placement purposes. Therefore, results from the research are evaluated to show their degree of support they provide for three inferences regarding test score reliability, namely *evaluation* and *generalization*, and test construct, namely *explanation*, forming a partial validity argument for the test. The first part in this section focuses on the test construct and uses. The remaining section presents the three inferences in the validity argument for the test.

The Test and Its Uses

The EPT Writing test is a computer-based for international students who do not speak English as their first language and wish to pursue a degree at a large Midwest university in the U.S. The test is intended to assess test takers' "source-based academic writing ability", with source referring to reading texts. Specifically, it aims to measure test takers' ability to (1) summarize and synthesize information presented from different sources, as well as (2) state and support their arguments with sufficient details and examples in a well-structured, coherent way with a good command of English vocabulary, grammar, and writing conventions. The 50-minute test requires test takers to read two texts on the same topic and compose (1) a summary of the two texts and (2) an essay presenting their own viewpoint using supporting details and examples from the two texts.

The test scores are used to determine students' readiness to be successful writers in academic contexts or placement into ESL academic writing classes. Depending on students' test scores, they can be placed into an intermediate writing course which emphasizes grammar and paragraph level writing, or more advanced writing courses that focus on source-based writing beyond the paragraph level for undergraduate students and research writing for graduate students. Those who pass the test are deemed ready for undergraduate, first-year composition courses or ready to be successful with writing in their graduate studies.

The Partial Validity Argument

The data used to evaluate the level of backing for the argument for the validity of the test score interpretation and use came from (1) the interviews with the raters, (2) the EPT Writing

operational ratings, and (3) the stimulated recalls provided by the raters. The backing is presented in three inferences, namely *evaluation*, *generalization*, and *explanation*.

Evaluation inference

The *evaluation* inference leads to the claim that test scores accurately and reliably summarize test takers' performance. It is based on the warrants that the rating rubric is appropriate for providing evidence of variation in source-based academic writing ability and that raters' performance is reliable. As shown in Figure 5.1, the first warrant is based on two assumptions: (1) raters consider the rating scale appropriate for assessing test takers' source-based writing ability, and (2) the scale steps are adequate to distinguish among the levels that appear in the scale. The second warrant is authorized by two assumptions: (1) raters are comparable and consistent in their ratings, and (2) raters do not exhibit significant bias against certain task types. Backing for these assumptions were obtained during the course of the study and presented in detail in Chapter 4.

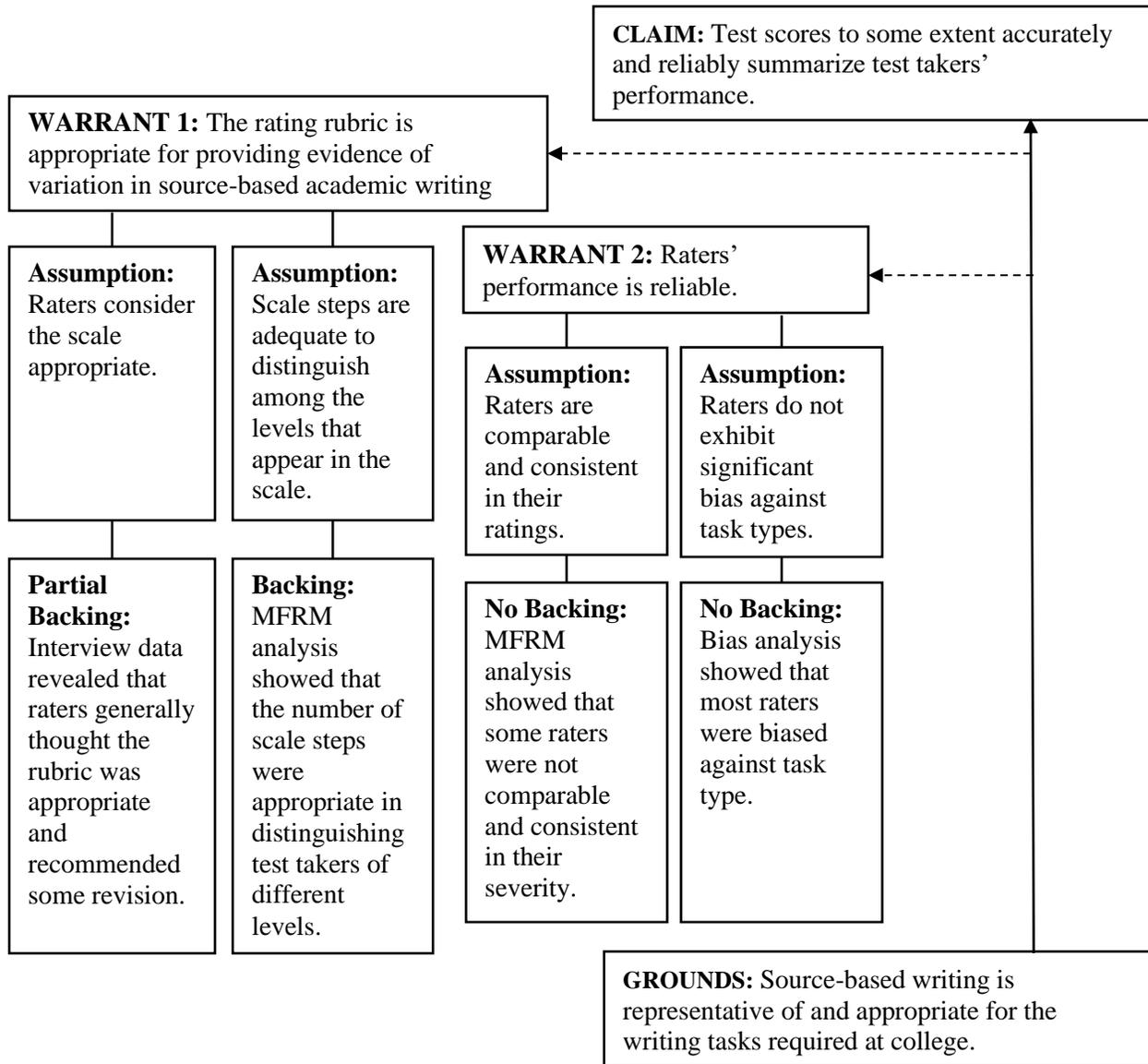


Figure 5.1. Evaluation inference with two assumptions and backing

Raters' opinions on the appropriateness of the rating rubric

The backing for the assumption that raters consider the rating scale appropriate for assessing test takers' source-based writing ability came from the analysis of interviews with raters. This assumption is partially supported because the raters generally thought the rubric was appropriate even though they recommended some revision. Specifically, the raters believed that

in general, the rubric was useful. They reported that it was mostly appropriate for assessing test takers' source-based academic task responses. However, a few raters recommended revisions of the rubric, especially regarding the description and weight of source use. In particular, they suggested that source use should have a clearer description and given more weight to emphasize its importance in a source-based academic writing test. On the other hand, three raters proposed that source use should not be considered a criterion for evaluating test responses. However, given that the construct of the test is "source-based academic writing", it is justifiable to consider ability to integrate external sources when rating test takers' performance. Additionally, two raters suggested that score band B+ should be excluded from the rubric because they did not find score band B+ very useful and thus, rarely used this score band when they rated. Two raters also questioned the relevance of spelling in the rubric or the clarity of quantifiers in the descriptors for each score band. Finally, two raters suggested relabeling the score bands to numbers instead of placement decisions (i.e., B, C/D, Pass). The positive opinions about the rubric in addition to these critical recommendations for improvement of the rubric are interpreted as partial support for the assumption that raters consider the rating scale appropriate for assessing test takers' source-based writing ability.

Adequacy of the rubric for distinguishing among the writing proficiency levels

Backing for the assumption that the scale steps are adequate to distinguish among the writing proficiency levels that appear in the scale was provided by the analysis of the EPT Writing operational ratings in *Facets*. Overall, this assumption was supported because of the linear growth of the average score band measures and the acceptable mean-square outfit statistics of the score bands. That is, the fact that the average score band measures increased monotonically from the lowest score band (i.e., B) to the highest score band (i.e., Pass) and that

the mean-square outfit statistic computed for each score band was below the acceptable threshold of 2.0 indicated that the rubric was effective in distinguishing writers of different levels.

However, the Andrich-Rasch thresholds did not advanced monotonically with the score bands showing that the raters did not use some score bands (i.e., score band B+ and C/D+) as frequently as others. This could have resulted from the fact that the score bands were labeled as placement decisions instead of numbers, which could have influenced raters' decision not to award a test taker with a B+ or C/D+. In fact, interview data revealed that some raters considered the implication of their decision during their assessment of test takers' responses, as Rater 1 reported that "*B and B+, I don't really know how that would be useful, especially if they're going to go into the same B class.*" (Rater 1, Interview, 02/02/2018). However, replacing the score band labels with numeric values would likely move the test to the norm-referenced testing framework, which conflicts with the purpose of the EPT Writing which aims to place test takers into appropriate writing courses and thus, follows the criterion-referenced testing framework. Therefore, more rigorous training is needed to help prevent raters who are also instructors of the ESL writing courses from thinking about the courses available while making decisions about test takers' performance.

Raters' comparability and consistency

The assumption that raters are comparable and consistent in their rating was not supported by the backing provided by the MFRM analysis of the EPT operational ratings. An examination of the Wright map, reliability and separation indices, Chi-square statistic, and rater fit statistics indicated that overall, the raters exercised a highly dissimilar degree of severity. Specifically, three raters were found to be significantly more generous than the remaining raters,

who were comparable in their severity. In terms of consistency, the fit mean square statistics showed that most raters were consistent in their severity. However, one exhibited inconsistency in their rating, possibly as a result of their overuse of the extreme score bands. Thus, it can be concluded that the assumption regarding raters' comparability and consistency was not supported.

Raters' bias against task type

The assumption that raters do not exhibit significant bias against task types was not supported by the backing provided by the bias analysis of the EPT operational ratings in *Facets*. Results from the bias analysis indicated that the majority of the raters did not display any bias against a specific task. However, four raters were found to be significantly more generous or more severe depending on the task they were rating, which as a result, negatively affects the reliability of the ratings provided by these raters. Based on this finding, it can be concluded that the assumption related to raters' impartiality about task type was not supported.

Based on the backing for the four assumptions, a conclusion for this inference is that test scores to some extent accurately and reliably summarize test takers' performance, as showed in Figure 5.1 above.

Generalization inference

The generalization inference can be made when the test takers' scores on EPT Writing test tasks are accepted as generalizable to scores they would receive on other similar tasks and scored by other raters. It leads to the claim the test scores reflect the desired level of performance consistency across tasks and raters. This inference is supported by the warrant that the score reliability is adequate for ESL writing course placement. As Figure 5.2 illustrates, this warrant

rests on the assumption that the number of raters and tasks is adequate to result in consistent scores.

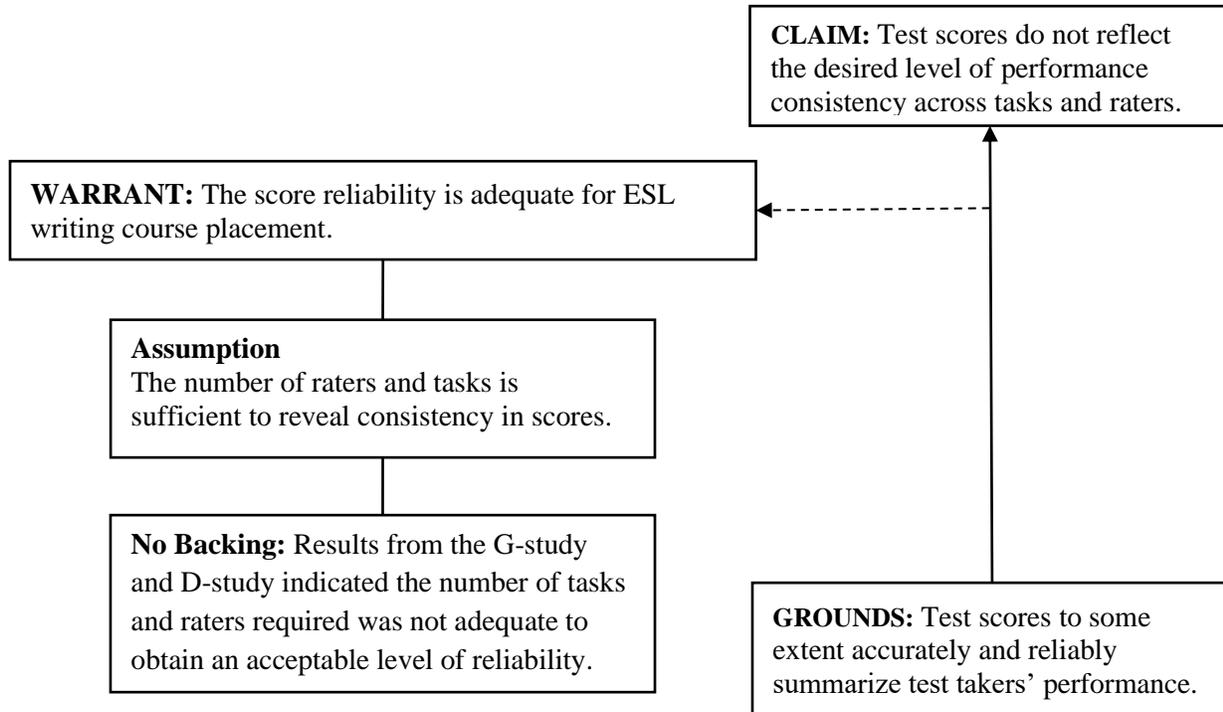


Figure 5.2. Generalization inference with one assumption and backing

The G-study and D-study of a subset of the EPT Writing operational ratings failed to find backing for the assumption that the numbers of raters and tasks used in the operational test are sufficient to reveal consistency in scores. Results from the D-study indicated that with the EPT Writing design consisting of two tasks and two raters, the score dependability was at .58, much lower than the desirable dependability of .7. Thus, the assumption regarding the adequacy of the number of tasks and raters was not supported. Therefore, the warrant authoring this assumption was not supported, leading to the conclusion that test scores do not reflect the desired level of performance consistency across tasks and raters (Figure 5.2).

Explanation inference

In the original IUA, the *explanation* inference leads to the claim test scores are indicators of the construct of source-based academic writing. It is authorized by the warrant that raters' attention during rating is aligned with the test construct. As displayed in Figure 5.3, the assumption underlying this warrant is that the writing features that raters attend to are appropriate in view of the construct of source-based academic writing ability defined for the test.

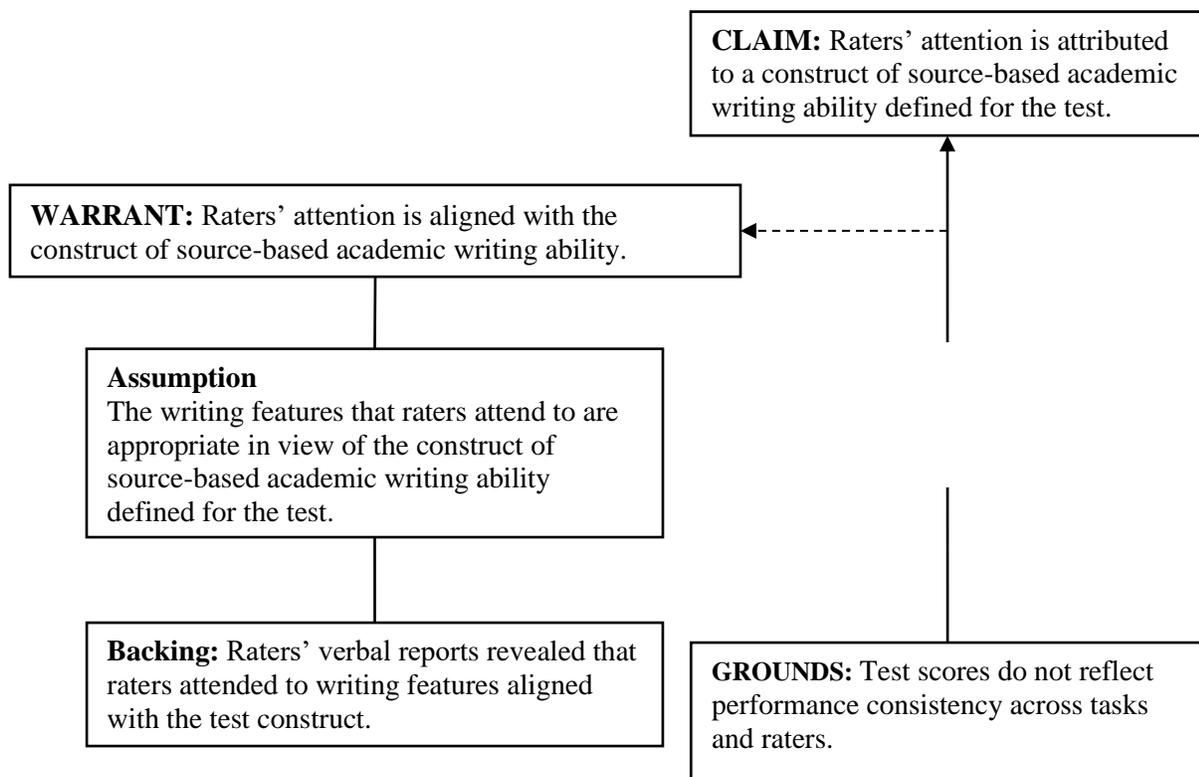


Figure 5.3. Structure of the explanation inference in the IUA for the EPT Writing

Backing for the assumption regarding the writing features attended to by the raters came from the analysis of the raters' verbal reports. This assumption was supported since the results showed that the raters paid attention to a variety of writing features that were consistent with the components of the language skills and processes defined in the EPT's source-based academic

writing construct. Specifically, they attended the most to *Grammar and Lexis*, followed by *Arguments and Organization*, *Discourse Synthesis*, and *Style and Conventions*. Therefore, there was evidence for the raters' attention to source use, which consisted of *Discourse synthesis* (i.e., source text understanding and citation quality) and part of *Style and Convention* (i.e., source use convention). These features were consistent with the components of the language skills and processes defined in the EPT's source-based academic writing construct. The raters also reported attending to *task completion* and *text length*, two features external to the rubric. However, these features received minimal attention from the raters, at 2.5% and 1.3%. Thus, it can be concluded that overall, there was evidence supporting the assumption that the writing features attended to by raters are appropriate in view of the construct of source-based academic writing ability defined for the test.

However, since the grounds for this inference is not solid due to the results of the D-study about test score consistency across tasks and raters (as indicated by the broken arrow in Figure 5.3), the original claim that test scores are indicators of the construct of source-based academic writing cannot be supported. Thus, my conclusion for this inference is that raters' attention is attributed to a construct of source-based academic writing ability defined for the test.

Implications

This study has useful implications for the definition of source-based writing constructs, data collection methods in cognition research, and validation research on language assessment.

A Construct of Source-based Academic Writing

An important implication of the study for the development of source-based writing assessments is related to the definition of a construct of source-based academic writing. In the

original definition of the EPT Writing construct, which consists of various types of language knowledge, source use was defined as the ability to identify key points and represent of reading ideas. This definition, together with how it is operationalized on the rating rubric, proved to be insufficient, as revealed by the raters who reported different interpretations of source use. Specifically, raters reported attending to different aspects of source use despite the fact that two of them did not think that source use should be evaluated. The source-use related features the raters attended to included (1) the accuracy of source-based information integrated in test takers' responses, (2) the quality of the citation, and (3) the conventions test takers followed when integrating sources. Clearly, the construct that guides the rubric development and rater training needs to be more carefully considered. A possible revised construct of source-based academic writing is illustrated in Figure 5.4.

In this revised construct, source use, an important component of the language knowledge, skills, and abilities the test is intended to measure, is more clearly and comprehensively delineated. First, source use ability entails ability to use information from the source texts accurately. In order to successfully accomplish this, test takers need to have good reading ability to comprehend the source texts. Second, ability to integrate external sources includes ability to use the information from the sources to support one's argument in a coherent way. That is, the purpose of source text integration must be apparent must be clear and the flow between the writers' own arguments and source-based information must be coherent to the reader. Finally, source integration ability also includes ability to follow conventions when integrating sources, such as using quotation marks for direct quotations and paraphrasing.

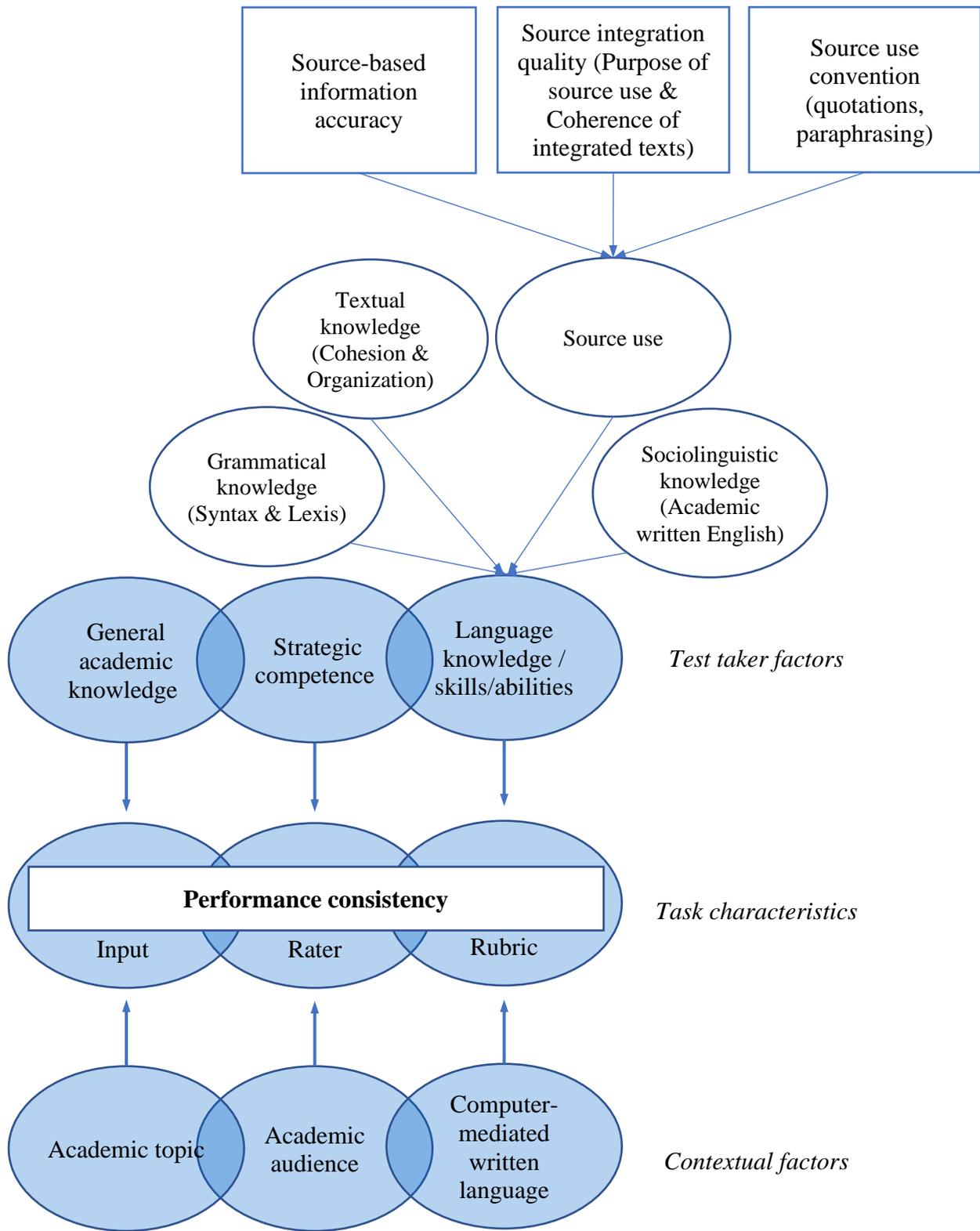


Figure 5.4. A revised source-based academic writing construct

These aspects can be used to describe source use ability for two reasons. First, these aspects of the construct indicate some of the features reported by the raters in this study who were appropriately focused on assessing the quality of source use in the responses. Also, these same features appeared in the findings of Gebril and Plakans (2014) who found that raters, when rating source-based task responses, attended to issues related to locating source information, citation mechanics, and quality of source use. Second, a construct of source use ability that consists of these features follows suggestions from previous researchers (Ascensión, 2008; Plakans, 2009) who proposed that a construct of source-based writing should include discourse synthesis, which could be seen as part of source use ability. According to Spivey (1997), when responding to a source-based, reading-to-write task, writers undergo three key textual operations in discourse synthesis: (1) organizing content of source texts by dismantling the organization of the sources and shaping their meanings with new organizational patterns, (2) selecting content from source texts that are relevant for inclusion in their own writing, and (3) connecting topical information from source texts to provide a coherent flow between the propositions selected. Test takers who undergo these operations efficiently use source-based information accurately (i.e., source-based information accuracy) as well as meaningfully and coherently (i.e., source integration citation).

Overall, it is important to include these three components of source use ability when defining a construct of source-based academic writing, which serves as a basis for rubric development and rater training. It is crucial to operationalize source use clearly in the rubric to communicate with raters to minimize variation in their interpretation of source use.

Data Collection Methods in Cognition Research

In addition to implications related to source-based writing test development, this study also has implications for methods of cognition data collection. This study demonstrated how eye-tracking technology can be used to prompt data collection on raters' cognition. In particular, the data showed that the recordings of the raters' eye movement trajectories as well as their viewing order of test materials (i.e., the test responses, rubric, and sources texts) helped elicit detailed recollection of raters' thought processes during their evaluation of the test responses. Previous research on raters' cognition used mostly think-aloud protocols, which interferes with the real rating process (Cumming et al., 2001; Lumley, 2005; Stratman & Hamp-Lyons, 1994). Studies on raters' cognition using eye-tracking technology (e.g., Ballard, 2017) have been criticized for their assumption that viewing means attention. Nevertheless, it was observed in this study that a combination of eye-tracking technology and stimulated recalls facilitated fruitful and detailed reports from participants. Based on the success of this study, future studies on raters' cognition should explore the use of eye-tracking technology as a way to elicit recalls of thought processes from participants. For example, future research could follow an experimental design to compare participants' ability to recall their thought processes in two conditions: (1) with eye-tracking data as stimuli and (2) without eye-tracking data as stimuli. Such studies would provide backing for my earlier argument that the combination of stimulated recalls and eye-tracking technology, used to elicit stimulated recalls from participants, is effective for gathering detailed information about participants' cognition.

Validation Research on Language Assessment

Regarding language test validation research, this study showed how reliability-related findings are connected to findings about construct validity using the argument-based approach to validation. Previous studies investigating the topic of rating processes on source-based writing tests have focused either on the reliability of the rating processes or the construct validity of the resulting scores. This study took into account the link between the reliability of the ratings because of the importance of the reliability of the rating processes and outcomes to any findings about the test construct. Findings from an investigation of a test construct, the intended explanation for performance consistency, need to be interpreted in view of the actual consistency of the scores, as revealed through reliability-related investigations.

Figure 5.5. connects these aspects of the validity argument by showing how the study results affect the warrants, assumptions, and conclusions for the *evaluation*, *generalization*, and *explanation* inferences. The figure shows that this study started with an investigation of the *evaluation* inference, which is based on the grounds that source-based writing is representative of and appropriate for the writing tasks required at college. Two warrants required support in order to make a claim about the quality of the rubric and rating. The first warrant had two assumptions related raters' opinions about the rubric appropriateness and the functionality of the rubric which were partially and fully supported respectively. Therefore, the warrant that the rating rubric is appropriate for providing evidence of variation in source-based academic writing ability is also supported.

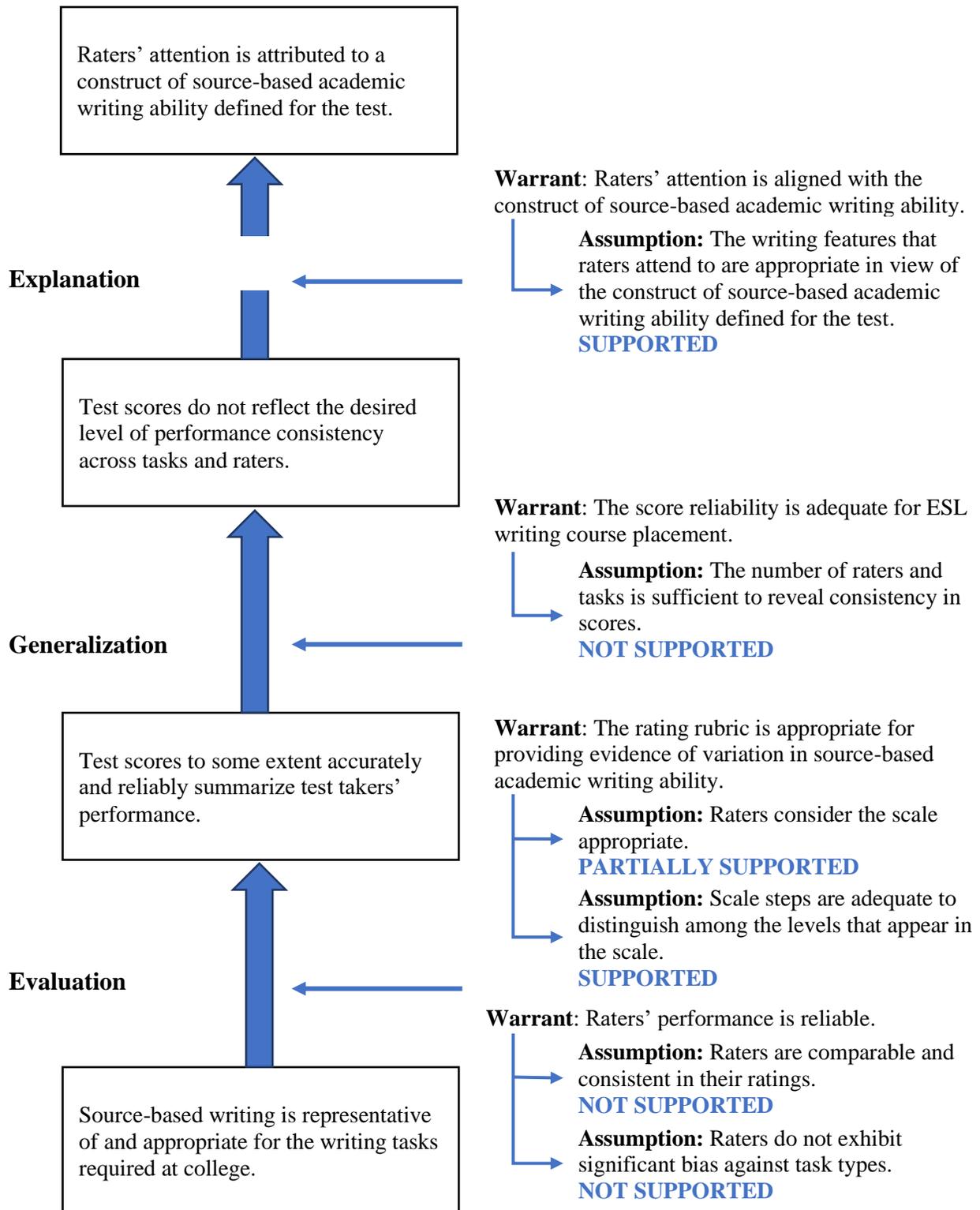


Figure 5.5. A partial validity argument for the EPT Writing showing the level of support obtained for warrants and assumptions for *evaluation*, *generalization*, and *explanation*

However, the second warrant about the quality of raters' performance required support for the assumptions about raters' comparability, consistency, and bias. These assumptions were not supported, leaving the warrant that the raters' performance reliable not supported. Therefore, the claim for the *evaluation* is only partially supported: that the results allowed for a claim that test scores to some extent accurately and reliably summarize test takers' performance. The validity argument shows the importance of these findings about the rating process by revealing that this claim also serves as grounds for the next inference, *generalization* about score reliability across tasks and raters.

Figure 5.5 shows that the *generalization* inference has one warrant that needs to be supported. However, its one assumption was not supported. As a result, the warrant is not supported, leading to the conclusion that test scores do not reflect the desired level of performance consistency across tasks and raters. The finding of low reliability of the test scores means that the grounds for the explanation inference is weak.

The *explanation* inference in Figure 5.5 is left without solid grounds, but it has one warrant. The warrant requires support for the assumption that the writing features that raters attend to are appropriate in view of the construct of source-based academic writing ability defined for the test. Sufficient backing was found for this assumption for the raters' verbal reports, leading to the support of the warrant that raters' attention is aligned with the construct of source-based academic writing ability. Based on this backing, one would like to draw the conclusion that test scores are attributed to a construct of source-based academic writing ability, the claim stated in the original IUA. However, the results of the generalization inference showing the low reliability of the scores did not present solid grounds for making such an inference. Thus,

a plausible claim for this inference is that raters' attention is attributed to a construct of source-based academic writing ability defined for the test.

The verbal reports provided some indication that raters attended to the response characteristics that were relevant to the construct definition. However, they also revealed that the raters varied in their interpretation and evaluation of source use. This variation fits with the picture of variability among raters that resulted in only partial support for the evaluation inference and lack of support for the generalization inference. Together, these results suggest the need for a construct definition that better captures the intended meaning of source use, a clearer description of source use in the rubric, and more rigorous rater training that attunes raters to this aspect of the construct and rubric.

In sum, the validity argument framework revealed specific aspects of the testing processes that fell short of desired goals as well as the implications of each part of the process for the overall construct interpretation when the scores are used. The IUA outlined the backing needed to arrive at each of the intended claims about rating processes, test score reliability and construct validity. The validation research provided results that allowed for a judgment about the plausibility of each of the intended claims. The overall structure of the IUA and validity arguments shows the connections among each part of the research to provide a basis for an overall interpretation.

Recommendations for the EPT Writing

Based on the findings of the study, key recommendations for the EPT Writing include revision of test construct and rating rubric as well as rater training.

Revision of the Test Construct and Rating Rubric

The test construct and the rating rubric need to be revised to better reflect the source-based academic writing construct that the test aims to measure. As presented in Figure 5.4 above, a construct of source-based academic writing must clearly define what source use ability entails. The rubric must be revised to reflect this change in the test construct.

First, source integration ability should be clearly described in the rubric. The verbal reports from the raters indicated that they attended to all components of the language skills and processes defined in the test construct. However, their interviews revealed that the raters had different interpretations of source integration ability. The findings related to source use from the verbal reports provided a clear description of source integration ability, which the raters understood to include accuracy of source-based information, quality of source use, and source use convention. The rubric does not describe these components clearly, leaving room for raters to decide what source use means. Also, as this is a source-based writing test, source use as a rating criterion should be more visible to raters on the rubric. That is, it should be placed as a separate category instead of being combined with *Conventions*. A clearer description in the rubric of what raters should look for to evaluate source integration ability, along with proper training of the raters, will help increase the reliability of the ratings.

Additionally, EPT Writing test developers could consider limiting the number of score bands to four instead of five. Because EPT test takers are students already admitted to the University, and thus, have met a certain level of English proficiency, their proficiency range might not be large enough to be categorized into five different score bands. In fact, EPT test takers are usually students whose TOELF score between 71 and 99. Therefore, reducing the

number of score bands might help ascertain that all score bands are used relatively equally by raters. This would help with the issue of disordered Andrich-Rasch threshold values presented in the Results chapter.

Rater Training

The results from the D-study indicated that score reliability for the rating design with two tasks and two raters rating test responses was lower than the acceptable coefficient of .7. To obtain a reliability coefficient of .7, three writing tasks and three raters need to be employed. Other scenarios resulting in a lower reliability coefficient, at .65 or .68, include using two tasks and employing three raters and using two tasks and four raters. In either case, more raters are needed to arrive at a reliability coefficient of at least .65.

However, increasing the numbers of test tasks and raters might introduce practicality issues. For example, more time will be need for test takers to finish the test and for raters to rate the test responses. Also, employing more raters implies more human resources and financial resources are needed. Because of these practical constraints as well as the fact the EPT has a very tight schedule to report test scores to test users, more thorough rater training could be the solution to the issue of low score reliability.

Rigorous training must be conducted to ensure that raters have similar interpretations of the rubric, especially regarding source integration ability, and to enhance their severity comparability and consistency. MFRM analysis indicated that some raters were significantly more generous than others and that they were not consistent in their own severity levels across ratings. Also, G-study results revealed the effects of the examinee-by-rater interaction as well as the three-way interaction between examinee, rater, and task compounded with other unmodeled

errors on the composite scores were large, indicating that the test takers were not scored similarly across raters. These findings suggest that more rigorous training of the raters is needed to help alleviate the effects of raters' idiosyncrasies and obtain more reliable ratings. In fact, past studies have shown the positive effects of rater training – training tends to homogenize raters and improve their rating quality by clarifying scoring criteria, increasing interrater reliability, reducing bias, and improving reliability (e.g., Davis, 2016; Kang, Rubin, & Kermad, 2019; Kondo, 2010; Lim, 2011; Shohamy, Gordon, & Kraemer, 1992; Wigglesworth, 1993; Xi & Mollaun, 2011).

Directions for Future Studies

This study suggests several directions for future research on source-based writing assessment. The ideas for these directions can be organized into two groups: rater cognition in source-based assessment and validation issues for the source-based academic writing test for placement purposes.

Rater Cognition in Source-Based Writing Assessment

Considering the fact that this study did not examine raters' individual differences in their cognition, future research could be conducted to investigate the relationship between rater characteristics, such as raters' experience with rating and familiarity with ESL writing courses, and their thought processes. Research has been done to identify how factors related to raters' background contribute to variability in their decision-making process (Barkaoui, 2010; Brown, 1995; Cumming et al., 2001; Elder, 1993; Eckes, 2008; Erdosy, 2004; Wiseman, 2012), few studies have explored this topic in source-based writing. Such studies will be beneficial for rater recruitment and training for source-based writing tests.

Additionally, future studies could investigate the relationship between rater cognition and the reliability of the scores raters award to test takers. For example, raters' attention to writing features could be examined in relation to their rating performance (i.e., comparability, consistency, and bias) to identify features attended to by more effective raters. Also, with the use of eye-tracking technology, researchers can observe raters' step-by-step access to rating materials, including the response, rubric, and source texts, as well as collect information about their decision-making process. Results from the analyses of these data could help build a model of effective raters, which could be valuable for rater training.

Validation Issues for The Source-Based Academic Writing Test for Placement Purposes

In addition to rater cognition in source-based writing, more research should be conducted to investigate other validation issues for the source-based academic writing test for placement purposes. This dissertation only focused on three inferences in the IUA, namely, *evaluation*, *generalization*, and *explanation*. To build a more comprehensive validity argument, other inferences, including *domain definition*, *extrapolation*, *utilization*, and *consequence*, must be investigated. Also, because each inference can be authorized by multiple warrants and assumptions, further studies should be conducted to examine *evaluation*, *generalization*, and *explanation* from different angles from the foci of this study. Table 5.1 outlines the additional warrants, their underlying assumptions, and backing associated with each inference in the IUA for the EPT Writing that could be investigated in the future.

Table 5.1. *Additional Warrants, Assumptions, and Backing Associated with Each Inference in the IUA for the EPT Writing*

Warrant	Assumption	Example of backing
<p>Domain definition inference Claim: Source-based writing is representative of and appropriate for the writing tasks required at college.</p>		
Observations of performance on the test reveal relevant knowledge, skills, and abilities in situations representative of those in the target domain of source-based academic writing.	Writing tasks that are representative of the academic domain and essential academic writing skills, knowledge, and processes needed to be successful can be identified.	Domain analysis (expert consensus, syllabus and textbook analysis, student interviews)
	Writing tasks that require important skills and are representative of the domain can be simulated.	Task design and modeling
<p>Evaluation inference Claim: Test scores accurately and reliably summarize test takers' performance.</p>		
Observations of performance in the test tasks are evaluated to provide test scores reflective of source-based academic writing ability.	Test administration conditions are appropriate for providing evidence of targeted language abilities.	Prototyping studies
<p>Generalization inference Claim: Test scores reflect the desired level of performance consistency across tasks and raters.</p>		
Test scores are estimates of expected scores over the relevant parallel versions of test forms.	Task, test, and rating specifications are well defined so that parallel task and test forms are created.	Systematic development of test specifications
	The writing prompts are comparable.	Expert consensus Statistical analysis of test scores from different prompts Linguistic analysis of writing prompts

Table 5.1 Continued

Warrant	Assumption	Example of backing
Explanation inference Claim: Test scores are attributed to a construct of source-based academic writing ability		
The linguistic knowledge, processes, and strategies required to successfully complete tasks are aligned with the test construct.	Writing features in test responses are aligned with the language skills and knowledge defined in the test construct.	Discourse analyses of test takers' responses
	The cognitive processes test takers employed during test completion are aligned with the expected process identified in the test construct.	Cognitive processing studies
Performance on the test relates to performance on other test-based measures of language proficiency.	Test scores positively and moderately correlated with test score from other academic writing tests.	Concurrent correlation studies
Test performance varies according to amount and quality of English learning experience.	Test performance correlated positively with duration and quality of English study.	Comparison studies
Extrapolation inference Claim: Test scores reflect samples of the intended performance in source-based academic writing.		
The construct of source-based academic writing ability assessed by the test accounts for the quality of source-based academic writing performance in the academic context.	Performance on the test is related to other criteria of source-based academic writing in the academic context.	Criterion-related validity studies
Utilization inference Claim: Test scores are useful in placing students in ESL writing courses.		
Estimates of the quality of performance obtained from the test are useful for making decisions about placement and appropriate curriculum for test takers.	The test scores provide useful and meaningful information about students' source-based academic writing ability.	Interviews with instructors, experts, and test takers

Table 5.1 Continued

Warrant	Assumption	Example of backing
Consequence inference Claim: The use of test scores results in positive consequences for test users and society.		
The consequences of using the test and of the decisions that are made are beneficial to test takers and instructors of ESL and other writing courses.	The using of the test and the decisions made based on the test scores will have a positive influence on test takers and how academic writing is taught.	Washback studies

As present in Table 5.1, the *domain definition* inference is based on the warrant that observations of performance on the test are representative of those in the target domain of source-based academic writing, eliciting relevant knowledge, skills, and abilities to successfully complete writing tasks similar to those in the target domain. The underlying assumptions of this warrant are: (1) writing tasks that are representative of the academic domain and essential academic writing skills, knowledge, and processes needed to be successful can be identified, and (2) such writing tasks that require important skills and are representative of the academic domain can be simulated as test tasks. Backing for the first assumption could be provided by domain analysis such as analyses of interviews and questionnaires with experts and students, as well as of syllabi, textbooks, and assignment instructions in various courses. To support the second assumption, expert judgment could be used in the process of task design and trialing. The combination of these evidence sources substantiates the conclusion that source-based writing is representative of and appropriate for the writing tasks required at college.

Future studies can also examine the *evaluation* inference from a different perspective. This inference is built on the warrant that observation of performance in the test tasks are evaluated to provide test scores reflective of source-based academic writing abilities. The

assumption underlying this warrant is test administration conditions are appropriate for providing evidence of targeted language abilities. Backing for these assumptions is drawn from the development, trialing, and revisions of multiple test administration conditions based on expert consensus. The test administration conditions could vary in terms of how many reading texts are used as stimuli, whether the five-minute reading time prior to writing should be mandatory to test takers, and whether the test time should be 30 or 40 minutes.

For the *generalization* inference, future research can investigate the warrant that test scores are estimates of expected scores over the relevant parallel versions of test forms. This warrant is based on two assumptions: (1) the task, test, and rating specifications are specific enough for developing parallel task and test forms, and (2) the writing prompts are comparable. To support the first assumption, experts' opinions could be explored to see if the task, test, and rating specifications are detailed enough for developing parallel task and test forms. Backing for the second assumption could be drawn from interviews with experts, discourse analysis of the prompts, and statistical analysis of test scores across different prompts.

In addition to the warrant investigated in the study, the *explanation* inference is also built on three additional warrants: (1) the linguistic knowledge, processes, and strategies required to successfully complete tasks are aligned with the test construct, (2) performance on the test relates to performance on other test-based measures of language proficiency as expected theoretically, and (3) test performance varies according to amount and quality of experience in learning English. The assumptions supporting the first warrant include (1) writing features in test responses are aligned with the language skills and knowledge defined in the test construct and (2) the cognitive processes test takers employed during test completion are aligned with the expected process identified in the test construct. Backing evidence used to support the first

assumption is collected from analysis of test takers' essays and their cognitive processing using eye-trackers, screen capture, retrospective recalls, and questionnaires. Specifically, test takers' essays could be analyzed to identify the features of language use indicating good source-based academic writing ability such as grammatical and lexical accuracy and complexity, coherent organization of ideas with sufficient supporting details, and appropriate use of academic written English conventions and external sources. Also, test takers' cognitive processing and strategies during test completion could be investigated with eye trackers, screen capture, retrospective recalls, and post-test questionnaire to provide evidence for the assumption that these processes are aligned with those expected in the test construct. The second warrant is authorized by the assumption that test scores positively and moderately correlated with test score from other academic writing tests. To support this assumption, backing evidence could be obtained from concurrent correlation studies on the relationship between the test scores and scores on other academic writing tests, such as the TOEFL and IELTS writing component. The third warrant is authorized by the assumption that test performance correlated positively with duration and quality of English study. Backing for this assumption could be gathered from comparison studies of group differences which examine the relationships between performance and English learning and writing experience.

Another inference that future research could focus on is the *extrapolation* inference connecting the test construct to the expected test score overall possible performances in the target domain. This inference is based on the warrant that the construct of source-based academic writing ability assessed by the test accounts for the quality of source-based academic writing performance in the academic context. The assumption underlying this warrant is that performance on the test is related to other criteria of source-based academic writing in the

academic context. Backing for this assumption could come from criterion-related validity studies investigating the relationships between test performance and test takers' performance on source-based writing assignments in other courses.

Additionally, future research could also examine the *utilization* inference that connects the target score with test use. This inference is authorized by the warrant that estimates of the quality of performance obtained from the test are useful for making decisions about placement and appropriate curriculum for test takers. This warrant rests on the assumptions that the test scores provide useful and meaningful information about students' source-based academic writing ability and that the meaning of test scores is clearly interpretable by academic advisors, ESL instructors, and test takers. The backing for these assumptions could be collected studies where different stakeholders are interviewed to examine the usefulness of the test in making inferences about test takers' writing ability and making decisions about test takers.

The last inference, *consequence*, could also be a potential topic for future research. This inference connects test use to consequences of test score interpretation and use and is based on the warrant that the consequences of using the test and of the decisions that are made are beneficial to test takers, instructors of academic courses. This warrant assumes that the using of the test and the decisions made based on the test scores will have a positive influence on the test takers and how academic writing is taught. This assumption requires evidence from washback studies where the impact of the test is examined through interviews with instructors, experts, and students, together with questionnaires for students.

Such studies will undoubtedly broaden our understanding about the validity of the interpretation and use of test scores from the source-based academic writing test for placement purposes.

Conclusion

This dissertation is the first study that investigates the issue of construct validity from the perspective of rating processes used in rating test response. By employing an innovative approach to collect raters' cognition data – one that employed eye-tracking data as a tool for eliciting raters' verbal reports – the study showed how raters understand source integration ability by documenting the evidence they saw in test takers responses. Based on raters' reports and previous research investigating source-based writing, I concluded that a source integration ability component should be explicitly defined as part of the construct of source-based academic writing for tests such as the one investigated in this research. Findings from the study also provides important implications for validation research in the field of language assessment and methods for collecting cognition data. It also gives useful recommendations to improve the rating process of a source-based academic writing test for placement purposes at a large Midwest university, including rating rubric revision along with rater training.

REFERENCES

- Anderson, J. R., Bothell, D., & Douglass, S. (2004). Eye movements do not reflect retrieval processes: Limits of the eye-mind hypothesis. *Psychological Science, 15*(4), 225–231.
- Asención, Y. (2008). Investigating the reading-to-write construct. *Journal of English for Academic Purposes, 7*(3), 140–150. doi:<http://dx.doi.org/10.1016/j.jeap.2008.04.001>
- Atilgan, H. (2013). Sample size for estimation of G and phi coefficients in Generalizability theory. *Eurasian Journal of Educational Research, 51*, 215–227.
- Bachman, L. F. (2007). What is the construct? The dialectic of abilities and context in defining constructs in language assessment. In J. Fox, M. Wesche, & D. Bayless (Eds.), *What are we measuring? Language testing reconsidered* (pp. 41–72). Ottawa, Canada: University of Ottawa Press.
- Bachman, L. F., & Palmer, A. S. (1996). *Language assessment in practice*. Oxford: Oxford University Press.
- Bachman, L. F., Lunch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*(2), 238–257. doi:[10.1177/026553229501200206](https://doi.org/10.1177/026553229501200206)
- Ballard, L. (2017). The effects of primacy on rater cognition: An eye-tracking study. *Unpublished dissertation*. Lansing: Michigan State University. Retrieved on April 9, 2018 from <https://d.lib.msu.edu/>.
- Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54–74.
- Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing, 28*(1), 51–75. doi:[10.1177/0265532210376379](https://doi.org/10.1177/0265532210376379)
- Barkaoui, K. (2014). Multifaceted Rasch analysis for test evaluation. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1301–1322). New York: John Wiley & Sons, Inc. doi:[10.1002/9781118411360.wbcla070](https://doi.org/10.1002/9781118411360.wbcla070)

- Barkaoui, K. (2015). Test takers' writing activities during the TOEFL iBT® writing tasks: A stimulated recall study. *ETS Research Report Series*, 2015(1), 1–42.
doi:10.1002/ets2.12050
- Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. *Language Testing*, 30(4), 441–465. doi:10.1177/0265532212473244
- Bax, S., & Weir, C. (2012). Investigating learners' cognitive processes during a computer-based CAE Reading test. *Cambridge Research Notes, Cambridge ESOL*, 47 (February 2012), 3–14. Retrieved on Sep 5, 2018 from www.cambridgeesol.org/rs_notes/rs_nts47.pdf
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.
- Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, 12(1), 1–15.
- Brown, J. D., Hilgers, T., & Marsella, J. (1991). Essay prompts and topics: Minimizing the effect of mean differences. *Written Communications*, 8, 533–556.
- Brunfaut, T., & McCray, G. (2015). Looking into test takers' cognitive processes while completing reading tasks: A mixed-method eye-tracking and stimulated recall study. *British Council Research Notes*. Retrieved on Sep 5, 2018 from https://www.britishcouncil.org/sites/default/files/brunfaut_and_mccray_report_final_0.pdf
- Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing*, 20(4), 369–383.
- Chapelle, C. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70). New York: Cambridge University Press.
- Chapelle, C. A. (2008). Chapter 9. The TOEFL validity argument. In C. A. Chapelle, M. K. Enright, & J. M. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language™* (pp. 319–352). New York: Routledge.

- Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple... *Language Testing*, 29(1), 19–27.
- Chapelle, C. A. (2021). *Argument-based validation in testing and assessment*. Thousand Oaks, CA: Sage Publications, Inc.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (Eds.). (2008). *Building a validity argument for the Test of English as a Foreign Language™*. New York: Routledge.
- Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice*, 29(1), 3–13. doi:10.1111/j.1745-3992.2009.00165.x
- Cho, Y., Rijmen, F., & Novak, J. (2013). Investigating the effects of prompt characteristics on the comparability of TOEFL iBT integrated writing tasks. *Language Testing*, 30(4), 513–534. doi:10.1177/0265532213478796
- Chubb, C. (2013). *Human information processing: Vision, memory, and attention*. Washington, DC: American Psychological Association.
- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research*. Thousand Oaks, CA: Sage Publications, Inc.
- Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework* (TOEFL Monograph Series N 22). Princeton, NJ: Educational Testing Service.
- Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, 33(1), 117–135. <https://doi.org/10.1177/0265532215582282>
- Deubel, H. (2008). The time course of presaccadic attention shifts. *Psychological Research*, 72(6), 630–640.
- Deygers, B., & Van Gorp, K. (2015). Determining the scoring validity of a co-constructed CEFR-based rating scale. *Language Testing*, 32(4), 521–541. <https://doi.org/10.1177/0265532215575626>

- Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language Testing*, 25(2), 155–185. doi:10.1177/0265532207086780
- Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt, Germany: Peter Lang.
- Elder, C. (1993). How do subject specialists construe classroom language proficiency? *Language Testing*, 10(3), 235–254. doi:10.1177/026553229301000303
- Erdosy, M. U. (2004). *Exploring variability in judging writing ability in a second language: A study of four experienced raters of ESL compositions* (TOEFL Research Report No. RR-03-17). Princeton, NJ: Educational Testing Service.
- Ericsson, K., & Simon, H. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, Mass.: MIT Press.
- Esmaeili, H. (2002). Integrated reading and writing tasks and ESL students' reading and writing performance in an English language test. *Canadian Modern Language Review*, 58(4), 599.
- Foucart, A., & Frenck-Mestre, C. (2012). Can late L2 learners acquire grammatical features? Evidence from ERPs and eye-tracking. *Journal of Memory and Language*, 66, 226–248.
- Frenck-Mestre, C. (2005). Eye-movement recording as a tool for studying syntactic processing in a second language: A review of methodologies and experimental findings. *Second Language Research*, 21(2), 175–98. <https://doi.org/10.1191/0267658305sr257oa>
- Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13(2), 208–238.
- Gamer, M., Lemon, J., Fellows, I. & Singh, P. (2012). Package “irr”. Retrieved on June 19, 2018 from <https://cran.r-project.org/web/packages/irr/irr.pdf>
- Gebril, A. (2009). Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing*, 26(4), 507–531.
- Gebril, A. (2010). Bringing reading-to-write and writing assessment tasks together: A generalizability analysis. *Assessing Writing*, 15, 100–117.

- Gebril, A., & Plakans, L. (2013). Toward a transparent construct of reading-to-write tasks: The interface between discourse features and proficiency. *Language Assessment Quarterly*, *10*(1), 9–27. doi:10.1080/15434303.2011.642040
- Gebril, A., & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. *Assessing Writing*, *21*, 56–73. doi:http://dx.doi.org/10.1016/j.asw.2014.03.002
- Godfroid, A., & Uggem, M. S. (2013). Attention to irregular verbs by beginning learners of German. *Studies in Second Language Acquisition*, *35*(2), 291–322. doi:doi:10.1017/S0272263112000897
- Godfroid, A., & Spino, L. A. (2015). Reconceptualizing reactivity of think-alouds and eye tracking: Absence of evidence is not evidence of absence. *Language Learning*, *65*(4), 896-928. doi:10.1111/lang.12136.
- Godfroid, A., Boers, F., & Housen, A. (2013). An eye for words: Gauging the role of attention in incidental L2 vocabulary acquisition by means of eye-tracking. *Studies in Second Language Acquisition*, *35*(3), 483–517. doi:10.1017/S0272263113000119
- Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge: Cambridge University Press.
- He, A. & Young, R. (1998). Language proficiency interviews: A discourse approach. In R. Young & A. He (eds.). *Talking and testing* (pp. 1–24). Amsterdam: John Benjamins.
- Holmqvist, K., Nystrom, M., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford, UK: Oxford University Press.
- Huhta, A., Alanen, R., Tarnanen, M., Martin, M., & Hirvela, T. (2014). Assessing learners' writing skills in an SLA study: Validating the rating process across tasks, scales and languages. *Language Testing* *31*(3), 307–328.
- Johns, A. M., & Mayes, P. (1990). An analysis of summary protocols of university ESL students. *Applied Linguistics*, *11*(3), 253–271. doi:10.1093/applin/11.3.253

- Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin*, *112*(3), 527–535. <http://dx.doi.org/10.1037/0033-2909.112.3.527>
- Kane, M. (2006). Validation. In R. Brennen. (Ed.), *Educational measurement* (4th ed.) (pp. 17–64). Westport, CT: Greenwood Publishing.
- Kane, M. T. (2002). Inferences about variance components and reliability-generalizability coefficients in the absence of random sampling. *Journal of Educational Measurement*, *39*(2), 165-181.
- Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement* *50*(1), 1–73.
- Kang, O., Rubin, D., & Kermad, A. (2019). The effect of training and rater differences on oral proficiency assessment. *Language Testing*, *36*(4), 481–504.
<https://doi.org/10.1177/0265532219849522>
- Kim, S., & Lee, H. K. (2015). Exploring rater behaviors during a writing assessment discussion. *English Teaching*, *70*(1). DOI: 10.15858/engtea.70.1.201503.97
- Knoch, U. (2009). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing*, *26*(2), 275–304. <https://doi.org/10.1177/02655322208101008>
- Knoch, U., & Chapelle, C. A. (2017). Validation of rating processes within an argument-based framework. *Language Testing*, 1–23. DOI: 10.1177/0265532217710049
- Kondo, Y. (2010). Examination of rater training effect and rater eligibility in L2 performance assessment. *Journal of Pan-Pacific Association of Applied Linguistics*, *14*(2), 1–23.
- Krippendorff, K. (2004a). *Content analysis: An introduction to its methodology* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.
- Krippendorff, K. (2004b). Measuring the reliability of qualitative text analysis data. *Quality and Quantity*, *38*(6), 787–800. doi:10.1007/s11135-004-8107-7
- Krippendorff, K. (2004c). Reliability in content analysis: Some common misconceptions and recommendations. *Human Communication Research*, *30*(3), 411–433.
doi:10.1111/j.1468-2958.2004.tb00738.x

- Krippendorff, K. (2011). Computing Krippendorff's alpha reliability. Retrieved on Sep 14, 2018 from http://repository.upenn.edu/asc_papers/43.
- Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, 28(4), 543–560.
- Linacre, M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement* 3, 85–106.
- Linacre, M. (2014). *A user's guide to FACETS v 3.71.4*. Chicago, IL: Winsteps.
- Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication research: An assessment and reporting of intercoder reliability. *Human Communication Research*, 28, 587–604.
- Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. New York: Peter Lang.
- Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing*, 15(2), 158-180.
- Marcoulides, G. A., & Ing, M. (2014). The use of generalizability theory in language assessment. In A. J. Kunnan (Ed.), *The companion of language assessment*. New York: John Wiley & Sons, Inc. doi:10.1002/9781118411360.wbcla014
- McNamara, T. (1996). *Measuring second language performance*. Essex, UK: Addison Wesley Longman Ltd.
- Moore, C. T. (2016). Package *gtheory*. Retrieved on Dec 2, 2018 from <https://cran.r-project.org/web/packages/gtheory/gtheory.pdf>.
- Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using Many-Facet Rasch Measurement: Part II. *Journal of Applied Measurement*, 5(2), 189–227.
- Ockey, G. J. (2012). Item response theory. In G. Fulcher & F. Davidson (Eds.), *Routledge handbook of language testing* (pp. 316-328). London, UK: Routledge.

- Paltridge, B., & Phakiti, A. (2010). *Continuum companion to research methods in applied linguistics*. London: Continuum.
- Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing*, 13(2), 111–129.
doi:<http://dx.doi.org/10.1016/j.asw.2008.07.001>
- Plakans, L. (2009) Discourse synthesis in integrated second language writing. *Language Testing*, 26(4), 561–587.
- Plakans, L., & Gebril, A. (2013). Using multiple texts in an integrated writing assessment: Source text use as a predictor of score. *Journal of Second Language Writing*, 22(3), 217–230. doi:10.1016/j.jslw.2013.02.003
- Pollatsek, A., Reichle, E. D., & Rayner, K. (2006). Tests of the E-Z Reader model: Exploring the interface between cognition and eye-movement control. *Cognitive Psychology*, 52(1), 1–56.
- Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, 62(8), 1457–1506.
- Read, J. (1990). Providing relevant content in an EAP writing test. *English for Specific Purposes*, 9, 109–121.
- Reichle, E. D., Pollatsek, A., & Rayner, K. (2006). E-Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research*, 7(1), 4–22.
- Révész, A., Michel, M, & Lee, M. (2017). Investigating IELTS Academic Writing Task 2: Relationships between cognitive writing processes, text quality, and working memory. *IELTS Research Reports Online Series* (March 2017). Retrieved on Sep 7, 2018 from https://www.ielts.org/-/media/research-reports/ielts_online_rr_2017-3.ashx
- Roberts. L., & Siyanova-Chanturia, A. (2013). Using eye-tracking to investigate topics in L2 acquisition and l2 processing. *Studies in Second Language Acquisition*, 35, 213–235.
doi:10.1017/S0272263112000861

- Sawaki, Y., Quinlan, T., & Lee, Y.-W. (2013). Understanding learner strengths and weakness: Assessing performance on an integrated writing task. *Language Assessment Quarterly*, *10*, 73–95.
- Shi, L. (2004). Textual Borrowing in Second-Language Writing. *Written Communication*, *21*(2), 171–200. <https://doi.org/10.1177/0741088303262846>
- Shin, S.-Y., & Ewert, D. (2015). What accounts for integrated reading-to-write task scores? *Language Testing*, *32*(2), 259–281. doi:10.1177/0265532214560257
- Shohamy, E., Gordon, C., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *The Modern Language Journal*, *76*(1), 27–33. doi:10.2307/329895
- Smith, B. (2012). Eye-tracking as a measure of noticing: A study of explicit recasts in SCMC. *Language Learning & Technology*, *16*, 53–81.
- Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. *Studies in Immigrant English Language Assessment*, *1*, 159–189.
- Spivey, N. (1997). *The constructivist metaphor: Reading, writing and the making of meaning*. San Diego: Academic Press.
- Stratman, J. F., & Hamp-Lyons, L. (1994). *Reactivity in concurrent think-aloud protocols: Issues for research*. In P. Smagorinsky (Ed.), *Speaking about writing: Reflections on research methodology* (pp. 89–111). Thousand Oaks, CA: Sage.
- Suvorov, R. (2015). The use of eye tracking in research on video-based second language (L2) listening assessment: A comparison of context videos and content videos. *Language Testing*, *32*(4), 463–483. <https://doi.org/10.1177/0265532214562099>
- Toulmin, S. E. (2003). *The uses of argument (updated edition)*. Cambridge, UK: Cambridge University Press.
- Trites, L., & McGroarty, M. (2005). Reading to learn and reading to integrate: new tasks for reading comprehension tests? *Language Testing*, *22*(2), 174–210.

- UNESCO. (2004). *The plurality of literacy and its implications for policies and programs: Position paper* (Vol. 13). Paris, France: United National Educational, Scientific and Cultural Organization.
- Vaughan, C. (1991). Holistic assessment: What goes on in the rater's mind? In L. Hamp Lyons (Ed.). *Assessing second language writing in academic contexts* (pp.111–125). New York: Ablex Publishing.
- Watanabe, Y. (2001). Reading-to-write tasks for the assessment of second language academic writing skills: Investigating text features and rater reactions. *Unpublished doctoral dissertation*, University of Hawaii.
- Weigle, S. C. (2002). *Assessing writing*. Cambridge, UK: Cambridge University Press.
- Weigle, S. C. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing Writing*, 9(1), 27–55.
- Weigle, S. C., & Parker, K. (2012). Source text borrowing in an integrated reading/writing assessment. *Journal of Second Language Writing*, 21, 118–133.
- Wette, R. (2010). Evaluating student learning in a university-level EAP unit on writing using sources. *Journal of Second Language Writing*, 19(3), 158–177.
- Winke, P., & Lim, H. (2015). ESL essay raters' cognitive processes in applying the Jacobs et al. rubric: An eye-movement study. *Assessing Writing*, 25, 38–54.
doi:<http://dx.doi.org/10.1016/j.asw.2015.05.002>
- Winke, P., Gass, S., & Sydorenko, T. (2013). Factors influencing the use of captions by foreign language learners: An eye-tracking study. *Modern Language Journal*, 97, 254–275.
doi:[10.1111/j.1540-4781.2013.01432.x](https://doi.org/10.1111/j.1540-4781.2013.01432.x)
- Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. *Assessing Writing*, 17(3), 150-173. doi:<https://doi.org/10.1016/j.asw.2011.12.001>
- Xi, X., Mollaun, P. (2006). Investigating the utility of analytic scoring for the TOEFL® Academic Speaking Test (*Research Report RR-06-07, TOEFLIBT-01*). Princeton, NJ: Educational Testing Service.

- Yang, H.-C., & Plakans, L. (2012). Second language writers' strategy use and performance on an integrated reading-listening-writing task. *TESOL Quarterly*, *46*(1), 80–103.
doi:10.1002/tesq.6
- Yi, Y. (2010). Adolescent multilingual writers' transitions across in- and out-of-school writing contexts. *Journal of Second Language Writing*, *19*(1), 17–32.

APPENDIX A: WRITING PROMPT USED IN THE EYE TRACKING SESSION

This test assesses your ability to summarize, synthesize, and evaluate information presented from different sources and to state and support your arguments with sufficient details and examples in standard English.

You have two texts written about **genetically modified (GM) food**, food that is the result of a laboratory process where genes from the DNA of one species are extracted and artificially forced into the genes of an unrelated plant or animal.

Task 1 (15 minutes): Write a summary comparing and contrasting the viewpoints in the two texts.

Task 2 (30 minutes): Write a four-paragraph essay *indicating whether you think that genetically modified food should be supported*. You should use information from the two texts and your own experience to support your views (300 – 350 words).

Notes:

1. Be sure to give credit to the author when you use ideas or examples from the texts.
2. You should demonstrate your ability to summarize or paraphrase the ideas in the text. Do NOT copy word-for-word from the texts.
3. You may NOT use a dictionary.
4. Your composition will be evaluated on development of ideas, organization, and language, including grammar and expression.
5. You have 5 minutes to read the texts before you start.

TEXT 1: Genetically modified food: saving lives, or lining corporate pockets? (Tom Chivers - *The Telegraph*, 2014)

Genetically modified (GM) crops have not overcome widespread resistance mostly because the industry is tightly controlled by biotech companies. That is, the real problem is that genetic engineering is hurting the poor. It makes cotton cheaper to grow for highly subsidized American producers, further undercutting the price of cotton and forcing West African producers out of business.

Major biotech companies have no financial interest in developing them for African crops -- and tightly control the technology. The methods of transferring genes were developed by universities, but companies now hold the licenses. The companies permit others to do research with the technologies but want control over any product commercialized as a result. Several poor nations are trying to develop improved versions of local crops, but these efforts have been damaged by the companies' control over the technology.

In fact, the companies which develop GM technology will have unprecedented power over the food chain. They have a clear battle-plan to achieve their goal of 'consolidation of the entire food chain': an aggressive patenting regime, patenting technologies and genetic material. Academic work has shown that how well people are fed is less to do with the actual quantity of food available in the world, and more to do with who controls the food chain, and how well the food is distributed. GM, and the ability to patent GM technology, place far more power in the hands of major companies.

As a result, there would be fewer competitors in the market. These biotech companies might also have more political power and might be able to influence safety and health standards.

TEXT 2: The Deadly Opposition to Genetically Modified Food (Bjorn Lomborg – *Project Syndicate*, 2013)

Three billion people depend on rice as their staple food, with 10% at risk for vitamin A deficiency, which causes 250,000 to 500,000 children to go blind each year. Of these, half die within a year. A British medical study estimates that, in total, vitamin A deficiency kills 668,000 children under the age of 5 each year.

Yet, despite the cost in human lives, anti-GM campaigners have denied efforts to use golden rice to avoid vitamin A deficiency. Indian environmental activist, Vandana Shiva, called golden rice "a hoax" that is "creating hunger and malnutrition, not solving it."

The NY Times Magazine reported in 2001 that one would need to "eat 15 pounds of cooked golden rice a day" to get enough vitamin A. However, two recent studies in the American Journal of Clinical Nutrition show that just 50 grams (roughly two ounces) of golden rice can provide 60 percent of the recommended daily intake of vitamin A. They show that golden rice is even better than spinach in providing vitamin A to children.

Opponents maintain that there are better ways to deal with vitamin A deficiency, saying that golden rice is "neither needed nor necessary." They call for supplementation (vitamin pills) and fortification (adding vitamin A to staple products), which are described as "cost-effective." However, this is not a sustainable solution to vitamin A deficiency. And, while it is cost-effective, recent published estimates indicate that golden rice is much more so. Supplementation programs cost \$4,300 for every life they save in India, whereas fortification programs cost about \$2,700 for each life saved. Meanwhile, golden rice would cost just \$100 for every life saved from vitamin A deficiency.

APPENDIX B: RATING RUBRIC

	B	B+	C/D	C/D+	Pass
Organization (30%) Unity Cohesion Coherence Relevance	<ul style="list-style-type: none"> - Essay is somewhat organized and maybe hard to follow. - Cohesive and transitional devices are used rarely or inappropriately. - Essay includes some irrelevant details. Paragraphs lack a focus or purpose of each paragraph is unclear. 		<ul style="list-style-type: none"> - Essay is adequately organized but requires effort to follow. - Simple cohesive and transitional devices are used. Some might be used inappropriately. - Paragraphs display some evidence of unity but some redundancy and irrelevant information. 		<ul style="list-style-type: none"> - Essay is well-organized and easy to follow. - A wide range of cohesive and transitional devices are appropriately used. - Paragraphs display unity and a clear focus despite occasional redundancy and irrelevant information.
Arguments & Details (25%) Supporting details Relevance	<ul style="list-style-type: none"> - Arguments are vague or underdeveloped. - Details and examples are mostly clear or relevant but need more explanation. - Sources are not integrated. 		<ul style="list-style-type: none"> - Arguments are mostly developed although more supporting details are needed. - Details and examples are mostly clear and relevant. - Sources are integrated though not skillfully. 		<ul style="list-style-type: none"> - Arguments are fully elaborated. - Details and examples are clear, sufficient, and relevant to the topic and task. - Sources are mostly integrated skillfully.
Grammar and Lexis (30%) Range Accuracy Appropriateness	<ul style="list-style-type: none"> - Writer shows control of simple grammar structures and attempts more complex ones. - Vocabulary is used with many repetitions or mostly based on the stimuli. Many words are not appropriately used. - Essay contains many grammatical and lexical errors which interfere with comprehensibility. 		<ul style="list-style-type: none"> - Writer displays good control of simple grammar structures and some complex ones. - Writer uses a range of vocabulary with some repetitions. Some words/phrases are not appropriately used. - Some grammatical and lexical errors might occasionally interfere with comprehensibility. 		<ul style="list-style-type: none"> - Writer uses a wide range of grammar structures and vocabulary appropriately and accurately despite some minor grammatical and lexical errors.
Conventions (15%) Spelling Paraphrasing Citing external sources	<ul style="list-style-type: none"> - Essay contains some misspellings which sometimes interfere with comprehensibility. - Text from the sources is used without paraphrasing. - Sources might not be cited. 		<ul style="list-style-type: none"> - Spelling is mostly correct. Spelling errors are minor without interfering with comprehensibility. - Text from the sources is mostly paraphrased appropriately. - Sources might not be cited. 		<ul style="list-style-type: none"> - Spelling is mostly correct. Errors are minor without interfering with comprehensibility. - Text from the sources is paraphrased appropriately. - Sources are cited appropriately.

APPENDIX C: RATER INTERVIEW PROTOCOL

Part 1: Background Information

1. Tell me about yourself: your major, previous education, L1, years living abroad, etc.
2. Tell me about your teaching experience.

Follow-up

- a. Have you ever taught any ESL writing classes at ISU? For how long?
 - b. How about first-year composition classes? For how long?
 - c. What other classes have you taught? For how long?
3. Tell me about your experience as a rater.

Follow-up

- a. Have you worked as a rater? For what tests?
- b. Have you rated for the EPT previously?
- c. How long have you been a rater for the EPT?

Part 2: Rating Scales

4. What do you think about the rating scale?

Follow-up

- a. How is it useful or not useful for you to grade the essays?
- b. What do you think about the scale in terms of language? Descriptions of the band scores?
- c. What do you think about the number of score bands?
- d. What do you think about the weighting of each category?
- e. How do you think the scale should improve?
- f. Is there anything that you find problematic? How would you improve it?

APPENDIX D: TRANSCRIPTS OF RATERS' INTERVIEWS

RATER 1

INTERVIEWER: First tell me a little bit about yourself. Like your education and background or where you come from, your first language. Like your teaching experience, your rating experience.

RATER 1: So, I'm from [country]. My L1 is [language] and my teaching experience is about 3 years for all ages but mainly in higher education, undergraduate and I teach English. I also taught TOEFL and IELTS.

But I've never rated prior to studying for my MA at [university]. And then during my MA, I did a little bit of rating training for the OECT at [university]. That was my experience rating, and then for my PhD I did quite a bit of rating for the speaking and the writing for the EPT.

INTERVIEWER: Ok, so how long have you been rating for the EPT writing?

RATER 1: Just once, more for the speaking. I did the speaking twice.

INTERVIEWER: Was that your first semester last time?

RATER 1: Yeah, last time. I never took the EPT myself because I had higher enough IBT scores that I didn't need to take any EPT or OECT.

INTERVIEWER: So what do you teach here?

RATER 1: I only taught one semester at [university]. This semester I'm only a research assistant. Last time I taught Critical Thinking and Communications. But mainly was in the form of written work. That's why I think I talk a lot about arguments. Presenting a good enough argument to make your text convincing.

INTERVIEWER: Was that 150 or 250?

RATER 1: It's 150.

INTERVIEWER: What do you think about the rating scale?

RATER 1: Definitely (useful). You need to identify like the criteria of a certain band. You can't just assign like B randomly. So definitely, when I was grading I would go and make sure that ok, just like some checklist thing. The grading scale itself is very important and very helpful.

But because this is a placement test it's nice to know what they are actually getting placed into. If you get what I mean. For example, B. If they're placed into B what are they learning?

INTERVIEWER: Oh so you need to know the curriculum of these classes?

RATER 1: Yeah, I feel like... because it's very different from a TOEFL IBT and IELTS. You're not placed into a class but it's more of a generic "are you ready for academic work?" it's a little but easier if you like but with this you're actually putting them into either a B or a CD, right? And I feel like I want to know if they're actually learning and I hope this would help in the decision whether

they're going to get a B or a CD. If I think based on their text production writing they would benefit from instruction and B I think that would help.

INTERVIEWER: Ok. What do you think about the scale in terms of language, descriptions of the band scores?

RATER 1: I have a hard time with the "many" and "some", "many", "some", "mostly". But if I try to visualize it as a certain number it can help sometimes. But that's totally on my part how I interpret "many". But yeah, in general, I think it's pretty easy to follow.

INTERVIEWER: Ok, how do you think the scale should improve?

RATER 1: So my question is like: what if it's worst than B. Of course you still put them in a B. I don't know really.

Probably pay more attention to the "many", "mostly" the "some" are used. Those worlds. Probably a number would be more appropriate. Probably a number would help visualize it cause like, I don't know. Like "adequately". What does "adequately" look like? But then with the training right? That helps.

INTERVIEWER: What do you think about the number of score bands? We have 5 right now. What do you think?

RATER 1: Right so I think the distinction of a B and a B+ isn't that useful. The position that I find myself in is the CD and the CD+. It's not quite a pass but it's not really a CD. That's the situation that I find myself in most of the time. B and B+, I don't really know how that would be useful, especially if they're going to go into the same B class. A CD and a CD+ might be a little but more important because sometimes you just can't give a person a pass. Like no no no. But they're not also that bad, I feel like.

INTERVIEWER: Do you think it is better to use labels or numbers for the score bands? Well if you think about this in terms of numbers like B is 1, B+ is 2, CD is 3, CD+ is 4 and D is 5.

RATER 1: Probably that would make knowing what course they're going to put into irrelevant right. Make it more like, less burden thinking process and constantly second guessing your choices. Cause if feel like I always do that. Am I being too lenient or am I being too harsh? So, I think it would be... probably a number would be better. Like the EPT speaking they use a number.

INTERVIEWER: What do you think about the weighting of each category?

RATER 1: I agree that argument and organization should have more weight. I feel like that the two components I place a heavy emphasis on is organization and arguments and details because grammar and lexis and conventions I think are pretty easy to teach. They're pretty easy to learn cause they're very... there's like a certain rule. We have rules vocabulary, grammar, convention citations. I feel like students when they try hard that can be easily acquired. But with organization and arguments those are two important concepts. Because they're really hard and it take time to, especially in academic writing, to sound natural

in an academic way and I don't know if that makes sense. Definitely I'm very harsh, if I don't see good organization, if I don't see the good arguments even if you're vocabulary and your grammar is good and you paraphrasing is good you can never get a pass from me if you're organization and arguments are not good...

INTERVIEWER: Anything else?

RATER 1: I think the ability to integrate source texts is very important, especially in task 2 because I feel it's very clear in the prompt that it wants you to summarize and incorporate your experience or any background knowledge not just rely on the summarizing part. Because then what's the difference between task 1 and task 2. So, I do grade pretty harshly if I don't see integration or I see integration but no paraphrasing. No paraphrasing or citing, because they can paraphrase but they didn't cite. So perhaps we should have more weight for it or maybe we should put it in a separate category.

RATER 2

RATER 2: So I've lived here for the last 20 years. And I'm currently a master's student in the applied linguistics program.

And regarding teaching prior to coming back to [university] I taught in middle school. It was a guest teacher for a non-profit organization and I did substance abuse prevention classes. And then for [university] last semester I taught English 150 and this semester I'm teaching English 250.

And as far as my experiences as grader; I helped grade papers for the going Spring 2018 semester.

INTERVIEWER: So have you ever taught 101B, C, D anything like that?

RATER 2: Nope, I also have do language exchange. So I'm in the process of learning Korean so friends of mine who want to work on their English we do a partnership where they help me with my Korean and I helped them with their English. I also had tutoring sessions with middle school or high school in which they didn't help me with my Korean, I just helped them with their English.

INTERVIEWER: Ok, so what do you think about our grading scale?

RATER 2: Overall, I think the scale was very useful. I actually did find the categories (B+/C+) very useful. One thing that I found when I was doing the grading, I felt like I had more options. So those essays I was like "I don't know if this is a C or if it's a Pass, it's a C+" and similar with "I don't know if this is a B or if it's a C, it's a B+". And so in that case it was very nice. Now I still had a few that I was debating "is this a B+ or a C?" at times. I don't think you can have enough categories I think is my point. And so ultimately a decision has to be made. But I liked it, I liked having the 2 extra categories for those cases... because they were also essay where I was like "nope this is a B, nope

this is a C” but for others where I was like “I don’t know” it was in the in between it was nice having that category to utilize.

INTERVIEWER: Do you think that the weight of the scale’s different categories are good now? We have 30 and 25 and 30 and 15 and do you think that should be changed?

RATER 2: I don’t know. Like I said the conventions, I think the 15 is appropriate. I personally would have put more weight into arguments and details. But then again I’m not sure which one I would lower to raise that one because I do think that the others are important.

INTERVIEWER: What do you think about the wording of the descriptors?

RATER 2: I had no trouble with that.

INTERVIEWER: How do you think the rating scale should be improved?

RATER 2: I’m debating on this one but one thing that came to mind was... and maybe this falls under arguments, I’m not sure. But one thinking both when I was grading at the begging of January and when I was going through this process was... I’m not sure how much the writer’s comprehension of the text factored in. And I feel like that should play a role because if for whatever reason... and I’m debating because it might be that they’re not expressing themselves properly, which might be a thing of itself anyway. Just my though process is if someone is reading a text but doesn’t understand it and then gives and argument and then we ignore the accuracy of the idea or the argument, but the language itself was good, so we’ll give them a Pass. When they come to say me when I’m teaching 150, they’re in trouble because then when I’m grading their paper it will be “no this isn’t what the material is saying”. And so that was something I was thinking about when I was reading and I’m not sure how you would do it but that was my thought was some way of measuring or evaluating their comprehension of the text.

INTERVIEWER: That’s an interesting thought, I think that is relevant to source-based writing. So, anything else that you would like to say?

RATER 2: No that was really the big one. At least that I had frequently coming up. And it’s not a lot of the students. It’s just a few that I’m like “I don’t think that you understood what you read”. But I’ve also had native speakers not understand texts. So, part of me is “I don’t really want to hold things against non-native speakers”. I guess one thing I thought too when I was going through this process... and I’m sure we would never do it because it’s too big of a hassle. I would be interested to see if you would incorporate native speakers writing into this. Kind of just to see... almost like a control group. But the graders don’t know these are native speakers’ texts and see what I grade this a Pass or a C, D or a B. Because there are times where I just wonder if maybe we would be surprised what we would find. Like I said that would be a nightmare to implement.

INTERVIEWER: Thank you for sharing your thoughts about the EPT grading scale and I appreciate your time.

RATER 3

RATER 3: So I think I'm going to talk about mostly my teaching experience. So I went into my masters; I took English 500, I taught 150, I taught 250 and I taught cross-cultural 150 classes because I did the TESL masters that was my primary focus. How can I work with international students in the composition classroom and that's even what my thesis even sort of looked at. It's the transition from 101C to 150. Also during my masters I did teaching IEOP, so I had experience working with those students; I taught at like advanced levels and also introductory level. Which was a really fun adventure.

INTERVIEWER: What did you teach in IEOP level?

RATER 3: I taught reading and writing. So then in my masters my experiences came from teaching reading and writing essentially. So then I was a lecturer for a year and I taught 150, 250 and then 314. And I think I had a cross-cultural section of 314. So then what I can see then clearly is when a rate essays I'm thinking about where the students are going and I have not taught 101B or 101C. I'm generally familiar with hitting the purpose of the course but I'm sure then its interesting to consider how I'm thinking of this maybe versus some of my peers who are teaching or who have taught 101B and 101C because again I'm really focus on where the student is going. So then after I was a lecturer I taught in Turkey for 2 years and I worked in the language prep program. The university is an English medium school but the students I mostly worked with high level students, they were up or intermediate and the TOEFL score to get into that university was a minimum of 80; so it's 10 points higher than it is to get here. Is it 79? Then maybe it was 89 there, cause I remember thinking the score for that university, the TOEFL score, was higher than here at [university] but I would have to go back and double check that.

So now I'm a PhD student and in the fall I taught 314 and now I'm mostly working with graduate students.

But sort of this trend I see as a teacher I prioritize organization over anything else when I teach writing. Part of it is because I don't feel like a have a strong control over English grammar, an understanding of that. And that to me is really secondary in teaching writing because if students don't understand how to organize their essay and set up their argument ultimately we're not even going to get to language if we can't work on the organizational aspects. And it's not that individual linguistic features aren't important but I just don't really care. That's not where I get excited or where I feel comfortable teaching maybe.

INTERVIEWER: Thank you. So have you taught 180?

RATER 3: No, I've never taught 180 or 101D. I'm working now as an English writing consultant; so I work with graduate students in that capacity.

INTERVIEWER: So have you ever been a grader for EPT? You have obviously but did you grade for us before that?

RATER 3: No, so the first time I ever grade EPT was this January 2018.

INTERVIEWER: But have you worked as a grader somewhere else?

RATER 3: Yes, in Turkey I worked as a grader because we all had to grade and we offered an exam 3 times a year. So then I would have worked as a grader 5 different times; I did both. So originally the first year I was speaking and then they moved me to writing.

INTERVIEWER: Let's talk about our grading scale then. So what did you think about our current grading scale?

RATER 3: I understand why the scale is set up that way. I would say it's hard as a grader to be limited to 3 categories because sometimes I feel like an essay has elements of maybe of a C and elements of a B. By understanding what that means, like placing the student in a B level. I mean that is a whole year of extra English courses and I think maybe grading an essay is more complicated than just having 3 levels. It'd be nice if there could be more flexibility, which is, why it was interesting for me to include like a B+ or a C+.

INTERVIEWER: So this is the grading scale that we used in our study. What do you like about it or what do you don't like about it?

RATER 3: I think the rating scale is very useful in helping me decide what grade I should give a test taker. So what I do like is between B, C, D and Pass I can see sort of how these different levels are differentiated from "somewhat organized" to "adequately organized" to "well organized". So that language to me is very clear, although I can't always necessarily say, "what does somewhat organized mean?". I always want to quantify it in some way, I don't know how you can do that but at least this rubric really works to differentiate between those different levels. Like "arguments are vague", "mostly developed", "fully elaborated". I mean that clearly gives me sort of a way to think about how I should grade this essay, there's not sort of confusion between what a C, D is or isn't and what a Pass should be. At least based on the language of the rubric.

INTERVIEWER: What do you think about the weighting for the categories? So we have numbers here.

RATER 3: I think that the numbers (%) are useful cause without numbers I think it suggests that all of these should be weight equally but clearly they're not. I mean I would agree that organization probably should have a little bit higher weight than say spelling. But I mean the rubric here does indicate that organization, grammar and lexis should have an equal weight. So it is telling me that I need to consider that, now where I actually do consider that or not is another story but I do think that having the percentages there is helpful to sort of guiding us.

INTERVIEWER: Would you change the weights if you were to revise the rubric?

RATER 3: No, I don't think I would change the weight because while I prioritize organization in an essay... when we're thinking about how we're setting our students to be successful if we don't prioritize grammar as much as organization, I think we're setting our students for failure. And I don't mean that as an insult to these international students who are coming in, I want them to be successful and I'm assuming that a lot of the are very frustrated that they're placed in 2 extra semesters of English courses but we're not doing that to be unkind, we're doing that to make sure they get the support they need. So that when they're in their classes they can be successful. That's at least how I see it, so yes. I mean if a student demonstrates that they don't have a strong control of grammar then an extra semester of grammar should help them. I mean that percentage should be present in the rubric.

INTERVIEWER: Is there anything that you find problematic? How would you improve it?

RATER 3: I don't think that there's anything problematic that stands out to me.

I think I would put... I think the citing external sources gets a little bit lost. I wonder if this would put up in arguments and details, which I know seems a little bit strange. Because for me paraphrasing and citing external sources is important. Spelling I don't know, I don't really consider spelling that much because that I almost group in grammar and lexis. Because I'm wondering if spelling is actually related to vocabulary. I wouldn't put spelling with paraphrasing and citing external sources. I'm not a really good speller so I don't want to be judged on it. That to me would really not have a huge impact on a student's ability to be successful because they can't spell. They'll figure that out. So that's sort of what I consider. Even if it's just at the bottom of the page I feel like paraphrasing and citing almost gets forgotten. And if you want that to be more important, I'm wondering if that was actually put in with argumentative details. If that would help because it's hard to separate our supporting details from citing and paraphrasing.

INTERVIEWER: I think there are some descriptions here in Arguments and Details talking about citing.

RATER 3: Yeah, and I maybe that's where I get confused because there's a mention here of sources not integrated but then there's another category when sources are mentioned. I think it would make more sense if they were together.

INTERVIEWER: Thank you

RATER 4

INTERVIEWER: Could you talk a bit about yourself, where you come from, your study, your teaching background...

RATER 4: I'm a second MA student in the TESOL/AL. I'm a native speaker of English. I worked for 2 years as a tutor for Chinese students in the US. In the last year and a half, I've been teaching first year composition 150.

INTERVIEWER: So you've never taught 101B or C?

RATER 4: No. I would like to teach ESL. The students that I tutored were ESL students, but I taught them one on one with them. So I'm familiar with it, but haven't done it with 20 students at the same time.

INTERVIEWER: What kind of rating experience do you have?

RATER 4: So last winter, I rated the writing and I rated speaking. I did it once for each.

INTERVIEWER: What do you think about the rating scale?

RATER 4: I think because my past experience only has three levels, so I think that threw me a little bit because I have to think about in-between levels. I don't think it's a problem. It just takes some time to get used to. I know it was hard for me with three levels. I think I got used to that. I think having the in-between levels is good. I don't think the change is bad. I think I just personally have to adjust.

INTERVIEWER: So, would you rather have a 3-point scale?

RATER 4: I think 5 is better.

INTERVIEWER: How do you think the scale is useful or not useful for you when you rated?

RATER 4: I think it is useful because it is not overly complex. 5 is still not a lot. But if we have too many choices, then it is more difficult.

I think the different descriptions in the different levels are pretty good. Some other rubrics I used in the past are a lot harder to use. They always say too much under each category, so maybe I agree with one of the things in there but not the others. So by only say a few things about the writing I think it helps you know because you're not going fit 100% in the categories.

INTERVIEWER: How do you suggest we should improve the rubric?

RATER 4: I can't think of any improvement. I think that one... It's not exactly the same thing but for somebody who is familiar with the classes that students will take potentially based on these ratings, maybe I think too much about the classes they are going to take. Maybe it's good, maybe it's bad, I'm not sure. But somebody who's more impartial maybe thinking the class placement is not helpful.

INTERVIEWER: What do you think about the descriptors of the categories?

RATER 4: Yeah. I like them. As I said before, they include enough details without including too much. Too much is confusing and not enough is also confusing. More information could be.... Maybe I shouldn't do this but I do think about this whether I like it or not. I subconsciously compare with the students in my 150 classes and I think "how good are they at integrating sources?" Even when they are a native speaker and a good writer, they might not be good. I don't think we should penalize students for that necessarily because they don't need to be great at this to take 150. They will improve when they take it. But in order to get out of the 101s, I don't think it's necessary.

INTERVIEWER: What do you think about the percentages?

RATER 4: I think I like them. I referred to them when I made decisions. Like for example, the last category, like convention is weighted relatively light, which is my intuition too. It might as well how I feel about it. So I wasn't taking that category into such consideration. I think the weights are good. It can be helpful. I think in my intuitive rating, I rated grammar and vocab more strongly in my own mind. But then it comes back to what's out target, what's our construct. I think it could be weighted more, but grammar is relatively easy to fix. But organization maybe harder to learn. So if there's a well-organized paper but they misused the past tense throughout, I think it's easier to fix the past tense. But learning how to organize the paper is harder. SO I think we tend to want to weigh grammar more because it is easy to check , especially as raters, it's easy to say that's wrong, that's wrong, that's wrong. But scale-wise, I don't know. I think it's good what it is. In my experience teaching, organization is a much bigger issue for students of all L1 backgrounds. So I think often, it's not weighed enough. On this one, I think it's good.

RATER 5

INTERVIEWER: Can you talk about your teaching experience?

RATER 5: I have taught EFL in Gram school, after-school school. I was teaching mostly grammar, then then having them memorize vocabulary. I was teaching various subjects like English, Math. I have taught [language] to young students. And then I have taught [language] in Texas. And now I'm teaching ESL. 101B and C.

INTERVIEWER: Talk about your rating experience for English tests.

RATER 5: I just have one experience of rating EPT test. I actually was asked by [name] to rate the writing in the EQUAL corpus.

INTERVIEWER: What do you think about the rating scale that we use?

RATER 5: We have to rate task 1 and task 2 with the same rating scale, rights? That's one concern that I have now. For task 1, they do not have to develop arguments. They can just summarize the texts. So basically, this part seems to be not relevant to task 1. Given that I do not have a lot of experience in academic writing, I think this is good, like a beacon for me.

INTERVIEWER: What do you think about the descriptors? Anything that should be changed or improved?

RATER 5: I don't think so. It's straight forward. But source integration is the one thing that I'm not sure about. I usually looked at if the students cite the sources correctly. In this case I only found one text which misinterpreted the source texts. In that case, I usually penalize these people. But as long as test takers say something about the source texts, I will consider it integration about the source texts. But I don't know how integrative it is; so, that's my issue as a rater.

INTERVIEWER: What do think about the percentages for the categories?

RATER 5: They're not really useful in the sense that I don't focus on these numbers. So, this (the score band) is categorical right, so we do not score 60, 70. If we were to score them, I think this is beneficial. But because we just classify the essays into these categories, I personally do not care about these numbers.

INTERVIEWER: What do you think about the number of levels?

RATER 5: I have two views about that. From the testing perspective, I think this is good. From the teaching perspective, I think students in 101C can survive the college life in terms of the writing ability. So, even without taking 101C, I guess they can survive the 4 years at [university] whereas in 191B they definitely need training. But for 101C, there are some students who I think can survive without taking 101C. So from the teacher perspective, I wouldn't say that 101C is not necessary, but based on my limited teaching experience, 101C students can survive the college but 101B definitely they need to take the class. When I classify an essay into C, I usually compare C and Pass. So if students cannot achieve Pass, then they go to C. So I think even without Pass, I think some students can survive. I don't know. So I think my opinion might be a big issue. I personally think the scoring can be binary, pass or fail.

INTERVIEWER: What do you think about the training materials?

RATER 5: It's helpful. Because I did not teach 101B and C here, it gives me a better sense of what is a B and C student. That helped me to create my intuition about students' performance. It helped my confidence in rating.

INTERVIEWER: What is the most useful feature?

RATER 5: Maybe writing samples with explanation why it should be a B or C.

INTERVIEWER: What do you think should be improved?

RATER 5: I think I'm happy. It's a good balance between the training needed and the time we can spend on it.

RATER 6

INTERVIEWER: Could you talk a little bit about yourself, like your education, your education background

RATER 6: My name is A. I got my BA in English literature in [country]. And then after being done with the BA level I taught 3-4 years English to different level. I started with kids then I taught adults such as 50-year-old women or young adults, different range of people. I also taught English to elementary students at school; but my main focus was teaching English at private instructions in [country]. And then I applied for a master program in [university], I came here in 8/2015 then I got my master then. There I also TA-ed and taught... There they have CIEP which is cultural intensive English, something like that. So, I taught that. They are international students who are not students at NAU but they want to. So, they are mostly F2 students, dependents of somebody who is

a student there. But there is not much of a difference, so ESL stuff. And then after graduating, I got accepted here; so, came here in 2017. And here I TA, I taught 101B and C. It's been 1.5 years. It's my third time teaching these classes.

INTERVIEWER: Have you taught 150, 250?

RATER 6: No.

INTERVIEWER: How long have you been rating for the EPT.

RATER 6: 1.5 years. 3 times.

INTERVIEWER: What other rating experience do you have?

RATER 6: I had to grade mid-terms and final terms.

INTERVIEWER: So, what do you think about the rating scale?

RATER 6: It's good, detailed. Very useful. Actually, I most like the portion like organization and grammar, I think. It's good because they are important things in student essays. But I kind of don't like the C/D level in between. So, most of the students fall into the C/D level. So, it seems that the B level is you really really should work on your writing and the Pass is you are really good you know everything, at least the basic things so that you can move to 150. But CD level is kind of not very clear to me. I just feel like this from last year. Because what I did now is that I just place many students in CD level, because it's not P, it's not B. So it's CD.

INTERVIEWER: So do you think having the pluses levels helps?

RATER 6: But there's no other classes for CD pluses, right? They are all together right? So I have this problem. In my class I have better students. They are very good, but you cannot say they are Pass. And we have weak students, but they are not B. They are better than B level. But they are all in the same class, so the better students are usually very bored, or tired, and they don't care that much of what I say. But their writing is good. So, most of the time I give them 90 out of a hundred or I give them some suggestions. So they are not exactly at the same level. In the B level, I have experience for only one class. It was a good class. Most of them are at the same ability. I think I only had 2 or 3 students that I think maybe they should be placed in CD level. They CD level has a mixed of everything.

INTERVIEWER: So, you think that the CD level should be divided in terms of classes?

RATER 6: I think the B+ can go to CD, and there should be another level for CD+. Even you have the same materials for CD, they should be divided just so that the students do not feel bored. I think in CD+ we can do more stuff. We can do videos from different websites. But with the CD, you can do the regular ones, like the materials and textbooks.

INTERVIEWER: So, let's talk about the descriptors.

RATER 6: Actually, once you get used to it, it's very good. When you're teaching those levels, it's understandable. It's very clear. You know what the rubric says. I think maybe some of these explanations need more examples. Or maybe not. I don't have any specific opinions.

INTERVIEWER: What do you think about the number of levels?

RATER 6: I think we only have 101B and C. I think the problem is B+/CD+. Maybe write some description for the plusses to say what they actually mean. As I say, B+ can go to CD. Because B is very clear. But B+ is something in the air; so, they can join CD+ classes IF we have CD+ classes.

INTERVIEWER: What do you think about the percentages?

RATER 6: Yes, I agree with them. I think organization and grammar are important things, so they might have larger percentages. I don't pay too much attention on their use of source text because by reading the essay, you can tell his or her ability. I actually haven't thought about what happen if the student use to language from the source texts. I think we are not going to evaluate how they summarize the text. We are looking for if students can write, I think. If they misunderstood the information and they wrote it in the essay, but everything is good, then I think it's ok. But if it happens in the real class, then I won't give them full credit. But it's different in placement test, I think. In placement test, you look at what the student's proficiency, even if they misunderstood the source texts. I'm talking about minor misunderstanding. If they misunderstood the whole general thing, then it's wrong. But if it's just a minor thing or if they don't understand some words and they think it's ok. But if they misunderstood the whole thing, then it's a problem. Writing and reading are related, but you should see in what way you're using this relationship between writing and reading. And in a placement test, you can just look at how well the test is written.

INTERVIEWER: How would the rubric be changed for an integrated writing test?

RATER 6: Maybe we should include something like finding or using the main ideas from the texts. And then support them, something like response essays (what we are doing now). They can take the main ideas from what they read, and then write their reaction to some. So, they could find the main idea or a good example from the texts and react to them, that means their reading ability is good right? Or finding the overall idea of the text. Did they say something about that? Then reading ability is important because they should know what the texts are talking about, the main ideas or the general ideas.

RATER 7

INTERVIEWER: Could you tell me a little bit about yourself? Like where you come from, your first language, what you study, what you taught?

RATER 7: ok, I'm a PhD student in the [program] and I'm from [university]. And I speak [language], my native language. And I speak a little bit German. And back in [university] I used to teach academic writing classes, academic reading classes and the reintegrating those two writing and reading classes. And I used to

teach 101 level basic grammar courses back in [university] and I taught a couple of ESP courses. And then here I taught 101C and 101B classes for international students.

INTERVIEWER: What are you teaching now?

RATER 7: Now I'm teaching 101C, 2 sections.

INTERVIEWER: I've never had these classes.

RATER 7: Really? They're fun

INTERVIEWER: Yeah, I could imagine. So how long have you been a teacher?

RATER 7: I would say 8 years.

INTERVIEWER: Have you taught 101, 150, 250?

RATER 7: No

INTERVIEWER: Mostly you teach writing classes

RATER 7: Yeah. I work with EFL students and ESL students not American students.

INTERVIEWER: So how long have you been a grader for the EPT in general?

RATER 7: Since I came here, like been on the very first day. So it's been 2 years, not 2 exactly, one and a half years.

INTERVIEWER: So when you came was this a test the we used? Or was it a different one?

RATER 7: It was not computer based, it was a paper based with the same format. So I taught one test for summary and essay.

INTERVIEWER: I see, so what do you think about the current EPT grading scale? We have 3 score bands

RATER 7: After you start grading the papers you feel like... you think about your own teaching experience like is this a good student for one of my B classes or my C classes? So I was always comparing the papers that I'm grading to my own students mostly. It is because there's a big difference between a C student and a Pass student. So I have to... since I feel like I have to put that student into one category I always think about my own students. So it was like another grading scale to me. I felt like that this is not enough because I can't decide because to me between B and C I kind of can't understand the difference between them because... for a student to be a B. You can easily identify if that student is a 101 student. For C and Pass there's a big difference.

INTERVIEWER: In your class?

RATER 7: The people I used to rate so I would like... Ok I have to pick one but I can because when you look at the grading scale it can go either way. So, it fits well with for example in one category and it fit well with another category another say pass. So I was always in between deciding so that's why I used to refer to my own teaching experience. in saying the rather that paper could be a good student in my class or not. So I felt that it's not really efficient so that's why in my mind I'm using another rating scale.

INTERVIEWER: So what about this grading scale? Do you feel like you have enough support to have you decide?

RATER 7: In terms of bigger categories definitely yes. It is very useful. The extra categories like B+ saved me a lot in terms of the papers that I was between two categories. I would like to see descriptions there, though. I mean it is good to see there is another category there. It gives an idea about what can the descriptors be, but I would just make use of those descriptors too to be confident.

INTERVIEWER: So how do you think the rating scale is useful or not useful when you grade? What do you see we could improve?

RATER 7: I do like the expressions in both because all the time I do it in myself. If I can't see any bold expressions, I just highlight myself. So it is good that there's is something there. Maybe the weight for each category. They might be... I don't know... there is a difference, for example I was always like "what if this student has more... fits very well and did 2 categories and is right about 2 categories. So should I be considering these rates? 50%? Should I be counting them?" so I'm not sure about that.

INTERVIEWER: So you think that that's useful?

RATER 7: I don't look at that because in my mind I know conventions are not really big issue and to me arguments and detail are more important than organization. So that's why I'm not really looking at the rates much.

INTERVIEWER: Should we remove the numbers at some point?

RATER 7: Might be, it doesn't affect my decision, but I just ignore them. Because I have in my mind that to me a paper should have a nice argument there other than just an essay format paper. So I just ignore them, doesn't bother me.

INTERVIEWER: Ok, you say the descriptors are good except for the fact that you need something here but the current ones do you think they're clear enough or is there anything that is to cage for you?

RATER 7: It is difficult to figure out what is the difference between "somewhat" and "adequately". I mean you can tell it like literally but when you look at the paper how can... it is difficult to understand "somewhat organizing" and "adequately organizing". The root level, that is a perfect definition, but it is difficult to reflect to the grading process. In this one for convention for example I was like I can't decide when one other paper when the spelling was important or not. So the spelling descriptor was exactly de same for C and D task. So if it's the same then there's is no need to put it there and if there is a difference then why is it the same?

INTERVIEWER: There is a difference for B C but spelling is quite the same for C D and Pass.

RATER 7: Normally I don't look at spelling much but in one of the papers the spelling error was really distracting, and it was too much. So I had to look at it maybe for the first time in my life. And then I noticed that it is exactly the same thing, so it really didn't help me. Other than that, it looks fine. I don't like

descriptors as a phrase, I like it as a sentence. It is more clear to me if you use a sentence instead of a phrase or a verb.

INTERVIEWER: So how would you change this rating scale?

RATER 7: I see, I would definitely add those extra categories, but I don't know if a student gets B+ what happens. I'm not sure about that. I mean if this grading scale like this format I would spend a lot of time to decide. Other than that, I think that the 4 categories are good.

INTERVIEWER: Anything else you would like to comment on the grading scale?

RATER 7: I didn't see anything other than that spelling one. I tried to look a lot at the rubric when I'm grading so I don't think there's anything.

INTERVIEWER: Anything else you want to comment on?

RATER 7: Maybe since we're doing EPT on the computer from now on, I'm not sure if we could send that factor

INTERVIEWER: Which one?

RATER 7: Typing

INTERVIEWER: As irrelevant?

RATER 7: So for example for that paper that I was talking. The spelling error. Should I consider spelling as that criteria because in real life all the environment is as real [something] spell function. Is to me it is becoming a little irrelevant because in real life spell check is always on so why are we doing it off during a test and if that is the case then why is spelling of the rubric. Because it's not going to be an issue anymore thank to technology. I don't know.... In the typing skills you can put it in the rubric but maybe in the training we can mention that it is a computer based and some people might not have any typing keyboard skills so that it might affect your overall paper quality.

INTERVIEWER: Or we need another study. Anything else you want to share?

RATER 7: No

INTERVIEWER: Thank you.

RATER 8

INTERVIEWER: Could you tell me a bit about yourself? Your educational and teaching background?

RATER 8: I'm a PhD student in [program]. I used to teach 101C, but now I'm teaching 101D for graduate writing class. My research interests are writing assessment, rater perception, and rater bias and how it's associated with the interpretation of the test scores. It's about validation, something like that.

INTERVIEWER: How long have you been rating for the EPT?

RATER 8: It's been about 4 years since I came here. I have a chance to rate the essays right before each semester.

INTERVIEWER: What other rating experience do you have?

RATER 8: When I was working at the secondary school in SK, I was asked to rate student speaking performance in the speaking English competition. It was me that designed the rating scale. But I think this was too basic thing. I wouldn't consider it rating experience. My real rating experience started when I was admitted to the PhD program.

INTERVIEWER: What do you think about the rubric that we use?

RATER 8: Yeah, I think it is very useful. It sets standards for test takers' performance and helps me consistent in my rating. It's easy to follow. I think it is very well-constructed. I think it reflects the distinct components of the writing ability of the examinees. I think it follows the previous studies in the multifaceted facets of writing ability. But I'm not sure about the weighting of each component and how the exact numbers for each rating criteria are decided because somehow I feel that the convention should have more attention. Because this is source-based writing, whether the examinee could paraphrase the original texts or not is a very important criterion to be considered.

INTERVIEWER: Did you look at the numbers when you rated? Did you take them into account?

RATER 8: A little bit, especially when I have conflicting decisions. For example, when I think they did well in convention, and grammar but not very well in organization. Especially, in the last two essays, because organization is 30%, so even though the students did well on other components, I didn't give them a pass but CD+.

INTERVIEWER: What do you think about the plus columns?

RATER 8: Yes, I think very beneficial because there are a lot of gray areas between the descriptors, especially for grammar and vocabulary and organization. The only difference between them is the adverb modifiers: some versus many, frequently, something like that. So, there are many continuums between these descriptors, right? And also, whenever I ran into some evidence that wouldn't fit into this or that category, that's when I chose the middle ground.

INTERVIEWER: Do you have any other suggestions?

RATER 8: I think this is good. I think the good thing about this rating scale is that I have something for the gray areas. Because I have to give a holistic score because of the nature of the rating, so it gives me some tough times. For example, this essay belongs to the B world, but it also belongs to the C world. How can I reconcile this decision with that decision, ok? Because adopting the analytic scale is the ideal situation, but it requires more man power, time. And also, I'm not sure if adopting the analytic scale is better than the holistic scale. But somehow I think the first three categories, organization, grammar and lexis, and arguments should be separated from convention because convention is a

distinct quality; so if we have more man power or time, how about rating the first three categories first, and then move to convention. It's just my opinion. By convention I mean source use. It should be a different category.

RATER 9

INTERVIEWER: So could you tell me a little bit about yourself? Like where you come from, education.

RATER 9: I'm a fifth year PhD student in applied linguistics program so I study language testing and I taught academic writing courses for undergraduate students for almost 3 years and 2 years for graduate students.

INTERVIEWER: So you taught 101C and 101D?

RATER 9: Right.

INTERVIEWER: Have you ever done 101B?

RATER 9: 101B no, I've never done 101B

INTERVIEWER: have you taught 150 or 250 courses?

RATER 9: No I haven't

INTERVIEWER: How long have you been a rater for us? For EPT

RATER 9: EPT... 4 and a half years

INTERVIEWER: Do you have any other grading experience before this?

RATER 9: Not a full time, from time to time I graded but it was not standardized

INTERVIEWER: And you said about writing, speaking?

RATER 9: It was writing but it was young learners, so it was quite different.

INTERVIEWER: So I'm going to ask you questions about the grading scale itself. So, what do you think about the grading scale?

RATER 9: I think this one is more useful than the previous one. Before, we had the 3-point scale. That's I think having 3 score bands... when we have some students on the bottom line and then their decision mostly done by which raters they have, depending on which rater they have I think it's increasing the rater errors. So, I think it's better to have 5 scales and then once we have those students on the bottom line then we can the final decision later. But with the 3 scale criteria I think we have many raters' errors. When I did rater training, it was useful but once I started grading, I just used my criteria. And only when I was not quite sure, I referred back to the evaluation rubric, but I didn't use that much.

INTERVIEWER: What do you think about the descriptors?

RATER 9: I think the language in the descriptors is clear. But my general intuition is that when we evaluate... I don't maybe it's just my personal thing probably but when we holistic score... this is a holistic score basically and then we have an analytic chop-off sentence. It makes me confused a lot. Because sometimes I

have very good grammar but poor organization. And it's hard to make decisions and we just can't give proportions of .3, .25 this is hard for the final decision. So, this is good when we do the calibration session so then we can find where we need to the language points anything that we need to focus on but... I don't know it's not that useful so that's why once I internalize the rubric and then I just follow sort of my professional instinct.

INTERVIEWER: So based on your experience as an instructor?

RATER 9: Right, I think I calibrate by myself with the current rubric and then I adjust my severity a little bit and then once I set it just follow that one, which is already in my mind.

INTERVIEWER: How do you think we should improve this rating scale?

RATER 9: I don't know if this is going to improve or not but I think we should minimize the words. I don't know we can do that but to many details for holistic score... the other raters really follow the details all of them....

INTERVIEWER: So, if we have a description for holistic scale then would you suggest that we only have one little paragraph saying that "B should be this"?

RATER 9: I don't know, I don't think it's about that way but I don't have any clear... I think it's up to the rater training part.

INTERVIEWER: Do you think that there are other categories that we should add in or do you agree with the percentage here?

RATER 9: Personally, I don't like the conventions here. It's not really writing, I'm not sure. Especially spelling, is that a writing skill? I don't really think... as long as you can tell which word it is, I don't think it's a writing skill and also if you cannot tell which word it is that makes the content part sort of ruined. We do have some other parts that can take care of these ones, so I don't personally like having spelling in the evaluation rubric.

And then source citation, I don't think this is good criteria for undergraduate students. And also, here in sense that the sources might not be cited not cited a properly and then I don't know where the raters are formatting the wording sources cited a properly. Does that mean that we have to put the year, the names or is it ok to use titles? It's not really clear this part so I don't whether the other raters really consider this one with the academic writing or not.

That's one thing and then grammar... I think grammar is ok. It's related with the comprehensibility. I like those kinds of things because writing is one of the communication vessels. So that's an important part. As long as it does not like any communication breakdown, I think it's ok.

And arguments and details... yeah I think that's academic writing style and organization... yes that is important. But sometimes I'm thinking "what is well organized"? In my case, context wise starting with a thesis statement. Not this statement itself but starting with the general idea and then comparisons or some list of examples. When we say organization probably some think that whether we have all those things. Probably some focus on that kind of things

whether they have topic sentences on each paragraph or the first part of the paragraph, probably someone thinks that way. But personally, I don't really give that much emphasis on putting topic sentence on the first part of the paragraph. As long as they have smooth transition I don't really mind. I know it is easier to have topic sentence but... generally it's ok.

INTERVIEWER: What do you think about the percentages? Like 30 and then 25, 30 and then 15 percent.

RATER 9: I think the flow should be smooth. I think that is the most important part of writing. I give about.... almost half of the points to Organization and another 30% for Argument and Details and then the other two. As long as they have acceptable grammar skills, I don't really care much for the grammar part, but if their grammar makes me confused, then there's a problem. If not, then it's acceptable. And then the last one, Conventions, I didn't really focus that much on that one. But sometimes the comma, the period makes the sentence different but if not, it was ok. I would put 10% for Conventions, and then, that 5% should go to the organization and arguments part. So, it'll add to 100%. 30, 30, 30, 10. We I think that's better.

APPENDIX E: TRANSCRIPTS OF RATERS' VERBAL REPORTS

RATER 1 (02/02/2018)

Rater 1 – Essay 1

Here I'm just refreshing my mind about, my memory about the rubric, so I was just looking at B and the pass.

I was just trying to read it very quickly

I always think okay, I'm gonna go back anyway.

And I read the introductory paragraph very carefully. I had to, in the first paragraph, I had to reread some things

because it feels like they're just paraphrasing a point, but they're not adding new information to it. It's like they're just repeating things. Right?

Then, I don't know, 'cause this is task two, I expect, in the introduction, for them to choose a position whether they think that people should use GMO or not. Right?

But I don't really see that argument, the examinee's argument.

So, I moved on to reading the second paragraph.

I noticed that there's an attempt to cite, but what I noticed is that what the person is writing based on the reading text, there's a misunderstanding of the content of the reading text. Right? They're citing, but their integration of information is not right.

I think I took points off for that.

They're adding new information, but it is not entirely clear whether this was based on something that they have read or their own experience.

I feel like you need to transition with something like, "Based on what I ..." to still acknowledge.

What I'm talking about is when they were mentioning bananas, where did that information come from?

Then, I moved on to the third paragraph. Yeah.

In the third paragraph it's clear that it's not done. Overall, this essay is very incomplete.

The ideas are very vague.

There's an attempt to cite sources, but the integration is not done very well.

And I don't really understand the point, what they're trying to say.

Here I'm going to the text to fact check.

I like doing fact checking just to make sure that they're not misunderstanding the text.

'Cause they didn't finish the text, they didn't really integrate this part.

I went to the second text.

They misunderstood. There's this environment activist, but just because she or he is an Indian environmental activist, it doesn't mean that golden rice is only used in India.

But, I don't know. There was a lot of assumptions there.

And the way that they're summarizing it ... They're trying to summarize the information, but they're just not including their own opinion.

So, it didn't feel like a task two.

In the third paragraph, they're putting the focus on India while GMO is not strictly used for India alone. It's just an example. Right? So they're making a generalization that is not in the text.

Yeah. The integration, the citation of sources just didn't really make sense. They're citing it, but I don't see the purpose of it.

Then I think I went back to reading it again, but I'm just skimming.

Then thinking, "Okay. I'm gonna give this person a B because it's just vague and it's not done."

I went to the rating scale, and I read the B scale.

I feel like if there were a lower grade, score band, I would give it a lower score band 'cause it's incomplete and it's very vague.

I feel like they show simple grammar structures and arguments are vague and the organization, there's a lack of organization between paragraphs and within the paragraphs.

So, I gave it a B.

Rater 1 – Essay 2

So once again, read the introduction carefully

and I was happy that I found where the position, the test taker's position is, so I have an idea of how they're going to organize their writing.

But I do recognize in their first paragraph, they use the same vocabulary over and over again like, "Should not be supported, should be supported, should be supported," Right?

And first impression is, the variety of vocabulary is a little bit lacking.

And then the words like "cost effective", well I think ... Yeah, I don't remember "cost effective", but I know that "controlled by biotech companies", there's exactly four words that appear in the first source text, so I know that. So, there's also lack of paraphrasing in there.

I moved onto the second paragraph.

So as I read this, I'm fact checking in my mind comparing the accuracy of the information you know.

So they went to cite source text two right away; but in this paragraph, they're actually just summarizing information from text two, and I'm not seeing any commenting on that information.

But then, when I was fact checking in my mind and I did think, "Okay, I have to check this in the text again."

There was some information that I thought was wrong; so, when I read it here I thought, "Well, that doesn't make sense for a Golden Rice only costs \$100."

And then they went on to say, "However...". That doesn't make sense because it's Golden Rice that is GMO, not supplement and fortification programs.

I was further confused because they said, "Which means all 500,000 children that goes blind each year could be safe." Like wait, shouldn't that be at the benefit of Golden Rice? It's not the benefit of a supplement in fortification programs.

I was like, "Okay, I'm gonna to physically to recheck that again."

But I know that this paragraph isn't really giving me other information beyond what's in the text.

And then in the third paragraph, I noticed that it's also not completed, when they say text two ... I think I didn't really go into this paragraph, because one, it's not done.

And then they cited the wrong text in paragraph three. They said text two, but this is actually talking about text one.

So I think after this, I went to the introduction I think. Yeah, I just reread it and I'm kind of thinking ... I think I'm gonna give this a B. Right?

It's just like the other essay.

It's not really giving me new information.

I was looking at keywords, that they're saying "it's cost effective, controlled by biotech companies".

I just wanted to make sure that they were hitting those points in the source texts, which they are but I wanted to see more information outside of the text. 'Cause this is task two.

I like rereading introductions, 'cause that's a very important part of text I think.

As you can see, I was focusing on that "controlled by biotech companies". At that point I was 100% sure that that was taken directly from the text.

I was rereading it again I guess.

But then I was like, "you know what, I'm just going to give this a B."

It's not done, there's nothing much beyond the text. This is supposed to be task two, so they're not completing the task.

I think I went to the text ... I found the "controlled by biotech companies", a phrase, and then I was like, 'Eh, lack of vocabulary

and just doesn't know how to paraphrase properly."

And then I checked ... I think I spent more time on this text

because I was a little confused of how they tried to paraphrase it. So there's a lot of ... They're taking words from the text, and using it like "deficiency, supplementation, fortification", and in the text, those terms were actually defined. So, I feel like you can actually paraphrase it pretty easily because even if you don't really know the term that's defined for you in the text.

And then my intuition was true, my memory was true, that it's the Golden Rice that... If they had implemented the Golden Rice, they would be able to save the 500,000 children that go blind each year. It's not because of the supplementation in the fortification program. So, they misunderstood the text. Or they understood the text, but they just didn't know how to write it up.

But at that point I was sure that I was going to give this person a B.

I was just skimming for the facts in the article, because it's everywhere.

Just very briefly looking at the essay. I wasn't really thinking about anything. I just like going back and making sure that I feel comfortable giving that grade. I read the second paragraph.

Okay, no new information.

I think I'm confident that this person should get a B.

I think I went back and forth a lot between the essay and source text 2 to check if this person has paraphrased enough.

'Cause I went to the text and like, okay, "vitamin A deficiency, supplement program, fortification program", those were all taken from the text. I think I was thinking like, "If I were the test taker, how would I paraphrase it?" So, I was thinking about that.

(reading on rubric)

It didn't take me a long to read the rating scale.

So, I was like, "Okay, I'm just going to give it a B."

It's not organized at all, and hard to follow.

So, I would also give it a lower scale if there was one.

Rater 1 – Essay 3

Okay, so I've read the first paragraph immediately,

and the first thought that came into my mind was that it was very ungrammatical.

But I feel like this person also has a little bit more of knowledge beyond the text about GM crops compared with other test takers.

The person did give their position, state their position, so I was happy with that.

It was just really hard to follow because of the ungrammatical structures.

There's a lot of spelling errors. I feel like this was not properly proofread before it was submitted.

They wrote a lot, so I think they just ran out of time.

For the second paragraph, I realized that there's no citation, there's no integration of the textual sources. They are citing from outside of texts when they say, "there's an experiment that is being conducted". I get what they're trying to say in the second paragraph. But actually, in the second paragraph's first sentence, that information also appeared in text one, I believe.

There is some information that are not from the text, in the second paragraph for example. I liked it. I mean, that's the point of task two. You don't want to just rely on the source text. You want to include your opinions. I remembered that was in the prompt, to include your own experiences, not just summarize, because summarizing is for task one. So, I was okay with new information. I actually wanted to see that.

Let's see if they're citing something from the text in paragraph three.

Here [in the third paragraph], I actually liked this point where they say "CM crops are highly tolerant to herbicides...".

There's a misspelling there. I think they meant "GM crops".

But still there's a lot of misspellings and I had to reread it several times to understand what they're trying to say.

So, it is the misspellings and ungrammaticality that was affecting the comprehensibility of the text.

Then I think I moved on to the fourth paragraph.

When they say "we're hurting those countries which depends on crops to increase their income", I feel like they could cite the text one, but they didn't. Because text one talked about that a little bit, but they're not citing it.

They were trying to describe it. It did bother me.

Task 2 wants you to integrate the sources appropriately, giving appropriate attribution, but they're not doing that.

At the same time, they were able to finish the task.

When I read the last paragraph.

I realized that they're using transitions, but they're not using it correctly. "First of all, secondly" is okay, but "further information" ... You don't really say "further information". "In addition" something like that. You don't just say "conclusion". So, they are using it rarely and inappropriately, like what the scale says, their transitions.

At first glance, it looks like it's organized, but actually ... It is a little bit organized, I feel like.

But it just takes time to read. Again, takes time to understand. That was my main problem with it.

I think I went to the task, and the last paragraph.

So, they're concluding that "they're not supposed to be supported because of the side effects to human beings in countries".

I was a little bit confused when they added "even they have benefits in other ways". What do you mean by that? Because you didn't really talk about some benefits of GM crops in your text. I didn't understand what they trying to do there.

But I liked "you have freedom to make the decision and your health is on your hand and for your future generations."

So, I think I'm going to give this person a B+.

I was just skimming the second paragraph again. I was just rereading things. I wasn't really looking for anything.

And then, here I was reading it, and kind of mentally doing a checklist of the rating scale, but I was focusing on B,

because I was 100% sure that it was never going to be a C/D.

It's somewhat organized, but very, not even maybe, but very hard to follow. Cohesive and transitional devices are used rarely and inappropriately. It didn't really include irrelevant details,

but it definitely lacked ... It was just really unclear. The ideas are mostly relevant but definitely needs more explanation.

I feel like they could have integrated ... The problem with this one is they didn't really integrate anything from the source text. Well, they actually mentioned some of the concepts that were introduced in the text. I feel like they could have ... Even if they were technically getting ideas from those texts, they could have strengthened their own argument using the source texts. And I feel like that would ... Even with the level of comprehensibility of the text, if they cited it right, I would give it a B+, I feel like.

The grammar structures were just really hard to understand. They're very, very ungrammatical.

And they tried to use this vocabulary but either they were used inappropriately, used it in the wrong way

or it was misspelled.

So it was really hard to understand.

And then when I read "sometimes interfere" on the rubric, I feel like, okay this is not getting a B+, because it interfered a lot with my comprehensibility. It was more a B than a B+.

I think I read the text again

and then just thought, okay, I'm not going to give this a B+. I'm going to give this a B.

I didn't really read the whole thing, I think. So, this is reading for confirmation of the grade that I'm going to give, right.

I like to do that. I don't like to just give it a B and then look at the scale, and then done. I sometimes like to go back to the text and make myself more confident that this person deserves the B grade.

Rater 1 – Essay 4

I just read it. I think I was just quickly looking at the overall.

... Because, you can clearly see that it's not the ... The task says that you need to have four paragraphs, and this is two paragraphs.

But I was like, "Okay. I'm going to read the content first before I try to assume things." Right?

I read it carefully.

First impression was that the vocabulary was better, the grammar was better, and then the position was clear that the test taker is for GM crops. They support it.

And then, they're citing the text. When I see citation of the source text I'm curious whether they're actually going to go beyond just summarizing what is in the text. Right? I do see that there's a summary, but the author is definitely trying to add more to that. Trying to provide their own comments. For example, there's this information from this text. The text that they mentioned was about the GM food... the companies hold a lot of power because of the GM technology. But, they're trying to give a solution.

"That GM technology should to be in the hands of the government".

And then, they follow it by several arguments for why that should be the case

and I really liked that argument. I think for the short amount of time that they were given, it was a pretty good argument. It was well thought out and I could understand it.

The main problem is for that whole big paragraph, here's several main ideas there that can be divided. Like, this can at least be divided into two paragraphs.

But I understood the flow of the arguments and I liked it.

I think I was just reading carefully here.

But, then the second paragraph they cite the second text, but what I noticed is that there's less commenting on this one. It's a little bit more summary than ... Not exactly summary, but taking information from the text. But, they're using the citation to efficiently argue for the support of GM.

There's a typo there, in "I can contain many necessary vitamins that are required for good health."

They transitioned pretty well too. The golden rice is a very good example. Right? I thought it was good.

So I was reading the grading scale again.

I was sure that it wasn't a B.

Because it was pretty well thought out. The content. The flow was understandable.

And it wasn't a B+, because B, you can't really can't comprehend it. B+ is a step beyond that.

But my interpretation on B+ is it's harder to understand, to comprehend with still quite a bit of grammatical inaccuracies. The range of vocabulary's still pretty small

but I felt like this essay wasn't that.

so I started reading the C/D. I read it

and I'm like, "Yeah. Well, this essay is adequately organized. It's not well organized." But, it doesn't require a lot of effort to follow.

It just wasn't divided into nice four paragraphs.

There wasn't really a redundancy in her woven information. I felt like all the information that was included in there in the essay helped me to understand. It was well described, so there wasn't any redundancy.

Then for the organization, it's more of a C/D+. That's what I thought. It wasn't C/D it was more of a C/D+.

And then, I went to the descriptors for arguments.

They were mostly developed and it's clear and relevant, and they did integrate sources.

Again, C/D+ not really a C/D.

For the second paragraph they could've added a little bit more detail, not just from the source text. Right?

But it wasn't a pass. It was a C/D+ for arguments and details.

And then, simple grammar structures and some complex ones. They definitely had the range of vocabulary. They didn't even have a lot repetition, so I felt that was good. There were some misspellings, but they didn't really interfere.

Mostly C/D+. I think at that point I wrote C/D+.

But I think I went to the text again and just reread. Because, I was like, "Is this a pass? No. This isn't a pass. I feel like I shouldn't give it a pass."

I was just skimming here

and thinking if it was better organized, like four paragraphs and a conclusion paragraph I would definitely give it a pass because, for the amount of time that they were given this was a pretty good text. Just because of the organization and the lack of conclusion paragraph I felt like ...
Yeah.

I was just thinking about that position that I'm going to give them a C/D+ and not a pass. But, I was happy. I was happy with that C/D+.

Yeah. I think I went back here to the rubric. I focused on the grammar and lexis there. Because I was reading from the grammatical structures

and I felt like it was pretty good. The complex ones.

I know they were comprehensible. None of the structures really interfered with comprehensibility.

So, I didn't think it was just a C/D. I thought it was more a C/D+.

I went to argument and detail too.

Well, second paragraph really needed ... I felt also if that second paragraph was more elaborated on it would be a pass.

I'm trying to find what would that person need to get a pass. "Why am I not giving this a pass?" I was trying to justify my own decision.

Okay. I was looking at the first paragraph again. I was just skimming, I wasn't reading carefully anymore.

I was thinking, "If I were this test taker, where would I cut this? I feel like this is too long. This could be two paragraphs."

I was rereading it again for, "am I confident that the grammar and lexis is enough for it to be a C/D+ and not just a C/D"?

I was just quickly looking at this paragraph

because I didn't have any feeling that they were not paraphrasing. They were definitely paraphrasing. Because, when students don't paraphrase it's very clear, because these source texts are quite technical texts. The source text is quite technical, so you can identify those words from the text very quickly. Like, "Oh. I read that."

But, the amount of description that was given to elaborate on the source text is pretty good in this one. But, not enough.

Okay. I was looking at the first source text-Skimming through it.

And then, I was confident that they paraphrased appropriately. They didn't use ... Actually, they didn't use a lot of the highly technical terms in the source texts

because they were trying to describe it a little bit more and give their own solution, I think.

And then, I just went quickly to this essay and looked at it very briefly.

But, there wasn't any copying of terms.

And then, I went to the second text, I think? I didn't check it

because there was very little copy of fact and they attempted to elaborate a little bit, but I think their attempt to paraphrase was good enough. I didn't really take points off for that.

And then, I was just skimming the essay again here but not really looking for anything in particular.

Yeah. I was looking at that "50 grams of the golden rice off of 60% the daily Vitamin A requirement."

That was the only thing that really came from the source text.

I liked this structure, "... and when compared to supplementation rate it is way more cost effective." Meaning, golden rice is way more cost effective.

That was very brief in the way that they paraphrased it

compared to the last person who completely misunderstood the point of that comparison between golden rice and supplementation fortification.

So, I liked this. Good understanding of the text.

I think I was done. It's a CD.

Rater 1 – Essay 5

First impression about this one is that there's a weird use of words "like no one would be strange about". That was the first impression I think.

"Someone think they might have some bad influence in the future and so as a result of"

so there's a run-on sentence

and I don't really know where this is going on or where this is going.

And "others hold the view that GM technology would do benefit to people for those who want cannot afford for the food."

I know what they're trying to say, but they're not saying it very efficiently and using ungrammatical structures.

It's very hard to understand

but it's good that I know the exact position of this person

because "As far as I'm concerned GM technology does do much more benefit right now".

Based on that, my assumption was "okay, I'm going to see discussion of how GM technology benefits people, right now."

But the first paragraph was just badly written. It was strangely worded.

So, I went to the second paragraph.

There was a mention of the second article, I would just consider that as an attempt to cite.

But the next statement, “many people who cannot afford food would survive in result of GM technology”. But the second article wasn't actually about that. It's not about the quantity but it's about the quality of the food. Yeah, I think this person is just not getting the point of the text.

And I really didn't understand where the person got the “10 percentage of global humans are suffered by hunger”. I get what they mean but I don't know where that came from. So, there's no citation there.

And then GM could provide requirement nutrition, so there's a lot of jumping between ideas so I don't really see it's not a transition between ideas, it's not very clear and there are, I don't know, per meters square it's that the M2 thing? So there's a lot of confusion here. I don't necessarily get what they're trying to say easily.

Yeah, so I was just reading carefully and then I went to the third paragraph.

Again, some of the expressions were weird like “the worry of the first article isn't fake”, “but if we take our eyes further”. I wouldn't use those. Yeah, they're very weird expressions.

And they are ungrammatical.

“The cheaper food no only provide for people”.

As I went through this text, I got even more confused as I just didn't see the function with each paragraph.

Again, the second and third paragraph, I didn't really see the function and the essay is missing a conclusion paragraph.

So this was at the end of the reading of the text, I easily picked a B.

Yeah, it was just really hard to understand. Efforts to go beyond the text doesn't make sense. It's clearly not done because finally most food would supply by just few companies and then nothing else without a full stop, so it's just not good.

Yeah, I was just skimming through paragraph 2 again

and like, “what's the function of this paragraph?”

I was trying but I didn't really understand.

I think I went to the first text. So, I was reading the text to just refresh my memory of what it's actually saying.

And it's not exactly saying that GM crops can produce more. It's not about that. I feel like maybe that was the function, but it's not the point of the first or second paragraph, but they're not talking about that, so the test takers not really talking about that. So I don't think they really understood the source text actually.

I think that's why it just didn't make sense to me.

They're trying to paraphrase

but it wasn't really clear why they wanted to include that information.

And I went to this text really quickly, the second source text. Yeah, just really quickly skimmed, but I was sure that this essay was a B.

It's kind of like the first and second text where it's not done, it's just very vague.

Yeah, skimming the essay again

and thinking, "okay this is definitely a B".

I reviewed the scale briefly. Yeah, the scale uses these like "somewhat".

I was like, "is this essay somewhat organized?". And then, I changed it with another adjective like, "this essay is very unorganized".

I was confident that this is a B. It can't be anything else than a B.

Rater 1 – Essay 6

I was looking at the introduction first

Yeah, clearly, it's very ungrammatical, and the flow of ideas ... it's not cohesive. That was my first impression, it's not cohesive.

This is very ungrammatical.

And then there are irrelevant information. Like, you don't even need to say that you're not a scientist.

"However, I'm not a scientist. I also have my own opinion for the genetically modified food. My perspective then genetically modified food should not be eaten."

So, yeah, at least I know where this person is going.

But, there's not enough information to contextualize the topic here in the introduction.

"Somewhere more scientists start to find if the genetically modified food is good for human bodies."

I don't understand why they have that assumption. I don't understand. How did they come to that assumption? Underdeveloped ideas, I feel like.

And I don't know how that connects to the source text.

Then, I went to the second paragraph.

And then I saw that they never cited the sources.

And they used phrases like, "As we all know", which is not the case. No, we don't all know.

"Genetically modified food is, give the food a new gene from other species although it do not change the food's look like and give them a new property."

What you should have done is to cite the source text because this is where this came from.

I think that was their main topic here. But, what I don't understand is ... So, they had that one idea, right. And then they continue commenting on its benefit.

"It is very easy to add a gene into another species to improve that species production."

But then, why is this person suddenly citing text one?

"In the text 1 the author mention that it makes cotton cheaper...."

So, there's two ideas here in that second paragraph, and I don't see the connection between those two.

The organization is really bad.

The idea development is really bad. I don't understand why.

And you can see here that they're not even bothering to paraphrase it. They're just doing the quotation thing. So, I made assumption that this person probably doesn't know how to paraphrase it.

And they don't really understand that quotation from the source text because it doesn't connect to what they previously said. I didn't like that.

Then I read the third paragraph.

Then they again talk about, "Influence the health of human."

The worst part of this third paragraph is the fact that they interpreted the source text wrong. Here they say, "So as the text two, the golden rice have more Vitamin A than the normal rice", which according to source text two, if you remember, is a good thing because there's a lot of people with Vitamin A deficiency. There's a lot of children with that deficiency. But then they commented, "So, if human get too much vitamin, it will cause the blind or dead." No, it's not the amount of vitamin that is causing people to be blind or dead. It's the lack of that vitamin. So this student misunderstood.

"Other species will have negative property."

I don't understand what negative property is.

"Some of them we cannot found."

I don't understand that.

Then the conclusion is just one sentence. I feel like they can do better than that. This definitely a very underdeveloped text.

It's the clear B for me.

I was just skimming looking at introduction again, making sure that I can justify me giving it a B.

I really didn't like this text. It was very hard to follow.

I looked at the rubric very quickly. But I looked at CD

and I was like, "No way that's a C D, this is a clear B. It's worse than a B."

Rater 1 – Essay 7

So, I read the first paragraph

and the first impression was that it's vague. Well not as vague.

"what should be widespread on earth".

I feel like it's kind of a weird expression to use. But yeah.

The second paragraph, "10% of people that have vitamin A deficiency"

Now this got me thinking, which news reporter? I know that it's talking about the second text. But they're not citing appropriately because both of them are news reporters, basically like journalists. But just saying news reporter is not enough to show like good citation skills.

And then "rice has become is a serious problem people were facing"

There's a lot of grammatical inaccuracy like. Why is it "were facing? People are still facing it right? And then "has been developed".

"It is a great movement for food."

I don't know why it is there.

So here they are quoting again directly. They are not paraphrasing.

And then there's just a lot of irrelevant information here and I just don't know where they are going.

"It means that the problem about vitamin A deficiency has been fixed."

I don't know where they got this conclusion. They're not phrasing the facts right and phrasing their arguments right. Yeah, I wouldn't say that it means that the problem about the vitamin A deficiency has been fixed. I don't think that was the meaning of the quotation here. So, they misunderstood it.

It bothered me cause they're not really paraphrasing there. They're just quoting American Journal of Clinical Nutrition.

"Everyday people only need to eat a little bit of vitamin A, then they will have the daily recommended intake of what they need."

They oversimplify things. They oversimplify the topic. They're misrepresenting the information in the source text because I think they don't understand the text.

And I'm reading the third paragraph again

and there's no transition; so, it's very jumpy, like why are you suddenly talking about biotech companies?

And then, they're not citing because this is from text 1.

"So if they don't have enough money to do then they can ask their people to collect money".

There are misspellings that make it very hard to understand these sentences.

I also felt it's hard to understand because of the grammatical inaccuracy.

And suddenly they conclude their paragraph with "it is helpful for poor people, also it's a better to stop the high rate of vitamin A deficiency death". So, they're going back to the idea of vitamin A deficiency. But that wasn't the point. That's the second text.

So I don't know. Very jumpy, not cohesive at all. Very bad organization.

I decided and I wrote down B.

And then the conclusion is not really a conclusion. Because when you conclude you want to summarize what you said and leave the reader something to think about.

And then it says, “at last I think people should care about their health”.

Why are you suddenly talking about health?

“Now the GM products have been developed, people should use this chance to rebuild their countries”.

It doesn’t make sense. I couldn’t really understand.

So clear B.

I went back to the introduction here.

And then I thought, well, they’re talking about GM food should be highly supported, but the supporting arguments are really bad.

I was just reading the second paragraph again, not really thinking about it.

I was like, “why they didn’t paraphrase it”. I didn’t like it. Direct quotations. It’s very clear they direct quoted it anyway. Yeah.

Then I went to look at the scale.

Actually, I was briefly considering giving it a B+. But for a B+ it should be more comprehensible.

But they didn’t paraphrase and for some of the information that they included, they didn’t cite. Organization, very bad. Arguments, very bad, not very clear. Some of them are irrelevant. The vocabulary itself, it wasn’t that it has many repetitions, it just didn’t make sense. They were just using just because. They didn’t really think about what they actually mean. So that’s why I couldn’t really understand the text. There’re misspellings, and it can make it hard to understand the text.

So, I like “no, this person doesn’t deserve a B+”. So, I give him a B.

I was looking at the essay, at the number like 10%, 50 grands.

I think I was looking at their use of the source texts to see if they copied anything. But I thought ok, beyond this one, that’s the only thing that they quoted.

For the other things, it’s just them trying to talk about the source texts but they didn’t understand the source text.

I was looking at convention at this point, and it says “text from the source is used without paraphrasing”.

So, there was direct quotation in the essay.

And then I’m checking my grade. I was thinking, yeah this is a B, not a B+.

I was just skimming the essay I think trying to find other things to criticize.

Ah, I think I was looking for transition words.

And then I was just looking at the first source text.

Yup. They're not really citing properly.

So, I'm done. So, it's a B.

Rater 1 – Essay 8

First glance, four paragraphs, finally! First glance, pretty good.

But then the first sentence of that introduction paragraph was clearly a misunderstanding of what genetically modified food is

and I thought this is a problem with highly specialized text is that, especially me, I feel like, as a reader, I grade down test takers who misunderstand the reading. So I feel like even if their grammar is pretty good, and the vocabulary is okay, I would grade their arguments, 'cause you can't really make an argument for the prompt, if you misunderstood the text. I feel like I grade really harshly if they misunderstood the text.

"It should not be supported because it takes the natural components like vitamins, fibers, blah, blah, blah, that provide good food composition for the body out of the food."

No. That's not what GM food does. GM foods actually enhances certain characteristics of a crop so that it can be more ... It can increase the quality of the crop and the quantity, sometimes.

"It compromises people's health and minimizes the longevity of people's life."

I don't know how they got that.

"GM food can be controlled by the government." "The control is not on the government side, but the companies."

And then they cite text one here. This suggestion should be that the government is the body that needs to have more control over GM crops, not the companies. So, misunderstanding.

And then the test taker talked about independent citizens

and I don't know why there's a connection with independent citizens.

And then the test taker is defining what "independent citizen" is

which made me more confused than actually clarifying things, because I don't understand why. I don't know. It needs more description and elaboration for me to understand the connection between this and what they said previously.

And then after that introductory paragraph, they introduce themselves as from the Dominican Republic. '

And then, "I have known about companies that utilize poor quality ..."

So, I was happy that they're including their own experience here, but it's not clear the purpose of sharing that experience.

"Because they pay enough money to the government, they are not punished as they should by the country justice system." And I'm like, "Money, power, certainly people act if they do not have high ethical standards. "That being said, cheap does not mean quality and GM food does not mean healthy food."

But I feel like you need to be more specific than just saying ...

Then if you're using that experience, I think, and not incorporating the text there, you're just being biased. I don't know. That's how it felt like.

And here, why should they be punished by the justice system? And what do you mean by utilizing poor quality GM food? Are the companies producing the poor quality GM food? How is it hurting the company? The community? So, I feel like they can go on explaining more about that.

And then, in the third paragraph- Suddenly in the third paragraph, they talk about that it is important to mention that, "I am a member of a church that encourages to have food stored."

I'm like, "What?" How's this have anything to do with the first two paragraphs? So the church encourages them to have food stored for emergencies, and to grow their own food if possible?

"Having their own garden increases the probabilities to improve..."

But that's not the point of the genetically modified food. 'Cause it's not feasible. Not everyone can have their own garden. So, their arguments are not good enough.

I feel like, and there's a missing condition, because I don't see the connection between this argument and the previous arguments. But I do see the connection when they're trying to push this idea of being independent citizens in the first paragraph.

And then here they're talking about growing your own food, but the talk should be focused on the GM.

And then they end, which actually this fourth paragraph doesn't really feel like a conclusion paragraph because they're still introducing some ideas where here.

"Based on the text, it is clearly observed that GM food is just a way to control people."

This I think is a very strong argument to make, because I can't see anything in those two texts saying that it's just the way to control people. Sure, companies have a lot of control over the technology, but a way to control people? I don't know. Yeah, the argument doesn't make sense.

"In text two it's observed that investigation and supplementation and fortification programs will cost more than just consuming a natural, well-grown, and healthy, golden rice."

If they say natural, then that's not true, because golden rice is GM food. So they're misunderstanding. They don't really understand text two.

"It is completely disappointing finding people that support such type of food that will end up hurting not only them, but their prosperity."

They completely missed the point that golden rice is GM food.

So, connection between each paragraph is not clear and the arguments that are presented are not really make sense to me.

No, I think I was just reading it [the second paragraph] again.

I was just trying to find the connection between each, and why is this person talking about controlling people's acts and having high ethical standards? Like mentioning their membership in church, like, eh. Very irrelevant.

I felt like they could talk more about the technology itself.

They tried to discuss the text, but they completely misunderstood the point.

Probably because they're pretty biased about it. Or I even think that this person is misunderstanding GM food, genetically modified food. They might have something else in mind, but I don't think it's genetically modified food. //

I was looking at that again. "Natural components like vitamins."

Wrong. I think I was thinking, "Nope, that's wrong". That's the wrong understanding of GM. "Nope, that's a wrong understanding of the text."

And then in between there's just a lot of irrelevant information.

I was trying to comprehend the text, but it was really hard for me.

I was trying to find the conclusion in the last paragraph. Where's the conclusion? I was like, "no, this is not a conclusion".

So they're going on, and on, and on, and they didn't really actually finish the task.

I couldn't ... Talking about independent citizens. Like, what? What's this got to do with genetically modified foods?

I went to review conventions right away.

There wasn't really a lot of misspellings. Misspellings wasn't the source of my inability to comprehend the text. And they did paraphrase, and they did cite.

So conventions is actually 15%, if I can say.

And then grammar and lexis is okay, too. But my problem is with the arguments and details and the organization.

So I feel like, no. There's not enough points to give it a B+. Just like briefly considering it for a B+, but then it's a B.

And I remember thinking that this person will definitely benefit from more instruction, B level instruction.

I was looking at the essay again, just very briefly, just skimming, again doing that process of making myself confident enough to give that person a B, or not a higher grade.

Rater 1 – Essay 9

I noticed in this first paragraph that there is a specific stance.

"I believe the negativity overweighs."

The wording is really weird. "the negativity overweighs"?

"Benefit from the state-of-the-art ways of improving life."

It doesn't feel natural.

"Agricultural industry is not then an exception and has been widely improved, especially through manipulating the genes."

The genes of what? This should be specific.

“There are both cons and pros of taking this ability of modifying genes of crops.”

I thought that was okay. I had an idea of where this test taker is trying to go with this statement here.

However, not a very good choice of words and expressions.

And then when I read the second paragraph. “This helps this industry... which can help reach more crops yields at the end.”

At first, I thought it was going to promote for genetically modified crops because they were talking about. But then the second paragraph does not make sense because they ended with something opposite.

"GM crops ends up in making poor nations poorer as it is less expensive."

But then where's the citation? Because that comes from the first text. Clearly, there's no citation there. I really did not like that, as usual.

This is just a summary. No specific stance from the writer.

And then the worst thing is in the third paragraph. That first sentence was like, 95% was taken from the second text. 95% of it.

And then they tried to paraphrase. There was a very small attempt of paraphrasing. In the source text, it says here, "as their main staple food"; so, they changed “staple” to “main course”. But that was there only attempt to paraphrase. Everything else is word for word comes from the text, so plagiarism.

And then in general, like the second paragraph, it's just all summary. What about your opinion? There's a little bit at the end of this ... There's again an attempt, but it wasn't enough.

This is Task 2. They're supposed to state their opinion.

"Instead of spending money on modifying crops' genes, more budget should be allocated to improve nutrition facts of foods."

The argument doesn't make sense. The whole concept of modifying the genes of the crops is to improve the nutrition of the food. So, did you actually read the text? So, again, they're not understanding the text.

They have this bias about genetically modified food that is making them misunderstand the points of the source text two. Yeah, they must have really misunderstood the text because that was in the text. It was clearly in the text.

“Half die within a year... A British medical study estimates that...”

I was looking, I remembered that, okay, this is actually in the second text, that copy and paste thing, yeah, here. They just copied from the second text. So, there's really no effort there to paraphrase.

I checked this last paragraph looking at the first several sentences in the last paragraph.

Because this one, two, three sentences, they're all copied. And then they just added a little bit of their own comments. That it was indeed copy and paste.

It bothered me very much so. Because I remember thinking, "Okay, if this is a practice that they're doing, then they will have problems in academic work."

Also, the prompt clearly states that you need to paraphrase.

I was just looking at it again and thinking, yeah, this is definitely a B because if you take that paragraph out, these three sentences out, there's not actually that much text.

And then I read this, the rating scale, and just went through the B level

I was thinking, "Is this organized? No, it's not organized." The very vague, underdeveloped ideas. Grammar is okay, but vocabulary ... for them to actually paraphrase, to take without citing, they had a hard time to come up with arguments. So for me that's an indicator for lack of vocabulary. Convention, I was like, "Without paraphrasing."

It was enough for me to give it a B.

Rater 1 – Essay 10

The first sentence, I thought this was okay.

But the second sentence was just a really, really long run-on sentence.

"Some people are favoring GM Foods as they are getting benefits out of it by having control over the entire food chain while others are finding deep trouble..."

So, there's no comma there, and I feel like this can be split into two sentences, right?

"This essay will describe both of them."

I didn't really know what "them" is referring to.

They didn't really state what their position was at the end. But I know that I wouldn't say their position until the end of the conclusion; so, I guess that I was okay with that.

But it already had problems in the way that they organized this introductory paragraph.

And then, "The GM technology are heavily controlled by the biotech companies."

So, they misunderstood the text, but I will check the source text and see.

So, okay, this is an okay second paragraph.

But it's missing the citation. They're not citing.

It's missing their own evaluations, comments, right? Again, this is task two. Where are your comments?

Then, when I read the third paragraph, right.

They suddenly jump into talking about golden rice without any transition to it, introducing why they are suddenly talking about golden rice.

In this third paragraph, they're just not exactly copy and pasting.

But they are just putting a lot of facts in, but you don't really know the purpose of those facts, right? What are you trying to say using these facts, right?

Interestingly, after the second and third paragraphs, I still don't see what kind of conclusion they're going to come to.

And then, suddenly, in the fourth paragraph, which I assume they think is the conclusion paragraph, but it isn't really a conclusion paragraph.

"However, it is found that golden rice is the most cost-effective source of vitamin A..."

So, I feel like this should still be in the third paragraph here, where they're talking about golden rice. But they're starting their conclusion paragraph with it. So not good in terms of organization.

And then they say, "Therefore, in my view". So, they are for GM products.

I'm like, "Okay, just because of that one argument, you're already for GM products?" So I'm like, "That's not a very good argument then." Sure, you want to show two sides of the coin on the issue, but when you have a stance, you want to provide enough arguments for you to justify your position.

This is not a very well-organized text. Yeah, it's just a very poorly written text.

If only prior to this conclusion, they introduced their stance, right, and like, say, I don't know, elaborating, giving their ... using their background experience or whatever, right, just to provide more comments. This is just summarizing, and then suddenly picking out one thing that they liked about GM food and then using that as an argument. I don't think that's a very good- Yeah. So again, this is just a summary.

I think I grade also really harshly when it's just a summary. What's the difference between task one and task two then? Right.

I reread the introduction. I reread the whole text actually, and trying to understand.

And then I was thinking, "Am I grading too harshly?" But you know, the text doesn't only need to have good grammar, good vocabulary, but it also needs to make sense because that's the point of academic writing.

And this didn't make sense to me. They're just, you know, putting all those facts and not paying attention to the cohesiveness, the organization, the structure, and what those facts are doing to their arguments.

I looked at the source text.

As I read the source text, they did actually understand it, that American farmers are being given a highly subsidized rate, and that hurts the other nations who produce it the traditional and costly way. So that's good.

I wanted to make sure that they're not copying anything. They didn't because I skimmed through it, and yeah, they didn't exactly copy anything.

But it was just like a summary. Their writing was.

And then, I went back and looked at the essay again, like, "Did I miss anything?"

I was checking the information in the second paragraph because this, the second paragraph is about text one. So, I'm like, "Okay, no, they didn't copy anything."

But they didn't cite anything... They didn't cite it.

And then, I read the third paragraph.

Well, they cited something here, which is good.

Let's see if they copied anything.

But actually they didn't. So, it was okay.

I'm like, "It's okay." It's still a B.

RATER 2 (02/03/2018)**Rater 2 – Essay 1**

And see first I was simply reading the student's submission. I just went right to reading the essay.

And I remember I stopped at "But some of the food that a lot of people eat is genetically modified". For some reason when I read it the first time, it didn't connect, and so I went back and read it again and like OK. I understand it now.

And then I read through the entire essay. I wasn't like looking for anything. I was just reading it for flow and do I understand everything that I'm reading.

I noticed the most were a few article errors.

And then of course, it stopped at the end right on a letter, so they didn't complete the essay.

So, I don't know if they had a timer and it was stopped so "hands in the air, don't type any further" or this student just... "I don't know what to say" and ended there.

I was reading the essay again because I wasn't sure that I understood the very first time.

The second time I read that point when I decided "okay, how much is this student taking from those articles and applying to the writing that here she is doing?"

And then I decided to look at the next two source texts.

Basically, I skimmed and went through text 1 very briefly.

I was checking, "is there anything that I think it is the main point that either the student incorporated or omitted?"

So I looked at "the price of cotton forcing West African producers", "biotech companies", "biotech companies uncontrolled."

Yeah. And then at this point I'm like OK I remember the rest of text 1. I'm like "oh yeah, I've got that."

So here, similar to what I was doing before where I was looking for main points. And so again "it was ten percent at risk".

I wanted to see if this author was quoted in the essay, and she wasn't.

And I focused on golden rice hunger. That sort of thing. Vitamin A deficiency in India. I looked in New York Times Magazine and yeah.

So, this was the point where I read the essay again

and I was focusing more on "What aspects of the articles again did the writer incorporate?" And so, I saw that they had Bjorn from text 2 but they didn't have Tom Chivers from text 2. And then I didn't have a lot of information that was included in the articles...

They brought up "bananas" which wasn't in either article. Not that it's bad but it was just extra. So yeah.

So, that's part of what they were supposed to do with summarize, to include important details from both source texts. I thought that there was more they could've summarized here. Yeah, I thought there was more they could have summarized from two articles.

Then, I spent a little bit of time reading it again.

Again, my mindset was I didn't want to miss anything or put something in the text that wasn't there. So I want to be thorough.

Because, my thought was, "okay, there are people who are actually going to be affected, so I want to do an appropriate job and not either A, put this person in the class when really their skill level a stronger or B, put them in a class and they're overwhelmed because their skill level doesn't meet that threshold.

"Bananas are hardly acceptable among people and for lack of taste".

But there was nothing in the articles about bananas.

I thought the writer might be going down the path of something like, "I know the bananas have gone through blight a few times; and so, because bananas typically are clones of each other". I thought they were going to explain more there.

I looked at that banana sentence again just to make sure I knew what the person was referring to.

So, okay, I know that this person was talking about the taste, not disease resistance or something like that, because that's what originally I thought this writer might be going. But that's not where they would go. (Laugh). But this is what I mean.

This is why I read it twice because I didn't want to impose my expectations on them.

And then well, I focused on "oh they stopped on the 't' and didn't finish their thought".

So, I was wondering about that [time was up. They had to stop. The system kicked them out].

Then, I decided this is a B.

Rater 2 – Essay 2

I first read through it, and mainly again for flow, and if there was anything that jumped out to me.

The first thing that I noticed is that this one had some more statistics and numbers than the previous one did "as far as the only hundred dollars", "4,300", "2,700", "factors thousand children", that sort things.

Even though we're not supposed to compare essays, I did.

So, for this one I was hoping to see a thesis statement, and it kind of got one here. It could've been clearer in my opinion, but I comforted it. I thought it was there.

"I think GMO should be supported for many reasons".

So, I jumped back up here to the first sentence again.

So, I'm like "Okay, so are we going for thesis again?"

And then here again I was taking note of there are some more specific. We've got “golden rice”, the amount of money that could be spent to save lives, how many children some could be prevented from going blind. So, I was paying attention to the details from the source texts.

Then I moved to the last paragraph.

I was reading this to see how it was for a conclusion. And my thought was, “basically what do we have for a summary or a closing statement?”, you know, some sort of a true conclusion.

then I moved on to the next thing, the first source text. So once again I was reading for points.

And in this case, I was actually specifically looking for the numbers the students cited because I wanted to see “okay, did he or she really pull them out of the article or was this made up numbers”? With this, I didn't find any of those numbers and I was little worried at first.

And I'm like “OK I'm gonna do one last scan” and then I quit.

So I spent a little bit of time on the second article.

Here is where I'm like “OK, I'm seeing the numbers here in the second text.” So this is where I was like “Okay. So, this is where the student got those numbers from”. So good they're using that source. It wasn't just, “Oh, these are good numbers to throw out there”.

And then I backed up to the essay again. I looked at the numbers again.

And then this is where I was debating for a little bit. Because for me, this is kind of between B and C/D. So, I was debating. I was really debating between B and C. So, ultimately I was like “OK, I was looking for basically looking for what in the paper that can tip me over the edge of one way or another, either okay I'm going to go with C or D, or I'm gonna go with B.

I came back and looked at the introduction. I was looking for anything to help me make my decision.

And so, I was looking at, “Okay. Do we have... are we struggling with command of English? Are we struggling with language? The language is not bad.

Again, I thought the thesis could be a little bit better, but I thought it was present and I didn't think it was horrible. I thought the thesis could have been more clearly stated.

There are a few places where the writer says “the author” but doesn't give a specific name. Yeah that was probably a big portion.

But they included the detail and well.

So, this is where I was reviewing the numbers and I'm like... I liked that the writer included specific details from the articles specifically mentioning “golden rice”, specifically mentioning vitamin A deficiency and then the numbers again with “a hundred dollars to save a person”, “supplements and other programs cost 4,300 and 2,700 dollars.

And in my mind, I was thinking, “I like this. I think I'm going to go with C/D”.

So, yeah. So, I again was really debating. It was the second paragraph that ultimately made my decision that “okay, I'm going to put this person in C.”

Rater 2 – Essay 3

Yep a little bit okay. Not a whole lot. But again, I just a quick scan or review. I looked at organization mostly and arguments a little bit.

So again, I'm reading the essay for flow just to understand.

And one of the first things I noticed was that this paper struggles more with spelling, word use, grammar to a certain extent as well

compared with the other two essays.

So, almost right away, I had “changing” instead of “changing”, “manning” - I wasn't exactly sure.

I think they meant “mapping”. But I wasn't sure. That's what my thought was.

“like prevent damage by insect”

So, I noticed that there are just grammar mistakes there.

And “I don't think genetically modified crop should be supported because they are lots of side effects on it”.

So, “they are” instead of “there are”.

So, I read here the second paragraph.

And this is when I was noticing like “I don't remember this being in the two articles that they were given”. So, I took a note of that and would check later.

I jumped back up here to the introduction again because I'm like “OK, I didn't read for the thesis statement the first time and so, I glanced back up this paragraph again. And then I came here.

“There's an experiment using conducted by GM that had suspected pregnancy a mother on in the side effect is affected the infant”.

So again, there's some language issues there.

But the biggest thing I noticed was okay, this is a very specific study. Was this in the article? And so, that was a thing that was going through my head.

Again, in the third paragraph, I was noticing some issues in language and spelling.

To certain degree, I was also wondering if these are just typos. So, for example “this will lea”. I don't know, just missed the “d” when typing. I don't know. The same thing with “herbicide to kill”. I'm assuming what that is supposed to be.

The misspelling bothered me a little bit, but only for a sense that I had to read it a little more closely.

So once again, I was reading for content and flow.

To a certain degree, I was also looking for... because at this point I really haven't found any information that I thought came from the articles that the writer was supposed to be summarizing; and so, I was looking for that specifically. And I ultimately couldn't find it.

And then I looked back here, so the student talks about poor countries struggling with companies, you know, that have the patents and that sort of things.

I'm like "Okay, so this was from the article." It didn't go into that much detail, but it brought up poor countries. And so, I remember I'm like "Okay so there's a little something from the source text".

By this point, I think what was going to my head was the English usage for this paper didn't seem as strong

as the previous one

even though it was longer.

And then again, I remember thinking "I need to re-read the articles because I don't feel like there's much in this from them". So, I was thinking that.

And then I did a quick glance here again. Oh, OK, I think I was looking for the thesis statement again.

I went to the articles and I checked them as well, just to make sure that I was interpreting it correctly or we're not missing something. So, this is where I was reading source text 1.

I was like, "Okay, was there anything in the first article was pulled from this article and put into the summary that maybe I had missed when I read this article the first couple of times?" That was going through my head as I was reading. Again, I found, "the poor nations are trying to develop the companies control the patents". Other than that, I didn't really find much.

So, then similar to the previous article, I then read this one again.

"Was there anything in here that I've missed the first time that was in the student's article or the or the writer's article?" So, I was looking for the specific situation with the pregnant woman.

Then, I was like, "Okay, these were numbers that were not present in this person's writing (the student's essay). So, no, I didn't see any of this in the student's writing, which was fine.

I also noticed that there was a lot of information here that seem to be outside of the articles.

So, at this point, I was thinking serious grading. And at this point again, I was leaning towards between B and C. But it was closer to B.

So, with the second paragraph, again, there was the specific part about the side effects with mother and children. I was thinking about that "I didn't find that information from the source text" and then I don't know where the students got that information from; so, did they make it up?"

And then, for this one they were mentioning about being highly tolerant to herbicides and then that causing environmental problems, which again were things that came from the source material.

So again, fourth paragraph, my mind says, "OK I'm reading through this to make sure I didn't miss anything".

And this is where I felt like "okay here's a little bit of information from the source text." It comes from Chivers' article about how companies own the patents and poor countries aren't able to develop this technology because of patent control.

But he writer didn't refer back to or cite the source.

They didn't delve into it (the source) is thoroughly.

But I'm like "Okay, at least I can say it came from that source. At least they use the source". And then I jumped back up to the first paragraph. Again, I'm looking for the thesis statement. It wasn't it clear to me.

I was looking to see if there was some sort of connection between those paragraphs. And then I looked at the conclusions.

So, then I was again reading for English usage.

"Just because you have freedom to make the decision and your health is on your hand and for your future generations".

I don't want to say it wasn't clear. I think I knew what the student meant. But I had to read it twice because it's not familiar expression.

I was leaning toward B now.

No. So, this is where I went back to the rubric and started checking what constitutes a B. The main thing I was looking at was arguments and grammar.

Because in this case, this student wasn't incorporating things from the article and was inserting his or her own information. I was looking to see okay how much should this affect my grading. And so, this is where I looked at – "sources are not integrated. Yeah, they weren't integrated.

So, in the end, I decided since the student didn't incorporate information from the source material, yes that would affect the score and basically that puts it in the B category for me.

And "arguments are vague and under-developed", and I agree with that too.

And of course, grammar. I noted several issues with that. It was already teetering between B and C for me already, and so then that just pushed it over to the B category.

Yeah, I've made my decisions and so, it's B.

Rater 2 – Essay 4

So yeah, so I just did a quick glance again of the rubric for a review.

So here, again, I read through the essay for flow, comprehension.

However, for whatever reason it didn't flow as well for me.

And so I just reread it.

And something that stood out right away was that the writer used one of the source author's names. I stopped a little bit at Chivers. I took note of that and thought, "Oh, they used that. Good job."

And then, I was noting that then, you know, not only did they use Tom Chivers' name, but then what were they talking about that he was arguing; so, it was regarding political power for the corporations.

Again, I was reading for "what am I reading as far as content?"

And it was flowing fairly well. So, I noted that.

And so again, as I was reading this, I noticed more information from the source material. And so, I was taking note of that. So here, again, I took note of the quoted, the other author, Lomborg, from the source material. And, yeah, so I noted that.

And I was making notes in my head to, "I'm going to check the articles again." Just because that's what I always do.

Okay, they quoted golden rice again, and that was from the source material. And so yeah, then I was noting the Vitamin A and nutrition, you know, information from the source texts.

So in the end, "I think development of GM crops should be supported."

I'm like, "wait a minute", because it felt like most of the essay, the writer was writing that no, this is a bad thing. And suddenly this conclusion. I was confused initially, because it felt like you spent, you know, this much regarding problems with GM technology, but then you said you think it should be supported. And so there seemed to be a disconnect for me.

And so, then I was looking up here to the first paragraph again and then I went back down here.

And I'm like, "okay, actually what this person said was, I think we should support research, but it shouldn't be in the hand of companies, it should be in the hands of the government." So, that's what I was thinking through as I was jumping all over the place there was- Trying to figure out that. I realized, no, there wasn't as much of a disconnect as I originally thought. So, the ideas are not organized well which is why it's hard to follow.

So yeah, then I went to the articles. I remember I was reading ... Okay, we have "licenses", and "companies", and "corporations", and "unprecedented power".

And so, I was ready to find those points that connected with the writer's essay. I did not see many of these words in student's essay, not specifically, but the concepts. So, I mean, "companies" and "corporations" were used in the essay, but "unprecedented power" I don't believe was. Although, the student did write that companies could increase their power and could possibly overthrow a government.

I guess another aspect to my thought process was, "so this is what the student is writing about, where did they come up with this? Or where did this come from?" And so in this case, these ideas or points were coming from these two sections of this article.

And so again, at this point I was looking for whether the information in the essay actually comes from this source.

The student wrote about vitamin deficiency and here in the source, they got Vitamin A here, and they got 60% here. So okay, this section is where the student got that information from in his or her essay, and so that's what I was looking at, or focusing on, or thinking about.

I also noted that there were pieces that the student didn't include in the essay.

So, again I was reviewing this (the essay) one more time. And again now I'm thinking "okay, now we have to get down to how am I going to grade this, or what am I going to set for a grade."

Okay, so again we've got concerns about corporations misusing it for political power. Again, that came from one the previous ... you know, one of the source articles. And then they also included the other author's name, Vitamin A deficiency. And I noted Tom Chivers' name, and again

companies have full control over the technology. So again, all of these were things that the student would have read from the source material.

So, I decided that this is a CD essay for me.

Rater 2 – Essay 5

Okay, here I was reading the essay.

And again, my initial thought was, "Okay, I'm going to read this for flow, just like I've done with the other papers so far."

I remember what I was thinking when I was reading the entire article is that the use of English was poor.

"Someone might think they might have some bad influence in the future in result of ..."

I think I ended up reading this a few times because of lack of clarity. I went back and forth a few times because I found this to be unclear.

"As far as I'm concerned, GM technology does do much more benefit right now."

So, I had to go through that a couple of times as I don't understand it because of the errors.

Because we don't usually say "does do". We just say "does"; so I was expecting a "not".

There were some other missing words there as well.

And then I moved on to this paragraph (the second paragraph). I went back up here because I think I read it yet again.

I was noticing the same things that I found in the first paragraph, where I'm like, "Okay, this isn't flowing as well."

There are some expressions that I'm not familiar with. There's some either words missing.

I wasn't exactly sure what per M2 meant. So I remember thinking, "Okay, I don't know what that meant," and then the writer says, "That is to say," and then explains here, although I wasn't exactly convinced that it really explained to me what M2 meant.

So then at this point, I'm like, "Okay, so this person is including material from the source text," so I was pleased to see that.

With the last paragraph, I was reading it.

Again, there were expressions that are not typically used. There were some clarity issues because of that. So, yeah, again I was noting lack of clarity, because of language, because of language.

"It also benefit to hunger people."

So, I had issues with this sentence. I could not understand it.

So, I went back to the previous line because it said hunger people. Then I'm like, "I think I saw ... oh, yep, yep. Hunger people."

I was noting that there's some consistent language issues that I'm seeing.

I'm like, "I know what's there. I know that this student included a few things but not much. And because the flow and the sentence structures and those sorts of things."

I'm like, "This is a B." For me it was a clear decision.

Rater 2 – Essay 6

"Today, the genetically modified food is more popular, such as the rice, plant, vegetable, even meats."

There was a flow where this doesn't sound how native English speakers would write.

So I re-read this first sentence.

"Some more and more scientists start to find the if the genetically modified food is good for human bodies."

So, this one does not sound right. It didn't flow.

And, "however, I am not a scientist, I also have my own opinion for genetically modified food."

I think the last part was fine.

And then I remember reading the second paragraph.

I was taking notes of language use.

I also noted, however, that content-wise the writer did use, mention the part about how genetically modified crops were undercutting West African cotton producers, which is from the source text. So, it's good that them make use of the source texts.

And then I moved on to the third paragraph, and I read this sentence about golden rice. "Golden rice have more vitamin A than the normal rice, so if human get too much vitamin, it will cause the blind or dead."

And I was, at first, I'm like, "okay good, this writer's including something from the text", but then it became clear that he or she did not understand what the author was talking about. I'm like, "no, actually, it was that people weren't getting enough vitamin A that was causing the blindness." And so, I took note of that.

I was pleased that it was there.

But I'm like, "oh, but this isn't a proper citation for quotations, but you need to know how to do it correctly."

There is also issues with language use.

And then, "The genetically modified food will make a bad influence for human body."

So again, that doesn't sound how a native English speaker would speak.

And then they had a very brief sentence as their conclusion.

It was a clear B for me.

I didn't look at the articles. So, my thought is I might've, because with the thought process being reading comprehension's important and if you read something and don't understand it ... Because I had already decided that this is a B, students' misunderstanding the source texts didn't really play a role.

Rater 2 – Essay 7

Again, my first, as I was reading I was reading for flow.

And this flowed fairly well.

There were a few points I noted where it caught me a little bit. But for the most part, I didn't have any trouble reading.

So, then I moved on to the next paragraph. Again, I was reading for content.

So, I was noting, "okay, yeah, we're getting information from the article". And I also noted, well, in the future, the quote from the American Journal of Clinical Nutrition and the 60 percent of recommended daily intake of vitamin A. And so yeah, so those were things I was noting was content. I liked that they were including this information from the article regarding the golden rice.

And again, the flow was going well. So, it didn't interrupt my reading.

And then I moved on to the third paragraph. "I think the government should give the companies more freedom to expand on their technology."

I noted that "okay, we're including information from the source material again with biotech companies."

However, one thing I noted was that the writer of this essay had a different take from the writer of the article as the writer from the source materials was very critical of biotech companies. So, the writer of this essay has a different impression of what should happen than the writer of the source article. So, it's a misunderstanding of the source text.

And then I thought, I also thought that even though it had a good use of information from the source material, it could've had more. So, it focused a lot on the ... I don't remember the author, but the author who wrote about vitamin A, it didn't provide a lot of information from the other source material. So it was a little heavy on one source than the other.

And so, then I moved on to the last paragraph.

So it's kind of a conclusion, but it incorporates some information from the previous paragraphs. I thought they could've done it a little bit better but I noted that it was present.

And so, after I read this paragraph, I'm like, "this feels like a pretty solid C/D."

Rater 2 – Essay 8

So again, I read for flow and for content.

And as I read this one I'm like, "this is reading really well." And so yeah, I didn't have any areas where I'm like, "I don't understand this."

There were very few mistakes as far as any missing words or grammar structures or anything. Spelling, even.

And then I noted the, "As read in text one," so the kind of a citation. And then talking about that content.

So, then I moved onto the next paragraph (2).

"Because they pass enough money to the government, they are not punished as they should be by the country justice system."

And so, this is where I'm like, this feels a little personal. And so I noted that. And I will say, when someone writes a summary it should be no emotion and no opinion. It should be summarizing the articles. And I think this task was a little different where they could give their opinions. So this is good.

Well, and then I continued with paragraph 3.

Again, the language was good.

I also noticed with paragraph two and three is again, we don't have much information from the source material again.

And then in the last paragraph, we have information from the second source text. So, I made note of that.

I also noted though that the writer, again I thought didn't quite understand the text or at least didn't make the connection that the golden rice is genetically modified. So, he or she has spent most of the essay arguing against GM food but then speaks very complimentary of golden rice. And so, I'm like, "I don't think you realize that golden rice is genetically modified."

And then again, the flow is very good. The writing is very good.

So I did a quick glance of some word choice, which I liked, "companies using poor quality GM foods." So, yeah.

And then I moved on to the source text at that point too.

And so, this was checking for information from the source material in the writer's essay. Again, because the essay didn't have a lot from this article but I was still checking, "okay, what they did right about, you know, is it here?"

And then I moved on to the second one. I wanted to make sure okay, I wanted to make sure that I didn't misunderstand the article. Again, I was making note of things like, okay, "company control" and "power" and that sort of thing. In this case I was actually looking for information here that I could then bring back when I looked at the essay a second time.

There were things that I pulled from the essay, the golden rice and because I had noted I don't think you quite understood, I did check that out.

And then I moved back to the essay. So again, basically I think I re-read the entire thing.

And again, I was noting okay, this is really good language and really good use of English.

And then again, I was noting content. And so piecing together what was from the source material and what was in the article. And then also noting extra information that was not from the source material but was from the writer.

And so, this part in the middle about GM power, it's written from the writer's perspective but one of the authors talked about it as well.

Now I'm on the second paragraph, so I'm looking at connections between companies that this person has experienced in his or her own country. And then the remarks that were made from the first article. So now I'm focusing on the, I don't think he or she understands, again it was the golden rice. Saying this is well grown and healthy.

At this point, I was debating between C/D and pass.

So now I'm at the rubric, and I am looking at every piece. And so at this point I was going point by point for each section. And in my mind I'm thinking, "okay, does this fall under the pass category, or does this fall under the C/D category?"

Okay. So, for organization, basically I was reading each point and for the C/D.

I was like, "no, it's better than that. No, it's better than that."

And then I would look over at the Pass.

I was like, "yup, yup, this fits. Yup, this fits." So it's a Pass for organization.

I was looking point by point and line by line for both of them CD and Pass for argument.

And this one was harder for me. And I found myself a little more on the C/D side of things as far as the argumentation.

And then I moved on to grammar and then again, it was reading each point, checking the box.

And in this case, I'm like, "no, this is a pass in terms of grammar and lexis."

Yeah, I looked at the part for convention this as well.

I decided that this part is not a pass.

Oh, and then I also noted the percentages on the rubric because then I'm like, "okay, so I've got these two that are a pass and they're the two biggest categories. I've got this one where I'm debating. And it's a sizeable chunk, but it's smaller". And then I'm like, "okay, and then this one's smallest." Well then, I'm like, "okay, this (convention) is only 15 percent anyway".

Now I was actually leaning more towards pass on this essay.

But this was more debatable, so I decided to come back to the essay. Just because I'm one of those crazy, thorough people who's like, "okay, I'm gonna read this one more time and just make sure I'm not missing anything" to basically confirm my decision.

I spent a little bit of time on the introduction. And because I was mainly looking at information from the source material because yeah, because that was the one area that I wouldn't have given it a pass on.

That (convention) I felt, that I was leaning more towards C/D as opposed to pass.

I was reviewing the rubric again. I started with grammar and moved to organization. And then again I'm looking at, okay, this [argument] is only 25 percent. Whereas the others, we had 30 and 30, so 60 percent. So looking back on it.

And then I'm like, "okay, no, I'd give this a pass. I'd give this a pass." So it's a pass.

Rater 2 – Essay 9

Again, I'm reading for just flow and information.

"I believe the negativity overweights."

I did get a little tripped up towards the end. So just a word use.

At this point I was still reading the second paragraph for flow, but also looking for information.

Things are still reading fine. The ideas flow well together.

Then I did note “okay, this is information from one of the sources,” as it brings up the plants and grows cotton, or Americans are undercutting West African growers, that sort of thing. Yeah, so those were some things I noted. Otherwise, yeah, I didn't really note much at this point.

The language seemed simple, I think, was probably something I also noticed.

The ideas are also simple. Not super simple, but simple.

Now I'm moving on to the third paragraph.

This is where I'm noting “okay, yep you're getting more information from your source regarding vitamin A and the different information.” I also noted that they brought up the British medical study from the source. So good.

“That is why more attention should be paid to increasing the quality of food people take around the globe. Instead of spending money on modifying crops genes, more budget should be allocated to improve nutritional facts of foods.”

I wasn't exactly sure if the argument was money would be better spent on the supplements rather than the genetically modified golden rice. I mean, I don't know if I agree with this.

The writer doesn't bring up golden rice, and so in my head I'm like “did you understand the article when it was talking about it was more cost-effective to do the genetically modified golden rice than it was to do the vitamin supplements?” I spent a little bit of time on that thought.

Then I reviewed the first several sentences for last paragraph again.

Again I was noting “okay, you seemed to understand these parts of the source materials.”

But I felt like was ... I wasn't sure what the writer was arguing.

Yeah, so I looked through the article and made notes of the main points.

In this case, there was a lot of information that was not in the writer's essay, for examples, things about companies having too much power, being in control of the food chain, that sort of thing. I thought it was fine. I wasn't holding it against it, but I was noting “okay, none of these was in the essay.”

That's why I spent a longer amount of time up here, was “okay yeah, here is where the West African cotton growers was brought up in the essay.”

Then I moved on to the next article. I was looking at the third and the last paragraph. This is where I was looking at the golden rice.

Again, I was trying to bridge the gap of that argument discrepancy that I was reading in the essay. The essay had written that governments should spend money on fortification programs than on modifying crops. That's where I was like, “okay, but it says here that it's actually more expensive to do the fortification programs than to modify the crops.” From an efficiency standpoint, it would seem that actually maybe governments should spend more money on modifying crops. So, this person actually misunderstood the source text.

I went back to the essay. I read it again. This time I was again making notes of “okay, what am I catching for flow? What am I catching for information? What is there for organization?” Those sorts of things are what's going through my mind.

I was reading this part [the second paragraph] about genetically modified companies

And so, I'm like "okay, obviously they gathered information about cotton and the West African growers from the source text. The writer would have also got information regarding the GM food or the GMO companies, biotech companies as well from the source text." And so, I was thinking about the fact that this person should have mentioned the source.

Yeah, I was moving to the last paragraph and I'm rereading that again for content.

And I'm noting, "okay, here are pieces of information the 250 thousand, 500 thousand and 668 thousand from the source." But again, they didn't cite the source.

Then I again I'm focusing on that discrepancy between what the source text says and what the student says that still hadn't been solved at this point.

For a phrase like "improved nutrition facts of foods", I wasn't even exactly sure what the writer meant by improved nutrition facts of food.

Then I'm busy thinking, "okay, how does this feel to me as a paper?" I was feeling between a C and a Pass.

Here I'm looking at the different points. Now I'm looking at the rubric to help me decide which level is it? I was looking at grammar and lexis.

Again, when I was reading this I'm like, "it feels simple, so for this C/D that fit better for grammar and language for me."

Then I checked conventions and one thing I noted was "sources are cited appropriately".

And one thing I noted was like "I don't think they actually cited anything. They used information from the article, but they didn't actually say the authors' names or even text one or text two." It was just "here is information."

So, for conventions, I leaned more towards CD.

Then I checked organization for a Pass and CD.

The organization again it was on the balance between Pass and CD for me. At this point I'm like, "well it felt fairly well organized." And I wasn't sure where to go for organization.

Then I think I moved on to argument, and then I looked at argument.

I felt like argument was also on the balance. Well, this one I felt was a little more C, D than pass

So then again, based on the rubric, I went with C/D.

Rater 2 – Essay 10

And so here, I'm again reading for flow and content.

I remember noting for the first paragraph, the person says that with this essay, we'll describe both of them and timely come to a conclusion.

I'm like, "this kind of feels like a thesis statement but is not very specific or clear." And so, I noted that.

And then still on the second paragraph, I'm noting "okay, yup we're talking about biotech companies. We're using information from the article." But I also noted that there wasn't any citation in the article or references to the article. It was just hearsay information. So, yeah.

Now, again I'm noting information from the article about American producers and undercutting other nations. Then I noted golden rice from the Vitamin A again so again I was noting information from the article. And then at this point I'd noted, "Okay, yup, we've got Lombard mentioned. And we've also got Vandana Shiva mentioned." So, I noted that as well.

Now I'm at the last paragraph. And, this is where I noted that at the beginning, the writer said, "This essay will describe them both and finally come to a conclusion." But in the conclusion, the writer does not really come to a conclusion other than to make sure that the benefits are reaching all people. And so, it was a conclusion but it didn't feel very strong. It's I guess what I noted.

And so here, I'm reading the source texts

taking notes of... "Okay, here's information that the writer used". It was basically confirmation to see if the information used in the essay actually comes from the source texts.

And the same thing here where I'm looking at information golden rice.

So, yeah, I was reading the essay again

checking for accuracy of the information they use from the source materials. So, the student referenced Vandana Shiva, but from the source text, I noted that she referred to it as a hoax that it's creating hunger and malnutrition and I didn't remember that being noted in the essay. So, even though the student referenced her, I'm like, "I don't know that this part was there in the source."

And then I noted the cost-effective and because I think that was mentioned in the writer's essay. And so then, I again was checking this paragraph for that information. Yes.

So, I was checking for the information about Vandana Shiva. And he did include that she was against using GMO foods, but he didn't go into any detail, which wasn't a big deal but it was something I noted.

And then again, I was looking at the conclusion with, "This doesn't feel like a strong conclusion."

So, I felt like this is a CD paper for me.

So, now I'm looking at the rubric. And yes, and I am going point by point and basically looking at each section. And I focused on C/D because that's what the paper felt like to me. And so, basically at this point I'm confirming "Okay, does it meet these criteria in CD?"

And so, then that's why I decided for ... ultimately decided C/D.

RATER 3 (02/03/2018)

Rater 3 – Essay 1

I first looked at the rating scale just to sort of reorient myself. And so that's usually where I want to make sure I am kind of aligning myself with the rubric to remember what the level B, level C, and a level pass. As far as the language used. The language is used for organization and sort of arguments. Then yeah, just to sort of ... I was trying to remind myself of the percentages for grammar and lexis and conventions. Because I mean that's only 45 percent of the total score, so I always have to keep in mind that if there are grammatical features that are confusing or potentially problematic, for me as a reader, I need to consider them. But I shouldn't be prioritizing them.

But I know what I do is I'm very concerned about organization and the argument that's developed, that's what I prioritize as a rater.

Yeah, so the first thing I do is try to read through the entire essay because it's really hard to start evaluating an essay if I don't understand the big picture of what's going on. I was looking for the overall argument, and how the writer is starting to organize that argument, I need to see where they're starting and then where they go.

I could not find a thesis statement or a topic sentence.

Well, to me this essay read more as a summary of the two articles. It didn't actually make an argument about which one was preferable.

I'm guessing the writer didn't get a chance to finish. So, the assumption I was working on is if they were given more time, they would have made that in the concluding paragraph, that final argument.

But I was thinking about where students need to be when they enter 150, since I've taught that what factored into my score is the fact that they didn't start their essay with a strong argument that then they defended.

I was thinking that I could see how ... When I think of an introductory paragraph, I think of that upside down triangle. Give us the context, move to the thesis. I could see that the writer was trying to give us that background. But then as a reader I was thinking, "Okay, so what? What is your point? Where are you going with this?"

The language for me here was not problematic. I thought it was clear. There were maybe some very minor errors, but nothing that I think is so concerning that it would at least put the student down in a level B.

So again, that this whole paragraph is just really a summary of the main idea.

Then I got into the second paragraph here

And I was glad to see that the student was summarizing one of the source, but that's all that the paragraph was doing. It was just summarizing. Maybe they're sort of an implicit argument that they're trying to make

but I didn't hear any of the student's voice coming through to say ... to make a clear argument and to actually answer the prompt.

"Bananas were hardly acceptable among people for the lack of the taste."

This last sentence really did confuse me, though. So well, I had to go back and read that sentence again, because it was very confusing to me.

But I should say that I was happy to see that the student was trying to come up with an example. But again, what is their argument? They haven't explicitly stated that. What they're doing here. I was again just reading to try and understand where it was going.

I immediately went back to the beginning

because I wanted to try and see, try to remind myself, "Okay, did the student actually try to make an argument? Was there anything in the introduction that indicates their opinion?" That was missing for me. Well, I think I was again trying to look for a thesis statement.

I had an idea then that it was going to be a level C essay.

So, I decided to sort of go through very quickly to make sure again that what my ... What I sort of had in mind was again aligned with the rubric

I went back, at least at the beginning of reading, and looked at this, at organization mostly.

Yeah, and I probably didn't even look at grammar and lexis because in my mind the student didn't struggle with that as much.

I was trying to come back to confirm was there a thesis? Was there any sort of an argument? Again, I did not see any sort of a strong argument. Here I'm trying to go back and double check any sort of a topic sentence, anything that I can pull from.

I felt like I have to double, triple check to make sure, because I really want to be in alignment with what everyone else would be doing and the rubric especially.

At this point I thought, "Oh, I should quickly go through the text to make sure the student didn't plagiarize?" Then I thought, "Well, that's not going to change my score." That wouldn't push them down to a B, or if they didn't plagiarize that wouldn't put them in a P, so that really wasn't worth my time.

I decided it's a CD yeah, not a C+, not a C-.

The way I sort of think of this is for an incoming student, since I've taught 150 and 250 and 314, I sort of have an understanding of what's expected.

After reading this I thought, "Well, it would be very helpful if the student could have maybe a course to understand essay development first before coming to 150." But I don't think that they need to go through both B and C.

The language did not to me demonstrate that there would be a need for an entire semester of grammar (101B).

Rater 3 – Essay 2

Okay, so again, I started by looking quickly at the rubric.

So then, as with the very first essay, I went and I should have read the entire essay all the way through.

This one was interesting to me because there was a very explicit argument here. Well what I consider to be an explicit argument. The writer says, "I think GM should be supported for many reasons." And then they list their reasons. So, for me, there's a thesis here.

The introductory paragraph was less well developed.

especially compared to the previous essay...

So then, I continued reading the rest of the essay

looking for a thesis statement, or a topic sentence and then looking for how the examples supported that topic sentence.

I also, I remember thinking that "Oh, I'm glad to see that they said the author of text two." But what I would have really liked to have seen was a clearer, in text citation, with the authors' name. Because that's something we like our students to do. But I was glad to see that.

So, then I just kept reading to the rest, to the end.

So, one thing I noticed about this paragraph was that, I think that the students was trying to set up an opposing point, that then they were going to maybe disagree with.

But I don't know where they were going because the paragraph didn't finish.

But that I thought was a little bit confusing to make one argument in the thesis, the first paragraph supports that argument, but then the second paragraph doesn't actually support that same argument. It's providing an opposing opinion.

So, then I would have gone back to the introduction to, again, think about the argument they were making. And then I went back to the thesis statement and then tried to look back at the topic sentence for the third paragraph, for the second body paragraph.

So, I was trying to see the structure of the arguments again.

And then, I was sort of concentrating on the citation again, thinking about that I was glad that the student was referencing the text and not taking it for their own ideas. But wishing for a little bit more from them. More of a correct, I'm using quotation marks here, "correct citation".

Sort of what we would be teaching in 150.

In my mind this essay was a C again. It was not a pass. But it was not a B.

And then yeah, then I went back to the rubric and just sort of quickly scanned down to see rating scale.

I was thinking, because the second student clearly made more of an argument than the first one

So, should he, or she, be put in the same class as a student who wasn't really able to make an argument? But I would not pass that student on to 150.

I thought that could be too much for them at this point. So, I wouldn't want to put them in a situation where they can't be successful.

So, I guess I struggled with how do I justify that a students should be in C or B, even though there's the rubric to help me make that justification. So, I did actually go down and sort of look at the, well 'cause I looked at organization, argument. But then I think I did keep going down and look at grammar and lexis to think about what the students was doing and how they were fitting onto the C. But I wanted to compare it with B.

Compared with the first student,

the language here was maybe not as strong. But I didn't feel that it was only simple grammar structures or that they had so many grammatical or lexical errors that I couldn't understand what was going on.

And then, I looked up at C to see sort of compare to that.

I just glanced at the C really quickly

And then I think that I don't care about spelling. At least, if there's a clear issue with a student's not being able to spell, then maybe if that was something more salient.

Then I did consider like this paraphrasing important. And I suppose I could consider this to be paraphrased appropriately. Mostly appropriately.

In my mind, it helped me justify that "yes, this should be in the C."

Rater 3 – Essay 3

This one was especially interesting to me because immediately I noticed that the language was very different than the other two.

Basically, after reading the first sentence, I knew this was going to be a B essay.

Not only were there issues with word form and word order, there were very bizarre spelling mistakes.

I wondered if that was actually an issue with the student not being able to type appropriately.

So, it didn't matter how well the student had organized the essay, the fact that they had such problems with word form, which made it very difficult to understand what they were trying to say needed to be prioritized.

So, knowing the structure of the classes, it didn't really matter how they organized their essay, they should be put in 101B.

Then, I did go ahead and go through and read the entire essay. I had to go back a couple of times and start rereading the sentence again.

I was having a more difficult time understanding what they were saying because there were weird spelling issues

and then word form errors that made it more difficult to read.

But I did see the student does make an argument. They say that they don't think genetically modified crops should be supported because there are side effects. They have a transition. They have topic sentences here. They're looking to support.

I do remember thinking that it was problematic that there were no citations. There were no references to the source texts in the first paragraph?

I remember laughing at "effect the pregnancy mother." I thought that was really funny because of the language. What's interesting is there's clearly an attempt to use more complicated structures, but the student is not able to successfully use those.

Again, basically the student tries to make an argument in the third paragraph that herbicides lead to environmental factors; GM crops use these herbicides; This is harmful to the environment. So, there's an argument that they're trying to explain.

I'm thinking about ultimately if this student were placed in 150 right now, the instructor would have a difficult time working with the student, not because instructors aren't prepared to work with second language learners, but the student needs to work on sentence structure and word form and these more complicated structures before they can really be successful in 150.

So, I already placed this student in B.

I thought the student could really benefit from a semester of grammar instruction.

So here, I was reading. I was looking for the main idea of what they're trying to say.

"those highly technology country", so I thought maybe it's further evidence that the student needs some additional grammar instruction.

I finished the third body paragraph. I was reading the conclusion to see what they had to say.

So, the student managed to write a full essay.

But as you can see, I didn't even go back to read the first paragraph again because I knew it didn't matter how strong their argument was, I was prioritizing the grammar here.

Rater 3 – Essay 4

Okay so, when rating this essay, I approached it like all the other ones. I started by reading. I had to go back and re-read a couple of sentences

Because the sentence structure was a little bit confusing.

“It can be also developed further”, so I think the word order here was a little bit confusing for me.

I was thinking that, overall the control of the language was especially much stronger than the previous one.

But I was again, sort of, influenced by my concern with organization and thinking, “wait so what?” when I was reading this.

I didn't see a clear introductory paragraph.

Then, this student ends the first paragraph, "All-in-all I think the government should have the right to patent the GM technology."

I remember thinking like, “well that doesn't answer the prompt. That's a nice opinion but, that doesn't answer the prompt.”

This paragraph for me, seemed to lack a clear focus. There are a lot of ideas going on. I didn't see a thesis statement that clearly answered the prompt.

So, after I read this first paragraph, I knew that his essay was not a pass

because the student didn't demonstrate an understanding of how to structure the five- paragraph essay.

However, this essay did tell me it was a graduate student.

So, he's trying to provide that contextual information. It almost read a little bit like an introduction of a research article. Maybe not a very well executed one but, more so than this traditional five paragraph essay that we teach undergraduate students.

This told me the student shouldn't be passed.

If it is a graduate student, from being in 101D. If it is an undergraduate student then, they would benefit from 101C in understanding this prescribed five paragraph essay structure.

I kept noticing the not capitalized pronoun "I", which drives me nuts.

In the second paragraph, the student started this paragraph by referring to Lomborg, which told me that they were going to spend most of their argument discussing this article. So, this whole paragraph was going to be based on this one reading.

As I was reading, I was concerned that the student had used language that was too similar to one of the readings. For example, the use of "fortification."

I just thought, "well I think it's a prudent thing to go and glance through the article."

I immediately went to that second article. The Lomborg piece. Then I started reading through trying to find where they talk about vitamin A deficiency in golden rice. Sort of looking at what Lomborg said.

I could see this is an attempt to paraphrase, but the author, Lomborg, does use the word "fortification." The student doesn't copy directly, there really is an attempt to paraphrase. I wouldn't call this plagiarism, I would just call this an attempt to paraphrase that was mostly okay. But I thought more practice could be beneficial.

Yeah, I decided to give it a C.

I thought the student could benefit from a semester of explicit instruction about how to structure an essay.

Rater 3 – Essay 5

So, I went ahead and jumped into reading. I had followed the same procedure, started by reading.

I felt like I was having a really hard time concentrating on this essay and I don't know if it was just the structure of the sentence that took me more effort to read.

I'm reading here.

I was happy to see that the student did get to a thesis statement.

"does do much more benefit."

The grammar errors were causing me some difficulty. It wasn't impossible to read, but it was just slowing me down and making me read slower, I guess.

Then I started reading the first body paragraph.

I mean, the student tries to cite the second article but doesn't do this in what I would guess expect because there's no reference to the actual author or the title.

"Many people who cannot afford food would survive in result of GM Technology."

So, I think it's an issue of word choice.

I think they want to say “because of.”

I sort of hovered over M2 for a second because I couldn't understand what they meant. I think they mean “meters squared” or “squared meter” but that took me a second to try and figure out what was going on.

I mean I can see that the student is trying to develop an argument here.

But again, I'm seeing problems I think with word order, word forms and word choice that is making this more difficult to read.

The fact that I'm struggling to understand the language is sort of at this point, I was already thinking like okay, well this is probably like B, B+.

“But if we take our eyes further, the worry of the first article isn't fake.”

This is again, they're trying to make an argument here. They're trying to say that this is a concern.

The thing is this is not what I would consider a very academic way to say that. So, the student doesn't have a strong control of academic language.

I was wondering if some of this is an influence from the L1. It would be interesting to know.

I was thinking of giving them a B. I knew I needed to give them a B.

But honestly, I felt bad giving a student a B because I know how hard many of them have worked to get here. So not only do they have to do TOEFL, they do the EPT. Then they have two semesters of ESL classes before they can take 150. Then they've got 150 and 250, I mean that's two years of English classes.

I felt like if I'm going to give a student a B, I'd really, really have to justify it in my mind.

So as you can see, I think I was bouncing around, going back and reading it again, sort of looking for ways that I can justify this decision to give them this B and not a C. I looked at the whole essay to again to make sure there isn't something that I've missed or maybe I've read it wrong.

After the first body paragraph, I believe at this point, my thought was that “I needed to go back and look at the rubric.”

So, what I was looking at was in alignment with what we say for B. I was looking at all four categories for B.

I was thinking that “I needed to consider all four of these categories under B, not just organization.”

Then I think I jumped up and look at C. I sort of like going up, I just quickly looked at C.

And then realized I wouldn't characterize this essay as “a good control of grammatical structures.” But I wasn't sure, I was still doubting myself cause with grammar, I don't feel like I can make that strong judgment. I was trying to decide.

Then I think I might've even gotten to look at the essay again.

But in my mind, I'm thinking this is not a C, but is it really a B? So I decided on the B+. I suppose that's where this B+ category sort of gives us some leeway to say, well what else can the student do, can we think about this?

Rater 3 – Essay 6

I immediately started by reading the essay. Again, I approached the essay the same way. I started by reading the whole thing.

As I was reading, I was looking for certain key features, thesis statement, topic sentences that support a clear argument that answers the prompt.

"In my perspective, the genetically modified food should not be eaten."

I did see a thesis statement

but it didn't answer the prompt. The prompt wasn't whether it should be eaten or not. The prompt was indicating whether you think GM foods should be supported. Immediately, I was thinking, "Well, the student isn't even answering quite the right prompt."

So, to me that automatically excludes the pass category. That automatically would put them in a C or a C+ range.

But I could see they were trying to structure their essay in a proper manner, whatever proper means.

I had to actually go back and read this first, yeah, I had to go back and read this first paragraph again, but then I moved on to the first body paragraph.

I noticed that okay, they had a transition.

They're trying to make an argument.

"It will hurt the farmer's benefits."

The student goes on to then explain what GM crops do and don't do.

I was wondering how is that totally relevant to the argument they're trying to make. I was a little bit confused as to sort of the overall scope of this paragraph and the purpose. I wasn't totally sure what was going on or how it was related.

Then, I went back, and it looks like I reread this.

Then I see that the student has a reference, not the exact reference we're looking for, but at least they say that. There is a quote.

But suddenly we're transitioning from a little bit of background information about the GMO's to West Africa. I was kind of confused about the connection. There seemed to be some cohesive devices that were lacking.

So, the overall clarity of this paragraph was confusing.

Then I moved on to the second body paragraph. I read the beginning of the second paragraph twice.

I think I was thinking, "Oh, I'm glad to see a reference," but again, the way they wrote it wasn't exactly what we want."

I mean I finished reading the second paragraph. Then I moved on to the conclusion.

I'm happy to see the student was attempting to structure their essay in this five paragraph structure,

but then I needed to go back to the thesis statement again to think about what was their overall argument

because it felt like there was an attempt to have a connection between the ideas, but they weren't fully connected.

I think at this point I'm thinking, "Well, this really isn't ... Maybe this needs to be somewhere between B and C. The language isn't what I would consider to be so indicative of a B essay, or indicating that the student would really need a whole semester of grammar instruction, but this essay doesn't seem as strong of C essay to me

(as strong of C essay to me) as some of the other ones."

So I wanted to go back and read again so that I could think about B or C.

Yeah, so I went back and I reread the essay. I moved to the introduction reading it again.

I was confused as I was reading.

There were a couple issues of phrasing that sort of tripped me up

but mostly I was trying to understand how all the different pieces connected. Whether it's lack of cohesive devices or whether there's sort of a divided focus I don't know, they're not actually developing one main argument.

I thought "I'm trying to give the student the benefit of the doubt and saying I might not understand rather than maybe the other way around, where I should trust myself more and be more critical of the student."

I was also thinking, "Oh, the student also looks like they're trying to summarize. Is the summary too close to the texts?"

They clearly tried to use their own language here, for example "It will cause the blind or dead." I mean these are their own words and they're not plagiarizing.

"Other species will have negative properties. Some of them we cannot found."

I was laughing. I feel like I'm evil, but the way they write is quite cute, actually.

But there wasn't anything suspicious that would indicated that they accidentally plagiarized

So, I decided that I wouldn't need to look at the articles.

So, I decided this is a B+.

Rater 3 – Essay 7

So, I started as I normally do looking at the essay.

By reading the very first paragraph, I'm looking for a movement from sort of general background into a clear thesis statement.

What I was really impressed with at the beginning here is that the student uses the language "supported,"

which I often find is very helpful when they answer the prompt in their thesis statement very specifically by using similar language of the prompt, or if they had used a synonym that was very similar to “support”.

To me, this says this is a very clear thesis. "I think the genetically modified product..." They have a strong thesis statement that answers the prompt. So, as a reader I knew where they stand and I have a sense of where they're going to go in the essay.

Well, interesting word choice there, "should be highly supported and be widespread on earth." So, at this point I was thinking, “oh, this might actually be a pass essay,”

So, then I moved onto the first paragraph

and I was thinking, “oh, the news reporter. So, which text are we talking about?” I think the student is trying to indicate that they are summarizing one of the texts, but there's not really a clear citation here and that was confusing.

I was also a little bit confused because there was no sort of transition

or topic sentence. So, to understand how this paragraph was sort of the first part of their argument was a little ... I wanted something more explicit.

And then they go on and actually cite the American Journal of Clinical Nutrition, and they've got a quote. So, I was happy to see that the student knew how to use a quote, at least not just take language and use it, and then they even have a reporting verb.

So, as I was reading I was thinking that I was happy to see that they were trying to pull from a source to support their argument

but they hadn't made this particular argument explicit in a topic sentence.

In my mind then I was thinking, well, okay, immediately I might've put this closer to the pass category, but now I think they're sort of moving down the scale a little bit more into a C, a C+ range.

Then I went onto the second paragraph

and, again, I don't really see a clear topic sentence,

and so it looks like I went back to read that a second time and even a third time then

because I was thinking, “wait, wait, why are we talking about biotech companies now?”

So, I think this also sort of told me like, “okay, this is not a pass essay anymore because now I'm a little bit confused as to what they're trying to argue.”

They go on to talk about like, "the government should change their policy," but to me this isn't clear how it wasn't clear how it was supporting their argument that GM products should be supported. Yeah, as I was reading I was thinking, “okay, where are you going with this? What is the point?”

I mean, there was maybe the language isn't quite as strong as in the first paragraph

but I think I was confused because I didn't have a clear marker, like a topic sentence to tell me what the purpose of this paragraph was.

I think as a reader, I expected the student to be really explicit and to guide me through their argument.

So, yeah, then I went on to read the concluding paragraph. "Now the GM product is as being developed."

There was something with the language that was a little bit confusing.

I immediately went back up reading at the essay to sort of look at, okay, what was ... the main argument here. And at this point, though, I was just kind of skimming very quickly ... through ... But it's almost like I just want to double check that what I remember from the thesis statement and the first topic sentence is indeed correct.

Yep. So, I ultimately gave this essay a C+.

Rater 3 – Essay 8

So yeah, I started by reading this.

And I was interested to see the very first sentence was almost like a thesis statement because "should not be supported because." So that kind of sort of surprised me that the student started with a thesis statement at the very first sentence of the introduction, which immediately indicated to me that maybe they don't fully understand how we typically expect students to structure essays.

And then they go on it looks like to almost try and support one point.

But yeah, they're talking about independent citizens playing their own lifestyle. I was sort of thinking like this seems irrelevant.

I was thinking the student is like into government conspiracies. That's the first thing that came to mind.

I wasn't quite sure why they were ending this paragraph in this way because it doesn't follow the expectations that I have for an introductory paragraph. I expected to see a thesis statement at the end. I expect to see a bigger context moving down into a clear thesis statement that makes an argument and lays out the structure for the paper.

So immediately I sort of ruled out pass as a category.

So, then I moved to the first paragraph and I was like "oh. Okay. Well, it's nice to know you're from the Dominican Republic, but where's your topic sentence? And this is not an appropriate topic sentence.

I thought it was interesting that the student was trying to pull from their own experiences to support their argument, and that starts to come through in the second paragraph. Clearly the student has very strong opinions about companies that produce GM food or grow GM food in Dominican Republic, which tells me the student has some experience with this, and I think that's a really interesting way to help them support their argument.

Actually, I think the language was quite clear here. They demonstrate control over that.

But I'm still nervous about the fact that there wasn't a clear topic sentence to sort of guide me. I was wondering what is the point of this paragraph? They didn't prepare me to understand that.

So then again, I was moving onto the third paragraph.

I was like, “why are you telling me this? Like I'm glad that you're a member of a church, but this is not a topic sentence.”

This suddenly sounds like a personal essay to me, which is not appropriate for academic writing here.

Then, I went back to reread this to make sure no that there's no topic sentence here.

The student's talking about having gardens and improving health, and so suddenly we've moved away from what I would consider the argument of the paper to be, just sort of irrelevant information.

So then, I went onto the concluding paragraph.

And I was again confused because now the student is summarizing the text. So, as I was reading this, yeah, I'm seeing summary of what is going on.

but then when I got to text two. Again, I saw fortification”, which I know is the language from the text, and then I focused on “posterity” because I was really surprised that they used that word. But then I was sort of like “oh, I wonder if they too heavily relied on the text or on the language of the text 2 in their summary.”

So, I went back again to reread this because I wanted, again, I'm sort of double checking to make sure I didn't miss a topic sentence.

Then next, I went to look at the second article that the students were given.

And I wanted to look at the language that the original writer used to see how it matched what the student had written matched this. I think again there's an attempt to try and use some of the student's own language, but I mean, “fortification” just really stands out for me.

So, this is an attempt to summarize, not a totally successful one, which to me indicates that the student could benefit from like a semester in 101C before they are put into 150.

I was at this point I was sort of struggling with C or C+.

It looks like I read the introduction again and then quickly topic sentence, topic sentence, or lack thereof.

Oh, then I did look at the rubric. I think I just wanted to quickly make sure that I was aligned with the rubric. It looks like I focus again mostly on organization and arguments, and then just quickly looked at grammar and conventions and then moved on.

Ultimately, I ended up giving the student a C.

I didn't think that they would have a hard time with their control of the language in 150, but the student doesn't necessarily understand how to structure an essay and in a way that we would expect in 150 and we teach. So, my thought was this student could really benefit from I think one semester in 101C where they would be able to focus on development of an academic essay.

Rater 3 – Essay 9

I started this essay by reading.

The beginning of this essay was a little bit confusing for me

So, I was already trying to work to orient myself to understand what the student was saying.

That was an indication that this essay is definitely not going to be in the pass category.

It looks like I even went back and reread this.

"I believe the negativity overweighs."

So, there's a problem here with the structure.

In the thesis statement, it was a little bit confusing.

But it seemed like the student was trying to argue that they do not support the use of GM foods, that was the assumption I had to make.

Here, I was already putting this into the "C" category and maybe even considering somewhere between "B" and "C."

The language is influencing my ability to understand their argument; so maybe there's some language problems here.

Then I moved to the first body paragraph, and again, I was reading and looking for a clear topic sentence, and it thinks it looks like ... Yeah, I went back and reread this.

To me, there wasn't a clear topic sentence here.

I didn't understand why they were talking about bio-tech companies.

Yeah, as I was reading this I was thinking, "If I have to work to understand what somebody's saying, that doesn't mean that they have a strong essay."

I think I'm on the beginning of the second paragraph, here, so I was trying to read this, but it looks like I had to go back and read this first sentence again

Again suggesting a unclear topic sentence to orient me.

And I was confused because the student said "On the other hand," which suggests that they were going to provide an opposing opinion, which I understand is one way to structure this argumentative essay, is to provide a counter point and a refutation but they don't do that very well.

Then, I was immediately, I was interested, because they said "A British medical study," and I was thinking "Okay, well, they're trying to cite one of the texts," but they didn't actually cite the author.

"Vitamin A-deficiency kills this number of children under the age of five every year."

In my mind, I was thinking, "I think this is exactly from the source text." It's not a bad thing, but I guess if they're going to use statistics, I would probably expect to see quotation marks.

The student goes on to talk about how budget should allocate improvement of foods

and I was thinking, "How is this supporting your argument?" I think they were trying to provide another perspective to say... well, but I shouldn't have to hypothesize what they were trying to say, right? That should be explicitly outlined for me.

And I think I was trying to, at this point, decide, "Well, should this be a 'B' or a 'C' essay?" Ultimately, I ended up with "B-plus."

Because of the "B-plus" I'm going to give, so if it's getting into that "B" category, I've got to justify it.

I went back to the introduction, I think I actually reread the whole essay again, because I was confused.

Because I didn't feel that my confusion was because of linguistic issues, it was because of the structure of the text, mostly. But, I still think there were a couple of issues with the student's phrasing.

Yeah, and I think then at the middle of the second body paragraph, I was like, "Okay, I want to go and look at the original source text." And I realized "No, it wasn't the first one, I need to go and look at the second source text."

I was looking for the statistic, and I think, yeah ... The exact citation. Sort of the exact, "Vitamin A-deficiency kills this number of children." So, they're trying to use the source text to support their idea, but they just can't take a chunk of language and insert it in theirs. Right?

Then I went back to the essay.

I was looking, again, for where they had the citation. So, there's actually two different areas where they try to bring in information from that second source text- And they're having problems trying to cite that, as ethically as I would like.

And then, it looks like I went back and looked at the text again. I think this was my way of sort of like double checking...

So, I gave the student a "B-plus."

Rater 3 – Essay 10

I went immediately to reading. I read this first paragraph.

I couldn't understand it

so I had to go back a couple of times, and sort of try to understand what the writer was saying.

Then, I looked for a thesis statement.

Here they said they're gonna describe both of these perspectives, and come to a conclusion.

Immediately, I thought, "Okay, this student doesn't fully understand that we would expect to see a thesis statement at the end of the introduction." It looks like the student was trying to give a thesis statement, but it's not actually a thesis.

So in this attempted thesis statement, the writer says that they're gonna come to a conclusion, which to me suggests that the concluding paragraph is actually where they make their argument.

I went on to read the very first body paragraph, but then I went back to read again.

I was a little bit confused to why there was no topic sentence, or at least what I would consider. Again, this is indicating to me that I was a little bit confused about something, or something was missing.

And already in my mind, I'm thinking, "Okay, not a pass essay because I'm not seeing these very clear markers of what a pass essay would be."

I went on to read the second paragraph.

So then, I was already confused. Okay, so the student starts talking about golden rice, but what is the point of this paragraph? I did not see a clear topic sentence.

Okay, I'm glad that we see an APA citation, that was very exciting.

Actually, this indicated to me, this was a graduate student.

When the student goes on to talk about Vandana Shiva, I'm like, "Oh, this is interesting. There's no elaboration here. This is just sort of an essay of a bunch of statements." I feel like there wasn't any argument. There really wasn't anything very strong.

So, I went back again, and read the first paragraph, or the first sentence of the second body paragraph.

I was trying to understand what is the point of this paragraph? What is the argument? How does it fit the overall thesis?

I went to the conclusion. I really wanted to make sure I got there, and to read that.

This isn't how I expect an essay of this type to be structured.

At this point I thought I wouldn't put this student down in the B category. But they're not in pass, so clearly the only option is like C, D.

I thought they could use that instruction in 101C.

So, I went back to the beginning, I think actually, I read the whole essay over again. I just needed to read it again, to sort of support in my mind.

Again, the language I thought used a wider range of vocabulary.

The sentence structure was more complicated than the others, which made me think it was a graduate student.

It looks like I went to the rubric then.

At this point, this is more just like a habitual thing to review it to sort of make me feel better that I didn't give the student a pass.

So, I decided a C/D.

RATER 4 (01/18/2018)

Rater 4 – Essay 1

So here I'm just trying to get a feel for what we're looking for, for each level. I spend more time on organization just to see what the main differences are. I don't look at each block, that's just me. I just read this recently, so I didn't spend a lot of time on it. So kinda moving back and forth to compare what else should I be looking for. In argument and details. Just kind of moving from the top to the bottom, honestly. So I did spend some time on grammar because I feel like that's a good way to ... it's a way of separating some of the levels if content is similar. I didn't spend a lot of time on conventions. I feel like I've got a pretty good understanding from doing this before.

And I don't think that's something I rely on as heavily when I do when I do my grading.

Okay, so, first I'm reading through to find out if the content is appropriate if they're referring to the right information.

So far it doesn't look like anything jumped out at me, it's a big problem, I'm not lingering on anything in particular. It seems pretty clear in the first paragraph. It didn't have any big problems, there are a few things but it seems pretty straight forward.

One thing I do remember looking for is whether they're taking the text too directly out of the source, so I think I was watching for that to make sure. Cause they'll have a structure that seems pretty advanced, but they just took the language directly, or almost directly from the source. I was look for that.

E:I went back, I think, because I noticed we have an error in the verb form there, "genetically modified food had been growing more and more in recent years", which should be the present perfect.

So I think when I saw that, it made me go back to check the previous ones to see if the verb thing, I'd missed it earlier. So I think that kind of caught my eye, so I thought that I better keep a closer eye on the verbs.

Yeah, I went back to the beginning. Yeah, I guess I didn't go all the way back to the beginning.

If everything's going well, then I can focus on content more than more.

Yeah. So I feel like I've got my feel of that first paragraph, then I went back to check the verb forms, so I kinda had an idea of their level that way. I think it's, I'm saying it's kind of intermediate level right now.

It's not using any ... the structure, something I look for is the subordination, things like that. If their clause structure is more complex because they're using relatively simple syntax, if they're using syntax and they make mistakes, it makes me think they're using simple syntax and its correct, but it means another thing, so I'm kind of comparing, seeing what kind of structure there is.

Because if they're trying more complex sentences, making a few little mistakes, it's better than if they're using very simple one and it's correct.

At this point I feel like it was a C D level, probably.

[reading second paragraph - inaudible 00:04:38] "... if it was clearly providing a ..." I don't know, I looked at that, to be honest. I think maybe it's just in the center of the page, and I was taking a second to think-

So I think things are going pretty smoothly at that point.

I think I notice my eye lingering on the verb there, so maybe I'm watching out for whether verb form's correct or not. Ah, yeah. Then "have to be", there. That's another verb, maybe that's why I'm sticking on it.

Okay, "vitamin ..." I think now I've finished the paragraph, but now I want to kind of, well before whoever, I finished it, just kind of glanced back over everything else, just to make sure I understood everything. And I see myself moving a little faster now as I get toward the end.

because I have enough information that I think that my idea of this intermediate level is fairly solid.

But, yeah, I lingered there on, "but to extent", because it ends so it shouldn't have a period there-so I noticed that that's incomplete.

So that goes back, I think, to the idea of what kind of syntax are they using? Are they using simple sentences or complex sentences?

I went back to the first paragraph at the time. I'm spending a lot of time on the first paragraph. maybe because I was wondering ... It seemed like, because that's pretty good but I want to make sure it's not ... make sure it's their words. I think that's what I was doing there.

I think it's always good after I get to the end, you have an idea, go back to check to make sure it's consistent. Just to make sure I haven't forgotten anything that might be important or to kind of confirm my intuition on it.

I didn't spend a lot of time reading this one [source text 1], because I just read it yesterday. So I just wanted to glance at it, too.

One of the reasons was to look and see if they're taking specific words or sentence structures-just copying them, basically. I'm looking to see if it sounds like they're synthesizing this information or if they're just reading and copying parts.

Yeah just seeing if it's, if it seems like they really understood it or they're just taking a few details and putting them in their own writing. I was looking for specific sentence structures. So if they're saying it the exact same way, it tells me they might not have completely understood it.

... then that would tell me to question whether they're higher level, cause if they're using some advanced words or structures, maybe I think they're high level, but then if I look and see it was taken directly from the original text, then that would make me question whether they are that advanced.

I think I was comparing to their first paragraph. Because that's where I saw some ... where it seemed like it may have been. I think "leverage", maybe that word. That's a pretty advanced vocabulary word. I don't think lower level student would know the word "leverage". So, maybe I was looking for that particular one I think. That was a word that stuck out to me. I don't remember if it was in the original, but I think I was checking.

I think checking this is important, but I don't think it would change it (the grade) dramatically.

It didn't seem like there was anything that was definitely copied.

I went back to the essay at the time.

I think I was scanning to see how what they said related to the original's wording, syntax, vocabulary.

Honestly, I remember wondering, "do they talk about bananas in the original or is this a student bringing in their own"? So I couldn't remember, I think I looked back (the first paragraph). I don't remember if that was in there or not. It does seem that I was lingering on that banana section a little bit. Because I didn't read it very carefully, the original source, maybe I should have, but I didn't remember bananas being a big part of it, but they do spend a couple of sentences talking about bananas.

Cause sometimes students are lower level, they'll kind of change the subject a little to talk about what they can talk about. Even though they didn't understand what they should be talking about.

If they know something about bananas and they can write about it, maybe they can avoid the difficult parts.

Rater 4 – Essay 2

Alright, so this time I spent a little more time on the rubric, just to think back on what I'd done on the previous one, to kind of check myself. Specifically, I took more time with conventions, because I did not look at it for a long time before something ... okay, maybe I should spend a little more time. But, given how it went before I thought I wanted to look at more detail.

So, yeah. I tried to look at the ones that I hadn't before. Or that I've only briefly looked at. Yeah, so I'm looking at the grammar and lexis part, because I talked a lot about that before, and I want to clarify that that's a valid way of judging, cause that is a big part, I think, of what I look at. What the structures are that they're using, so I did spend some time looking at that carefully.

So just getting started, seeing what kind of content we have here, what their main argument is.

But the way the sentence is structured, it says, should food production, "GM food production should not be supported" and then "GM production should be supported". So, the way that that's written, it's easy to miss that because it's so parallel-

-I wanted to make sure I understood what they were saying because I almost thought they were repeating themselves, so I want to make sure. It's like the same sentence twice, with that not being that's the only difference, so I looked at that more carefully cause it sounded like repetition.

I think I looked there, I noticed it lingered on "supported", cause they already said "supported" twice, if somebody is repeating the same vocabulary a lot that shows me that they don't, they're lower level, they don't have a wide vocabulary.

Yeah. I went back (second paragraph)

to make sure they were talking about the right source text

and make sure they have their arguments clearer.

I think I kind of revisited that sentence "for example, golden rice" because it seemed like they're just listing some facts, so I thought I want to make sure this is their thought, they're not just pasting in facts, basically, from the source text.

And when they're given these numbers, when I see a lot of lists of numbers and facts in the essays I look at, that makes me question how much are they synthesizing the information or are they just making a list. So I think that's why I looked at that carefully to make sure it was ... they're doing some kind of analysis or synthesis of it.

I looked at "deficiency" a little bit more. I think that's a word that's not common vocabulary, even for a native speaker this low level, maybe, they might not use that word.

So that may have also been a signal that we're looking at a repeated list from the original text.

I think I linger there on "supplement" because they left out the article (grammar). I think, can't see it.

One thing that I remember, though, they say, "which means all 500 thousand children who goes blind each year could be saved", I thought that "all" stuck to me. Like, that's not necessarily what

the source text was saying. "All of them could be saved", so I was thinking "why did they think this means 'all of them could be saved'"?

I think that's why I kind of scanned back over the paragraph to see if they know what it is they're trying to say.

I think I looked at the number there, like comparing these different numbers, why do they think this means "all of them", cause there's no amount of money available. So why are they thinking "all of them"? That maybe why I was, I don't know why I'm looking at what it mentions text use specifically, but that might've just been my starting point to go back and scan-

Okay I think I went back there to the beginning for the last paragraph

because we have this signal phrase of, "on the other hand", but then it mentions text two, and I expected them if they're making a contrast, to talk about the other text.

If they're comparing two arguments, now it looks like they're comparing one argument with itself.

Okay then I went back and there's a verb error there. "There are chance". There's no article and it should be singular, so the verb agreement within that one is wrong.

I'm reading that one more carefully, trying to register, because we're really only talking about one of the arguments right here, I think. So now when I went back to the beginning, at this point, I'm thinking, are they talking about both of them? So I wanted to see how they framed the overall, are they laying it out as two different sides of an argument, or not?

So yeah, I was trying to clarify what their understanding of the source text is. Because to know if they're high level, part of that is whether they're understanding all sides. Because they could write enough text about just one of them, but they should be discussing both. If they have good reading comprehension and all that, should be bringing in both sides.

That's part of the assignment.

Yeah, I think after I finish, I like to go back and at least just scan everything to make sure I didn't miss something important, especially if something was a little unclear. If it's all very clear and good, I might not spend as much time reviewing.

Then I think, maybe I was looking back here at some verb agreement issues, but overall it looks like I'm just kind of checking it.

Oh. I think I remember, I noticed the "however", and I was wondering, are they using a variety of these transitional words. And I remembering seeing "however" before, so I went back and I found one here, and I found one here. So that tells me they're not using a variety of the structure words.

I did read the whole paragraph, the first paragraph again. Yeah, yeah.

I remembered that this task should bring in their thoughts on it.

So I think that's why I lingered there a little bit on "in my opinion", so I could see if they are expressing their opinion, if they're bringing in these arguments to support what they think about it instead of just summarizing, like the other task.

Yeah, going back to the third. Maybe just kind of checking myself, honestly.

I know I earlier on talked about the “on the other hand thing”, and that is that showing contrast between the two texts, so I was just thinking about that again.

I probably, at this point, have an idea of the grade I want to give, but I know on this one I was on the fence between a B plus and C, B minus, so I'm trying to find something that can help me make that decision.

Now I went back to the second paragraph.

Yeah. I think that “was a part where I had thoughts of, are they just taking something right out of the original, or is this ... also, how much are they including?” because there's a lot more that was said in the original.

And I think shortly after this, I'm going to go back and look at the original to see if they're missing any of the important points.

(reading the source text)

So now, scanning that argument, the original text I again to see what ... now I notice return to that because

they didn't talk at all about the effect West African producers. I think the fact that it appears in the first paragraph of the source text, means that it's kind of important, I would have expected them to include it, maybe.

I then I kind of go here about African crops. So that wasn't included in their paper at all.

I was looking for things that they may or may not have included.

Yeah, source text number two. Also seeing if they included the important points or just taken the first thing they saw and put it in there, and that's kind of what it seemed like. They didn't talk about [Vandana 00:11:11] Shiva, they didn't talk about the amounts, like how effective it is.

I think that was an important part that they didn't talk in detail about whether the golden rice is a good ...

cause some other essays I saw they misunderstood or didn't understand that whole section where they're saying that one source said you need to eat a lot of it to get the right amount of vitamins, but another source said that you can eat just a small amount and get enough.

Some students just took one side of that, or didn't understand that it was saying that it's actually helpful. So I was checking with that to see if they included anything about that. So that was about content.

Rater 4 – Essay 3

It started kind of jumping around. Uh, I think I noticed this "the alter of the gene", it's kinda confusing that first sentence. I spent some time going back over it at least once or twice. I needed to make sure to see what's the problem here. To see if I can understand what they're trying to say.

Looks like, can't tell it's off a little bit but, I might have been going back again to that first sentence little as I was looking at the next one. Looks like it was jumping up and kind of relating out to the previous part, maybe.

I think at this point, I'm still doing a more broad lookat it. I tend to come back to details later. I think I will come back to that paragraph, but I like to scan it first, usually, or at least I think I do

and, then later look at more details. I think there's more to look at that first paragraph, but I need to see what else is there first.

I'm moving to the second paragraph now.

I'm starting to have a, I'm lingering there toward the beginning probably because of a verb issue. "Being produced will be chaging" (*misspelling on student essay*). I think that's "being produced by changing", I guess, is what they should be saying but I was just trying to clarify that.

So moving further there's this, "remained" is misspelled and I think that made me stop.

And also "toxins" came out as "toxidness", you know the spelling.

I do remember specifically seeing that word and thinking it was a little funny.

So, I remember this sentence, "has been affected" is not the correct structure there. So, "it has been connected that GM", so it shouldn't be, "GM has been affected". They're using the passive there but that's not what they should be saying.

So, I remember spending more time in that sentence because I don't think that was part of the original, something about pregnant women, I don't think that was a part of the original set.

I don't know where that came from so I put a little more thought on that part to see if they bring in stuff out.

Because they're supposed to be making references to the original two texts.

They're using the passive a lot and I don't know why. "It has been affected".

So maybe they meant to say "it has been affecting" or "has affected" but I think they're using the passive excessively. So they're trying, one of the parts of the rubric talks about, trying to use more advanced structures but failing to do it.

So, that's why I looked back at that so hard.

Are they trying to use more advanced verb structures and are they doing it correctly?

Now moving to the third paragraph.

I looked at "intolerate" because it was misspelled.

I don't think that spelling issues are a huge consideration for me honestly, because they're (students) timed. It's something to look at but I don't think it is as that important as other factors but it does make my eye linger (on the video recording his eye movements when rating the essay). Even though I don't think it's important it does catch my eye.

So this part [the first sentence of the third paragraph], I went back

because as I finished it I thought, "does this make sense logically?" Because they're saying, "these crops are resistant to herbicides and so GM crops..." Then they say, "that it will be a problem because they use more herbicide to kill those crops" but why would they want to kill these crops? They're food crops. So this person is saying that they'll need to use more chemicals so they will hurt the environment which that's a bad thing. They're not trying to kill these crops so I thought, "why? Am I understanding that right? Why would they be saying that?"

So I guess it shows some confusion about the topic.

And I also noticed that they said "cm" (*misspelling of GM on student essay*). I saw the "cm"

and I glanced back up

to see if they've been using it correctly before.

The whole paragraph, the whole third paragraph, is kind of logically confusing so I spent some time on that trying to understand, "is that what they're saying?"

"Harmful to the environmental" so that was a word choice issue.

Yeah, so I did linger on "strongly use herbicides to kill crops." So I lingered there on "kill these crops" because these are food crops we're talking about. We don't want to kill them.

And then at the end, "causing the extinct to those animals".

So that's not correct at all, there are some problems there, so I was just lingering at that one and it's spelled really strangely.

I'm kinda bouncing around there (reading), nothing jumping out too much.

Oh, yeah, "Highly technology countries." So there are a couple of syntax or grammar errors in there that I caught.

Paragraph four. Beginning with, "further information". So far I'm just noticing some word errors

But then, "those under privilege (misspelling from student essay) countries" I don't know why it stuck there.

And I noticed the, "can't" part, even though that's just a silly typo that I would do.

"conduct an experiment" so I think here I'm trying to see if they understood what they read. So they're bringing in the part about those underprivileged countries; and are they talking about that in the same way that the original text did?

The word "monopolize", I do remember noticing that word.

I imagined at the time that it probably came from the source text because there are simpler words that are causing problems.

"Because of poorness", oh yeah. "Because of poorness", that sounds of like a lower level structure, so I didn't look that.

Then went back and looked at "a poor country"

to see if they're repeating the same word or nearly the same word a lot.

So I think the second to last sentence in the last paragraph is says, "even they have benefits in other ways."

So, when I saw that, i thought of this subordination here.

so I went back to the sentence before

to see if it should be connected with that sentence, if it's an error, or just a punctuation, or it's a syntax thing. So that shows that they're trying to use subordination.

So I think at this point, I finished it; so I'm kind of glancing back through it

to remind myself of what I saw before and sort of confirm or change my decision about the level.

In paragraph 1, I'm noticing the word "scientific". So, it is spelled really strangely and it might be what caught my eye.

Yeah. The second paragraph, scanning it again, nothing is staying for too long at this point.

Yeah. Paragraph three seems to be going pretty smoothly.

But now I'm going back to that same thing I talked about before where it seems like there getting the wrong idea of what the topic is about really. Because these are supposed to be good crops not crops we need to kill.

If we're trying to test their ability to do academic work, they need reading comprehension. It's important and I think it's part of why we're testing here. So, if they miss the argument by a pretty by misunderstanding I think, it should be penalized.

I'm revisiting that [the fourth paragraph].

I'm going back to three again for some reason.

I think in trying to decide how important that should be if they misunderstood that argument how big of a penalty should that be?

Because if they're timed, I can understand how reading it quickly how they can misunderstand a little. So, it's sometimes hard to decide.

If they're describing it well if they're making an argument that is reasonable based on one small misunderstanding, then I think that it shouldn't be a big penalty because they're expressing it well.

I think that this is a debate in my mind right now. Yeah, spending a lot of time in that.

So, looking at that last paragraph, the conclusion there.

I think at that point I decided a B+, and I have read the source texts, so I don't need to spend a lot of time on them.

Rater 4 – Essay 4

I did glance back at a few things (the rubric). Something that I had in mind was, I've seen essays are organized but they're organized in a really simple way. I was thinking about how should I consider that. If they say "firstly" and "secondly" and "in conclusion", how should I treat those kinds of structures? I looked back in the organization. So, I wanted to see what, what level and this one I lingered longer on the CD because of the talks about specifically in the part that says "simple, cohesive and transition devices". That's what I had in mind. If they're doing that structures would probably B, C/D in terms of an organization. I had what I wanted. After that I moved on. I didn't look at the pass one.

Okay. Here we go. Just kind of reading through, not lingering on anything in particular.

I lingered there because there's a "be" missing.

"They have yet to be genetically modified". So I stopped there because I noticed something was off and I looked at it again.

But I did think that was a more higher-level structure to say, something have yet to be like that unless they just learned it as a whole phrase. I think it shows that they're at least trying to use more advanced structures even if they made some mistake.

The word "exponentially", I don't know if that was in the original text, but it's a good word to use. So that was important.

I think here, I'm not really stopping on anything in particular, but the sentences are longer and more complex than some of the other writing has been, so maybe I was just checking to see if they accomplished it correctly.

This is fairly complicated sentence. "I think it must support the research for GM crops plants".
 "But as we can see in the paragraph by Tom Chivers, the corporate companies have gained the most full control over this technology and this is detrimental for".

This sounds like a pretty advanced sentence. I think I was reading it carefully to make sure everything is being used correctly.

There's nothing that's really tripping me up. It's pretty high level. Honestly, I don't see anything there that seems like it was a problem.

It looks like I went back to the beginning of that sentence starting with, for example, "the farmers". I don't recall thinking about any problems there.

"Therefore, I think the GM technology should be in the hand of the government so that it could look after."

Moving forward. The content is getting through clearly there's not any form issues that I'm noticing to stop.

I think compared to other ones where I saw a lot more going back and stopping

this one is moving along pretty smoothly, which shows that their ideas are getting across without distractions of formal problems.

There's a giant dot on "next". I think maybe at that point I was considering these transitional words. We have "therefore", "next", "for example". They're using some variety, they're logical. Sometimes I'll see students throw these words and just to have them there, to check them off the list rather than for a function. This seem to be using them with the correct function.

"The company's gonna hoard supplies" ... I think that was interesting because it wasn't in the original texts exactly. The idea of the companies who hold these patents or whatever, hoarding the food.

If they have control over the crops, then the idea that the students talk about hoarding supplies because it wasn't in the original shows me that sometimes they bring in extraneous stuff just to fill space. But this seems like they're thinking about it. So they understood what was going on in the original text.

The point of this exercise was to give their own thoughts on it.

They seem to be analyzing it and that's why I think I stopped there to think about if it was the same original text. And if not, then this is a good sign I think because they're extrapolating from the original information possibilities. It could happen.

I'm kind of examining that section a little more carefully to see if that is showing original thought or just copying something.

Now did this second paragraph, "supports the development of GM crops"? Here we're talking about "support". I guess that's a word showing that we're giving a side of an argument. I was checking that to make sure.

Earlier they said they were against it. I think that's something I would think about, these parallel structures or something to structure their argument or their description.

I stopped there with "the GM crops". There's a little issue there but I didn't spend a lot of time on it. There haven't been a lot of issues of verbs and stuff so I haven't done much stuff in there. I think that was the case there.

I kind of stared at "50 grams". I was thinking about... If I recall correctly, thinking about whether the information they're taking from the second source text

Because it seems like they're relying more heavily on one than the other. I think they could have said more judging from their other writing (source text).

But maybe they ran out of time.

What I recall doing is after I finished it all, I re-read the second paragraph and then normally I'll go back from the beginning of reread it all together. But that time I reread that second paragraph immediately after reading it once just to see what was there. And then I went back to the beginning to look through everything.

"There can be also developed further". I looked a little bit at that second sentence in the first paragraph.

I think too, now I'm looking for a complexity of the verbs or if there's any mistakes just as a last check.

By this point I had a good idea of what this essay, what level it fits.

So I'm not stopping a lot on anything. There weren't any big problems to make me stop. Just checking to make sure. After I kind of get an idea of what I think it is, I look again, make sure that I stopped here with "government so that they can look after the people". I looked at that again.

I was interested in the idea that... This is more about concepts and I think that's important for the Pass level.

And I'm thinking about the concepts they're talking about and not how they're saying it.

They're getting their ideas across pretty clearly.

I can think about, this is an interesting idea rather than look at how they made a mistake. They talk about here. "Therefore I think the GM technology should be in the hands of the government". I think that's an interesting idea politically.

At this point I think they've accomplished and they have a high enough level that we can think about their ideas rather than the structure.

I did scan the source text number one a little bit

to see if they're including the important, what they are taking out of it basically. In the case of this one, they didn't talk specifically about the effect on poor countries in Africa, which was mentioned it here.

The goal of this one (Task 2) isn't necessarily to list all the arguments. This isn't a summary. They're supposed to give their own thoughts about this issue and then supporting with the source texts. They don't need to mention everything. I didn't think there was a problem that they've left out some of the specific details because they had plenty to say and it was supported by it. They didn't need to use filler, listing with the original argument. So they're able to build on it with their own ideas.

I did spend a little time looking through, I think both of the source text, just to see which part they choose.

If they chose to take out just a random detail. Some students will just bring in one random detail cause that's all they read or understood. But if they'd take out, I think, the most important overall idea, then that shows more reading comprehension ability.

Text 2 the source text. I scanned through it, but I knew that they didn't write that much about it, so I didn't need to spend as much time looking at it.

Rater 4 – Essay 5

So, this first sentence I think I'd go back over it a few times because it's long.

It has a few things in there that are confusing, just to understand what the writer was saying took me some time, then to figure out what it was that was making it hard to understand.

The idea of using a longer, more complex sentence can sometimes show a higher level, but it might just be a mess. They need to be able to use it skillfully if they're writing longer sentences.

Let's see. "No one would be strange about".

I mean it's an error, but it's kind of - I don't know what kind of error it really is ... the vocabulary it's a word choice error. But it just sounds unnatural.

"have some bad influence in the future."

As I was looking at this it reminded me of students I've worked with in the past, and this might be irrelevant, but it reminded me of the kind of errors that my students that spoke Chinese would make, and I thought I wonder if it's students from China? So sometimes I'll see patterns of errors that are very characteristic of different L1s, so I thought about that a little bit right there.

"Benefit to people for those one who cannot afford for the food," third line, first paragraph.

it's kind of a weird construction there, so I looked at that more carefully.

"Others hope that the GM technology would do benefit to people for those one who cannot afford ... "

I can see what they're trying to say, but I think it's getting toward more complex structure, but it's not doing it right.

Okay so the first part there (paragraph 2), they are making reference to one of the source texts - the first part of the second paragraph.

And that is one of the criteria: Are they incorporating the original texts in some way? So, I paid attention to that.

There are a couple things, small errors that I kind of skip over. I didn't spend a lot of time on them because they're less important, at least at this point.

Like when they say "in result of" instead of "as a result of". Prepositions cause a lot of problems, even for high level learners, so I didn't spend time on that, but I did notice it.

And I think that could come into play later on, depending if I need to make a finer grade decision, but at this stage, not really something I'm doing ...

"people who cannot afford food before could eat more and have more chance to survive,"

I think I just looked at "eat more" because are we talking about eating more? In here I don't know, but not super important I guess.

I looked at "nowadays", so that's kind of highlighted because that is a very common word I saw in my learners that I used to work with before coming here.

"So nowadays ... " It's also used at the beginning of this essay. To me that shows, it's not necessarily low level or high level, but it's if they're using it more than once. I mean it's not a common word for L1 writers at all. "Nowadays" is rare, so it shows maybe intermediate level. It is some kind of, I don't know, it can be used as a transition effectively, but I don't think it's a high level one.

I stopped there, it says "ten percentage of global human are suffered by ... "

That was kind of interesting to say they are "suffered by" something, that as a passive, I guess. Suffering from hunger ... but then maybe it's prepositional thing ... but I did look at it a little more carefully.

And then we stopped there by "besides" that's I think a signal word, and that's a good ...

Although this essay has some problems, I do think it has good structure to like logical structure, and there's words like that are showing that ... that's what made it kind of tricky. There're a lot of grammar or other issues like this in text, but I think the ideas are structured pretty well, so I thought about that a lot.

I lingered on that last sentence of the second paragraph because everything else is mostly longer sentences with more complex sentences, but then they also have this little sentence at the end of the second paragraph. - "It's also benefit to hunger people ..." Not only is it wrong, but it's this little, very simple sentence, so I kept that in mind.

"If we take our eyes further the worry in the first article isn't fake ... " (paragraph 3)

So that makes sense, I can understand what they're saying, but it's very non-idiomatic. "We take our eyes further" is not something people really would say as a native speaker. And we wouldn't say "it's fake". So it's not ... I guess I was thinking about how bad is this if it's just kind of choice of words rather than something being wrong, so that's something to consider too I think.

There's this, "not only, but also" structure that I thought was good. It shows a higher level

So this last sentence, looks like I was reading it a little more slowly

because it is structured pretty well.

"Because the technology is high tech nowadays which means most ... "

I see a subordination there.

" ... which means most common companies can use it finally most food."

There's a complex sentence there, and structurally it's mostly okay.

I was debating a lot because there are a lot of errors in there, but how much should I weigh those kind of errors versus the ideas being put in and organization.

So soon I went back to the rubric to try to get some help on that because how much should I weight grammar.

There are a lot of grammar mistakes, or word choice mistakes, but good organization, good ideas - then where does that put it?

I went back to the essay to just review, especially the first paragraph. And that's that first sentence again.

So I think when I'm thinking about how this complexity of sentences is versus how well its executed. The first and the last sentences are both good examples of that.

Especially that first one, I went back again and read it a couple more times because it's hard to read, but how wrong is it. I guess is what I was thinking about. So that's probably why I went back and spent time on that one ...

In the second paragraph, there's nothing that's tripping me up too much. We're seeing a relative clause in here. They're using it correctly. So that's saying "people who cannot afford food before could eat", so that's a complex way of structuring it there. So they've got "food" called a relative clause ... it's evidence of higher level maybe ...

Just reviewing the rest of that second paragraph ... looked to that "suffered by hunger" again, but otherwise just looking more slowly. Looks like I'm stopping every few words. Looking more carefully now instead of just ... now I'm looking for those details of form, rather than just the concepts, I think.

Back to the third paragraph. The first sentence "takes our eyes further, the worry of the first article isn't fake" is getting a lot of attention again.

I think because it feels unnatural.

I went back to look at the grading rubric again. I looked at the intermediate, the CD level a lot because I thought maybe this is where it should be, so I went in detail to look at that, but well just see, cause I was unsure, this was kind of a hard one to grade.

Started from the top, I looked at everything I think in that CD level to see if it's a good fit. Pretty carefully, I mean I've read this a lot but I really spent some time on this time. And then I went to the lower level, the B level, and I particularly looked at grammar and lexis.

Because the argument, organization those were pretty good in this one.

So I thought how much should I penalize them for having grammar mistakes, so that was a tough decision to tell whether it should be in the B level or the CD level.

At first when I look at the B level, I looked at grammar and lexis a lot because that's their main issues, and then I looked at organization and argument and details because those were their strong points I think. Trying to compare those two and get an idea... if the grammar and lexis is in the B level, but then if the others are in the CD level, then how should I balance that, so that's what I was trying to do there.

I didn't really look at convention, I feel like convention's pretty straight forward I don't spend a lot of time on that. They did have issues with that.

But I don't weigh it this heavily, and I mean it's not weighed as heavily on the rubric anyway ... only 15%.

Back to arguments and details, now I look again at the intermediate level because still unsure, not 100% sure.

Okay, now organization again. Now I'm trying to decide if organization is at the C D and argument and details are at the C D level and only grammar and lexis is B level, does that mean that it's a C D or C D-, B , or B+? Trying to find that grade zone there.

I did spend a little time looking at the essay again, not any particular parts, just seems like I'm just sort of getting an overall view of their argument and not of the errors as much, maybe because I've been talking about organization.

The second time though I do remember now that I was trying to look at that structure and how they are using signal words and things like that and relative clauses.

At that time I think I had decided that it should be B+... yeah ... but as you can see it's not 100% confident.

Rater 4 – Essay 6

Okay, so first paragraph, I lingered on that first sentence because it seems very simple. It's more "products such as the rice plant vegetable, even meats."

So there are issues with articles and stuff but it just starts off very simple.

So I think, I think that gives me the idea, is this a low level, is this the B.

And then I will look for more evidence going forward to confirm that or not.

Now in second paragraph, I guess I went fast through the first paragraph.

Oh okay so this part in the second paragraph "although it do not change the foods look like and give them a new property." So I looked at that more carefully because "although", the word "although" suggests that it might be. You know, it's a good word to use, it's structuring the argument.

But, there's a verb error with the "do".

"Change the foods look like", so that's, that should be appearance or something like that so that's kind of showing lower level vocabulary.

"Give them a new property," so that's confirming the lower level that I was thinking for.

I revisited that first part, "to prove this species for production". [crosstalk 00:01:20]

It stopped on "in the text one". So they are making reference to one of the texts so that's a good thing I paid attention to that, noted that. So they're citing the source. And I read that part twice because they actually quoted it directly, which isn't that common in these. They're usually paraphrasing.

I don't remember if paraphrasing is required or if they're allowed to use direct quotes.

But I thought that was significant cos it shows the difference in their own writing from the original.

It's because that section that's in quotes is a lot more high-level. Okay, like they don't know words like "undercutting", most likely, or "subsidize".

Okay, read that second paragraph again, the second sentence. I don't know why I looked at that, I might've just been sort of scanning that whole paragraph a little bit.

Okay so here, I think. So they're talking about, it says "So and the African people will gain less money."

So, I knew "gain less" seems weird. You know, instead of saying "lose money" or something, they're saying "gain less money".

I looked back at how it was worded in the original quote, so, I think they did understand somewhat. They understood what was going on there in the quote.

Um, talking about going out of business, so, they are synthesizing information somewhat, but it's in a very simple way.

"Second", so we have one of the simple structures in the third paragraph to start off. You know, they do the "first", "second".

So it is organized, but in a very basic way.

Also, here we refer to text 2, so I stopped at that. So, they're citing, kind of like citing their sources. They're referring directly to the 2 source texts.

Oh, I looked at that "it will cause the blind or dead" so that part; I kind of chuckled at that, maybe I shouldn't be chuckling, but...

So, "if humans too much Vitamin A, it will cause..." So that seemed like very unsubtle understanding of the original. They're not saying that it will. So what this writer is saying is, Golden Rice having more vitamin A is bad and that it will kill or blind people, which wasn't the case, it was people with a lack of vitamin A becoming blind or dying, so they misunderstood that.

"Other species will have negative..." I looked at the word "species" for some reason.

Okay, going down to the last paragraph. So, I read it, went back to the beginning; so "according to my two reasons".

They're just kind of summarizing, they're talking about the two main points.

So that's, yeah, structured but in a very, very simple way.

I kinda had my idea of what it was. It seemed like a little like a B.

I went back to the beginning. Yeah, think I'm just reading a little more slowly now, carefully to see if I've missed anything relevant or important.

The second paragraph, I'm just kind of scanning through it.

And then the third paragraph. Down to the third paragraph, moving quickly through it now.

So it seems like I have made up my mind by that point.

And I go back to the rubric to confirm what I'm thinking.

So, you see I started on the B-Level, because that's what I was leaning toward, giving this a B. So I wanna see organization wise - is it fitting that description.

Argument, details, moving down the B description.

Yeah, moving through each of the categories, down to the grammar and lexis. So that was, I spent some time on that

because that was where a lot of the low level feeling came from was either simplistic wording or incorrect wording.

Didn't really spend much time on conventions. I spent a little bit more time than before.

But now, so I did look at intermediate level too, to see if it should fit in there, you know. Cos I'm leaning toward B, but I took a relatively quick look at - not too quick, I looked carefully at the CD section 2. I was having some doubt maybe, but I'm just seeing if it might fit into that category.

So, I was trying to confirm my hunch. After looking at the rubric, I knew what level I wanted to give.

Okay so now I went back to take another quick look at it (the essay). I clicked back onto it but I don't recall looking for anything in particular. So, kind of just glancing through it, I guess.

I had stopped on this part, it says "make a bad influence." So my question I had for myself was "What's their main argument?"

So, I think what I was looking at here. They say, "the genetically modified food will make a bad influence for the human body".

So, it's grammatically wrong.

But I think I was more thinking about "are they talking about the positive side too?" So then I looked back a little bit at the beginning of that third paragraph to see what they said about the positive sides of it. So, are they covering both in a balanced way.

But then, I noticed that they do talk about how Golden Rice can have vitamin A, they misunderstood there thinking that that would hurt people.

I think now, it was pretty well decided.

Came back to the end of the second paragraph, I think I was looking for evidence of those negatives, because it seems overall, their argument was that it's bad, so I was looking for their evidence.

And that end of the second paragraph was talking about how it's hurting people financially.

I looked back again at the rubric, very briefly. Arguments and details, grammar and lexis - both, but not organization and convention.

I guess just that last question I had for myself, was if I was choosing the right grade.

Rater 4 – Essay 7

I looked at, the f- the first sentence. I just read through it for ideas.

But I did stop at “discussing about”, which I don't think is a big problem, but it does catch my eye, because it's not the right preposition.

"[reading from essay] supported but for other ...". so they do lay it out pretty clearly in this first paragraph the two sides of the argument. So that's a good point here.

And then they, so they show the two sides and then they show what their thought is. So I think it's pretty good. The thoughts are organized.

You know, it's not being said in a complex way. It's got the necessary points.

IOkay.

"News reporter has reported ... " Oh, so they say, "The news reporter has reported." (the second paragraph).

So that's kind of a, I think I looked at that because I was like, "Who is this?" They're not specifying which text. Yeah, they are saying the news reporter. So it's not very clear where that's coming from.

"Rice has become a serious problem that people are facing." Yeah, so that sentence there, second line, second paragraph, "So rice has become, is a serious problem that people are facing."

So that part is confusing,

and I think they mean to say maybe, "Rice lacking vitamin."

So idea-wise it's problematic and structure wise.

So I think that helps me make decision. I think it would be a piece of evidence.

IOkay.

Yeah. So I stopped and looked at those, well they're using these parenthesis, I think that's not like incorrect or anything but sometimes people, writers, will use, add information using parenthesis when they don't know how to integrate the words into a sentence in a different way using like a subordinate clause or something like that. So I did pay attention to them doing that.

Okay so I had stopped at the end there, "what they need." So I think, well it's not just “they”, “what they need” is kind of vague, you know, that's I think why I thought about, like it's not everything they need, it's just the Vitamin A.

But I might have just been pausing to think about the overall paragraph.

IMm-hmm (affirmative)

So the third paragraph, kind of scanning through for ideas. I did, I went back to read the second line more carefully because they say "the government should change their policy and give the biotech companies a chance to..."

So the point of the original text was not that the government's not allowing these companies to do this, so they (the student) kind of misunderstood that.

And then they talked about if, like there's not enough money to do it is an idea they suggested and the government could just have, take more taxes maybe. Ask for people to collect money.

So I think this is just showing that they're not really writing at a higher level here idea-wise.

There are some grammar problems and stuff.

But, they're also not getting into the idea as deeply as maybe they could.

At the end of that third paragraph, "The high rate of vitamin A deficiency doubles." I don't know, I don't think that's a big problem. I guess I just stopped, maybe I tend to slow down at the end of paragraphs.

IAre you moving on to the last paragraph?

Final paragraph yeah. Reading through kind of quickly at this point. And now I went back to the beginning. okay, I guess I must have finished that last paragraph and not noticed anything that was big problem or anything that was especially great. So, going back to review from the beginning. Let's see if I find anything more detailed.

So I think that first paragraph looked pretty good. I didn't spend a lot of time on it because it's clear. They're no big problems. Good structure.

Scanning the second paragraph, and looking again at that part about how rice has become a serious problem.

Because that part had two errors about the content and the way it's written with verbs.

Okay, so, kind of scanning through further. Yeah, looking at that whole section there.

So I mean, well they're doing something good. They might have made that mistake but, they're showing how there's a problem, but here's the solution. So the argument structure wise, that's good. So they're saying, presenting this problem, but saying how the genetically modified food can help. So maybe I was thinking about that part of it. How they are, you know...

I don't want to let the errors get in the way of the parts that are working well.

I looked a lot at paragraph two. Not sure what I was doing. I was kind of going all over the place. And then it looks like I move on to paragraph three.

I think at this point I'm checking more content wise because there aren't any big errors in this section, in paragraph three. But they are suggesting that the problem is resulting from the biotech companies not having enough power, but actually the issue that people are talking about is that they would have, the original text said, they might have too much power.

So, "It may sounds a little bit difficult." So there's this sentence in the third paragraph. "It may sounds a little bit difficult to them, but to care about the people's life I think this plan is needed for them."

That part has some issues, but it is, it does show a structure.

I think it's good the way they're comparing different sides of this.

IMm-hmm (affirmative)

So I think I was looking at that to try and think of what, you know, how to balance those like mistakes and then the stronger parts, you know, when they're doing something well, but maybe there's a verb error in that sentence. So, think about how to balance that.

IOkay. Mm-hmm (affirmative)

Oh, okay, so I looked again at this part in the last paragraph, "Healthy is the most important thing for a family."

So that's not a huge problem but, it's just, "Health" would be the better choice there. I looked back at that.

Oh, okay. So at that point, I don't think I did look at the source text. I think I've read them so many times now that I feel pretty comfortable with it. Yeah so, I think it's a CD.

Rater 4 – Essay 8

Normally at the beginning I start to just read for content and kind of scan through it more quickly. This one, I did slow down on that and return to the beginning of that first sentence.

Because they're kind of bringing in information that wasn't from the original text. They're saying that GM foods take out vitamins, fibers, carbohydrates and many other elements. But I don't know that that's even true. It's definitely not part of these texts.

Okay, so I did pause there because he or she makes reference to one of the texts specifically, so that's there, that's important. They say something about text one as ... "According to text one". Yeah. Just checking that they're making reference to the real text.

There aren't really problems grammatically that I'm noticing.

So I think at this point of reading "reduces the probabilities of having independent citizens", I think I may have slowed down on that just because we're kind of getting away from the original ideas in the text, but that's not necessarily a problem. It's just noticeable.

All right. Second paragraph, kind of just moving through it fairly quickly. I think I reread that for content, just to see kind of what they're trying to say there.

they're bringing something personal here, which is not necessarily that common. I don't know...that's something I thought about, is this a problem with it or not?

They talk about kind of government corruption basically. So, I think I was trying to think back to if that's relevant to the original text.

"Certainly control". It should be certainly controls. So that's just a little blended morpheme thing or something. Yeah, should be an "s" on the end of "control"

Okay, "money is power and certainly control people's act if they cannot have it".

So "that being said", I think I went back to that because I thought that was a pretty good structure, you know, the connecting to a previous idea but also showing contrast, so it's a good structural thing there.

"It's important to mention" in paragraph 3 I think, is also showing some maturity, the writing there.

"member of a church" I also looked at that more carefully because it's bringing more outside personal information, but the question is whether that's contributing or not and if it's making it a better argument.

Moving pretty quickly through there, "having our own garden increase probabilities to improve our health because we know what we're consuming... "

So this is a pretty complex sentence, and they pull it off. It doesn't have any real problems, I think. "Increases the probabilities to improve our health," it's not super idiomatic, maybe not the way a native speaker would say it, but it's not wrong. I think just reading that part carefully to make sure they did do it right.

Moving to the last paragraph, they say "based on the Text" with capital "T". When we see that, you think of a religious book. And I thought "oh, is this related to, they're talking about the church before, so maybe that gave me the idea of religion". The fact that they mention the church made me see capital "T" in "Text" and think of the Bible. So I don't think that's what they meant.

Then I think it looks like I paused there on that part, probably for that reason because I thought...what kind of text are they talking about. But then I realize it's just the reference text that we're writing about here.

So the sentence, the second line of paragraph four, it's a pretty good sentence actually. We have structural complexity. So we have two layers of complexity, kind of. "On the other hand", "it is not only", so we have this "not only but also" structure, and also this "on the other hand" relating it to the previous...

So they really do a good job, I think, of connecting the different points they're making in a logical way.

I looked at the slash (/) thing. So they do say "economy/world", and I encourage my own students to not do that because it's kind of a way of avoiding using...it's easier than writing out the actual words. So I did glance at that.

"As I mentioned before, cheap"...So they did mention twice this cheap does not necessarily mean good or excellent thing. They say that twice. But as I read that I thought well, nobody thinks cheap means quality. I don't know if it's that good of a point because nobody thinks cheap means excellent. I don't know if it's a great point to make, and they do refer to it twice. However, I don't know if that's relevant to what we're doing here grading these. I just thought "well, it's not the best way of expressing your point". I think I know what they're trying to say, but yeah. By saying cheap doesn't mean excellent, that's obvious.

They did refer to text two, specifically, so that's good. I might have seen a quick glance up. I think that might've been a line there, looking at text one, so seeing that they referred to both texts specifically.

Okay, so I think by that point I could see that this is pretty well put together. I had my idea of what grade it was by that point, yeah.

I think I looked back at the rubric if...yeah, okay. So I did look back, and as you can see, I looked at the pass section primarily because I felt that this is a pass. I wanted to check my intuition through all the sections of the pass part of the rubric.

Okay, so I came back to scan through, I think given the information I just got from the rubric, to confirm the choice I made. I was thinking about as I came back, some essays there will be a big problem that I really have to consider, but this one I was pretty confident in my choice.

Rater 4 – Essay 9

Okay, great. I looked at "nowadays" there because everybody says "nowadays." My students always say "nowadays." It usually doesn't indicate a high level. But it doesn't mean it's bad, either.

"Agriculture industry is not then an exception".

I don't know why I stopped at "exception"? Maybe that's a good word and maybe that's why I stopped. Yeah. Like it's higher level than predicted.

Okay so, "manipulating the genes" is not the way that I would first have thought of saying it, so maybe I looked at it, but I think it's good, so...

"There are both cons and pros". That's interesting because everybody always says, "pros and cons."

Okay, so, "however, I believe the negativity overweight"

The "however" I think is a signal that they're comparing things. I went back to the beginning to see if they actually using "however" correctly to show contrast. I looked back to see that they say that it was improved, so then it would make sense to say, "however, I believe the negativity outweighs it."

So they're saying, "I believe the negativity overweight," which is the wrong word (overweight).

Second paragraph "in the first place, farms which are genetically modified to grow crops on draw more attention."

So that I think it's okay, but at first glance, the "on" being at the end. You're not really supposed to put a preposition at the end of a clause. So that's why I think it was confusing. So "farms on which", that would be clearer.

I think I'm not really picky about the preposition at the end of a sentence thing. It's not a real rule. But I think if it was at the beginning, it would be a lot easier to understand.

So that is like a complex construction, "farms which are genetically modified grow crops on draw more attention", so they are using a complex sentence there.

"Biotech companies become more interested in ..." So this is I think, yeah, going back to the first sentence of paragraph 2 to try to understand that part. So I went kind of quickly through.

Okay, last sentence of paragraph two. Finishing paragraph two, I'm glancing back around and maybe I'm just trying to connect to what they'd said before "making poor nations poorer as it is less expensive".

So the "American process", I think I looked at that because it's not really about America. I mean, it is mostly American companies, but they frame it that way, which is not like a mistake or something. It's not relevant to the grading, but I thought it was interesting.

And then, okay, down to the third paragraph. "Billions of people rely on rice as their main course food".

So that's, I think a pretty good sentence.

So I saw "on the other hand" at the beginning of the third paragraph, and I felt like I had seen that before, so I glanced back up just to make sure they're not using the same phrase. And they don't use it earlier.

"On the other hand" is overused sometimes with students, so that's what I was thinking about.

I did stop there because they don't say "the quality of the food people eat", but they're saying "the quality of the food people take", which is kind of not the best choice there.

So I they're a little confused here. So, they say "instead of spending money on modifying crop genes, they should be used to improve nutrition", but that's what they're modifying the genes for. At least in the case of golden rice. Some of them it's for like insect resistant or something, so that's what I was kind of trying to figure out "are they misunderstanding something or not?"

So, back to the beginning (intro). "The agriculture industry is not then an exception".

So that part was worded a little awkwardly, so I looked at that more carefully. "Agriculture industry is not then an exception".

But I think it's actually, with a comma in there, it would be fine.

It should add "the" in "agriculture industry", but other than that, I think it's okay.

With punctuation it would make it more clear.

But they're using a higher level of structure, they're not mastering it 100%, but I think it's pretty good.

Oh, I stopped on "taking this ability". "There are both cons and pros of taking this ability of modifying genes of crops into work".

So, "taking it... into work", that's strange, that doesn't make so much sense. So, I thought about that a little bit, how important is that mistake or choice.

Okay, I looked back again at the second paragraph. Kind of scanning through a little more slowly and carefully but ... "This helps the industry to be developed faster with more modern ways which can" ...

This part, I remember now. I thought this sentence shows some complexity, which is important to think about. "This helps this industry to be developed faster with more modern". So you got a prepositional phrase there and then which, so we're using a relative clause I think. "Can help reach, help reach that should be, more yields at the end". So I was checking this structure that if it's correct, and then thinking okay, that's pretty good. I think at this point in the end that's what I was doing. I was looking for that syntactical complexity.

"Submerge", oh yeah. The word choice of submerge was kind of strange, but I understood. They mean that the business will not do well.

Okay, looking at the last paragraph again. Oh, I think at this point I was looking at things like "which". Like how are they using relative clauses and stuff.

So, "rely on rice as the main course food, with 10% risk for Vitamin A deficiency" ...

So, that structure I think is good.

They're talking about these people rely on the food, and these people also have this risk, and then they modify the deficiency.

That noun too, so, it's pretty good.

Okay, kind of just reviewing the last part. "More budgets should be allocated to improving" ...

I think I did linger on "allocated" a little bit. That was a good word choice. Budget allocation.

So that might be, I maybe should have looked at the original text to see if they got the word, but even if they knew the word well enough to include it and integrate it into their own writing, I think that's a good sign.

The first line of the second, okay. If I remember right, I was looking for that sentence complexity. At that point, going back and looking for examples of that.

So, "farms which are genetically modified to grow crops draw more attention in order to apply"

So there is good, not only that syntax

but also these logical relationships, like "in order to".

And then, I think at this point, moving kind of quickly through 'cause I'm looking at how the sentences relate to each other as I was saying in the first place. And then this helps the industry, so they're building on what they had there. So this is logical following from that.

And then the "however", which we have in that last sentence of the second paragraph.

So, I feel like they're using their evidence in a clear way, which is good.

They're connecting the ideas logically.

Oh, okay I did look back at the rubric.

So, my initial thought is that is the CD - it's pretty good.

It's good, but is that where it should be and then I looked into the pass level because I would say a good CD, so trying to decide if it should be a pass or not.

Yeah. I looked at organization and the argument and details.

I think were strong suits on this one, because they do organize their ideas really well. It's logical, they're making good points. So those were thoughts in my mind for sure.

Okay, so looking at the pass level now there's those details, a little bit at grammar and lexis ...

But, okay so they did take into consideration the intermediate level grammar and lexis. I think I looked at the low level just for reference. I don't think this is going to be a low level paper, but it's good to get a comparison.

I think it's a CD+

Rater 4 – Essay 10

Okay. So I'm kinda reading through quickly.

But I think I lingered a little bit on, "in recent days". "Getting benefits out of it". I looked at that. I don't think it's really a problem but, I did look at it because it's not standard. "Finding deep trouble out of it". Also, it's understandable, it's not interfering with comprehension, but it's kind of strange, so I paid attention to those things.

And then, looked back. So this I think here, what I'm looking at is, this is a long sentence, the second sentence of the first paragraph.

So they have the "neither nor" structure, so I went back after I saw the "nor" and read that part, went back to check the "neither" just to make sure that that's done correctly. Because if so, if

they do it right, then that's a good sign of their level. If they're using, if they can write a sentence that's that long and have control over their grammar. So, are they able to use these structures correctly? And there are some issues, but overall it's nothing major.

So, second paragraph, "GM Technology are heavily controlled." So, that was a mistake that I - So, you think about the count, non count nouns and that's what I was looking at there.

"Rich nations, [inaudible 00:01:30] benefits of such technology. Getting benefits of such", yeah, I don't know, something to look at briefly. Maybe "get benefits from" would be more common.

So I think I looked at it a little longer, just because it's not the most frequent way of saying it, but I don't think it's a problem.

"Is it providing it to their farmers" - I think there, it's not, I think I looked at that longer now because it had a grammatical problem or something.

But, because, I think there, this shows that they're analyzing and not only quoting from the original text, but paraphrasing ideas and then reading into them. Because talking about giving subsidies to farmers is, it wasn't directly said that way, I think the fact that they mention it that way is good.

"American producers are arguing, oh giving a tough competitions". So, that is just structured weirdly.

So they're saying that they are giving, "making it difficult for other people to compete". So, it's a hard thing to word. As I said it I kind of had a strange phrasing. So, I was looking at the way they chose to express that.

Okay. Third paragraph. "Much cheaper cost". In general I'm reading through the third paragraph pretty quickly.

There aren't any major issues.

I did look at, I think the "Project Syndicate". "A study of Lomberg found that." So, that kind of confused me at first. Is the Project Syndicate, is that the name of a study? I was looking to clarify what they were talking about.

But, they did cite their source there so that was, I think, even if it was worded a little awkwardly it shows that they're integrating their sources and paraphrasing rather than just pulling out quotes.

Kind of moving quickly through here - "not sustainable solution". Pause there a little bit. "Not sustainable solution".

I think that last sentence also, I pause on it because they're comparing two sides of it within that same study so that's - some of the students didn't catch that. That there were, on one of the original texts it showed two interpretations of the helpfulness of the golden rice. So, they did a good job with that.

So, they're showing, not only the two main sides of the argument, but they're going into a little more detail.

Last paragraph. "Very good, has such technology but at the same time [inaudible 00:04:02]". So, the last paragraph I went through that pretty quickly.

I think because there weren't, at first glance, there weren't any big problems or anything. I mean, there are a couple little things but nothing major. And it seemed to be a good summation of their point of view on it.

That's good. They managed - some of them don't have enough time to finish and get around to their point of view on it so that was good that they were able to present both sides and give their take on it.

Alright, back to the beginning. Review. So, kind of quickly going through that first paragraph to
—

I think I'm still not 100% sure the grade I wanted to give at that point. Going back to see if anything might change my mind.

Second paragraph again, [Reading inaudible 00:05:03] Glancing through it quickly now, I guess I'm not - I'm really glancing a lot through that second paragraph.

I think they made some good points in that second paragraph in how they're - I think they did a good job digesting the original argument, so I think they understood the original text well and I think they're explaining it well.

and they did a good job re-wording it in their own way and they mention this traditional and costly way. So, that was - it's different than the original way it was worded.

Maybe I was using that paragraph as evidence of a higher level.

Now in paragraph three, I return to that so I could see if they're balancing both sides of the argument. Paragraph two talks about the, kind of, economic side of things and then paragraph three talks about the nutritional side of the argument. So, seeing if they're talking about both equally.

Last paragraph, not really spending a lot of time on that.

It's pretty simple and no big problems on it.

Okay. Now, I first looked back at the rubric at the pass section.

This seems like a pretty good paper.

Started with organization. Moving from the top down. Reviewing the characteristics of a pass paper.

Something I was, remember thinking about, was I think I've lingered on the grammar and lexis because how many errors would still- It says in the description 'some minor grammatical and lexical errors. So, is that what they have? Or do they have more?

And I will look also at the C D grammar and lexis because it specifies that they might occasionally interfere with comprehensibility. So, that's the key thing.

I know there are errors here. But, are they actually interfering with comprehensibility is my question to myself.

I looked at convention, briefly. I did look at it a little bit because of how they integrated the sources. So, are they paraphrasing? I think I was wondering - are they supposed to refer explicitly to the original text or just paraphrase their information? I think they did a good job

paraphrasing. They understood and digested the information and were able to say it in their own words. I think I was checking if they needed to specifically mention who said these things.

So, I did kind of compare those two levels of convention, too.

Yeah. Grammar and lexis. Comparing those two levels. The C D and pass again to kind of get an idea of what, how many errors can they make to fit into the pass? “Wide range of structures” versus “good control of simple and some complex”.

So, trying to find where they sit in that spectrum.

Okay. What am I doing there? Oh, argument details. I think at this point I'm just reviewing things 'cause that was not as much of a concern.

Yeah. Okay. Back to the essay to kind of do a quick review.

I don't remember exactly what I might have been looking for, but I think again that balance of complexity and comprehensibility. Are they using simple grammar? Are they using complex grammar?

How many mistakes are they making? Does it get in the way of understanding?

Skimming again through the second paragraph because it's a little bit of a tricky one. It's high level, I think, but is it pass or is it? Now that we have these plus options. Is it a pass? Or is it a C D plus is my question.

I went back there (the rating rubric), but pretty briefly. Looking at that C D level grammar and lexis.

I know they did make some mistakes, but how significant were they is the issue?

Convention. I mean, I guess I looked a little bit at that. But, probably just about whether they needed a direct quote or I mean a - specifying essay one or essay two. They did. I don't think that was a big concern.

RATER 5 (09/06/2018)

Rater 5 – Essay 1

So I was reading a text here trying to understand what this person says.

And in here in the introduction I thought this person's writing is very generic.

So that's why I asked whether or not this is task 1 or task 2.

And then this part is really introductory and something introductory comes again here, so there's kind of weak moves. So that's what I thought. "genetically modified food gets a lot of different views", so this is very general thing. It's good as a first sentence (second paragraph) but in this sentence is also kind of general. This could have been in the first paragraph, and then the first sentence of the second paragraph could have been more specific to introduce how genetically modified food benefit people.

Right (reading first paragraph again)

so I think that's what I thought. Because this is general, I was wondering, “well this could have been in the first paragraph” and then I think I tried to remember or tried to refresh my memory

about the first paragraph because I thought this would have been before in the previous paragraph.

That's what I thought.

This paragraph and going back second paragraph. I think I'm just reading to understand the text.

And then here, I thought that the paragraph (the second paragraph) cannot conclude with an example. Yeah. So, this person keep talking about the bananas and how GM food, like the relationship between GM food and bananas. And then there should be some concluding remark to conclude this paragraph, but this person ends with the example; so that lowered the grade of this person. Maybe I paid attention to "bananas" because I was so attracted by this example. You know, example can't finish the paragraph. I think I keep thinking about that.

And then at this point of time (second paragraph) I thought this person should be B, even without reading the rest.

And then I think I spent less time on the last paragraph.

I'm not sure but I guess I kept thinking about this sentence. The example. Yeah this attracted my attention pretty much.

Yeah. I think that was a really, clearly B.

Not really because I was familiar with this.

I think I was looking for the rubric.

So first thing that I thought of is the argument and then the argument were not clear. So arguments are vague.

And then this is a really crucial for task 2.

I checked the difference between B and C well as.

I knew it's B.

I was thinking about that accuracy of language. Grammar accuracy.

So yeah I was thinking how this person's performance is in line with criterion and then checking, comparing with the C and D level of criterion for grammar.

Yes. I looked at the essay again.

Maybe this one I wanted to make sure that the accordance of the introduction part. So I was checking whether this was really awkward. So as I said there are kind of too many introductory sentences in this paragraph. So maybe I was checking that.

And then yeah I finalized my grade.

Rater 5 – Essay 2

I'm reading the text, the introduction.

I thought it does good as an introduction

and then I... these two ... these two phrases caught my attention. "however", "in my opinion".

And then I thought, “oh, this is good introduction, it shows this person's claim instead of the summary”. So, yeah. That's what I thought. I think that it's good because it shows her or his perspective. Opinion.

Yeah. And then here, “there are chances to will get better overtime” (paragraph 1). I was just wondering if this is grammatically correct?

In “for example”, yeah, in here I thought even though this person tries to express his or her ideas in here, I thought that this is from the source text. Those numbers. So, I thought even though this is task 2 and then even though this person uses “in my opinion” and “I think”, what he is writing in here is about summary of the source text.

And moving to the last paragraph. And then here, I think this shows my eye-tracking in here, I guess because this person use this phrase and then twice. This phrase, I mean, it's exactly the same “there are chance that”. So, I thought that would be repetitive. There should be an alternative way to express the same thing in different ways. So, that's what I thought.

So, that's why I went back to this part (the introduction)

I think “so, this person hasn't finished the text yet”. That's what I thought.

And then I think this always happen to me, but I usually look at the arguments and details first. Spend some time over there.

Yeah, I thought this text would be B

but I'm just checking the criteria for B.

Yeah. And then I'm rereading the text. I think I'm at this point in time, I'm focused on the arguments. However, right this person's argument is.

I think in the second paragraph, they told all of a summary but in the ... well, then in the third paragraph, it also says both are from a text that I mention that blah, blah, blah. So, it's part of summary but at the same time, however, there blah, blah, blah, is maybe this person's opinion. So, I was just evaluating this like summary and his opinion.

And moving back to the rubric, I spend some time on the argument part, I believe. So, “vague”. And it seems I spent some time on grammar and checking to see the criteria for grammar.

Yeah. And then moving back to the text. To the essay? Yeah, I read the last paragraph again.

This one is ungrammatical.

And then moving back to the rubric. And then checking C and D criteria.

At this point of time, I thought this is B.

But probably just in case I'm just reading the C and D criteria. Argument again. At this point of time, I focus on that grammar.

So, the last paragraph or the last sentence for the second paragraph, "Which means on the 500,000 children of those blind", I thought it was ungrammatical.

Since I have knowledge of a second language acquisition, usually I do not count like preposition or article errors because they are really hard for second language writers. But this one (relative clause), I think this is easy to acquire.

So, I didn't care about like this error, "cause children". This is something we usually make. I'm tolerant with those errors.

I recorded the grade at a time. Yeah, I was determined.

Rater 5 – Essay 3

So, first thing that I thought is "this is long". Actually, this is first thing that I think of.

So, it could be C or, yeah, that's what I thought.

And then the previous two texts are really short.

And then that really affected my rating. If it's too short, that would ... Not always but mostly B. That's really my rule of thumb.

I think this person is good, but there are a lot of grammatical mistakes, which say "additional nutrient rich being modified to benefit to prevent damage by insect or bring", you know.

I found this is really hard to follow. Yeah, the first double relative clauses are really quite difficult to follow. That's what I thought.

Yeah, and then I, also, paid attention this "I don't think that" ... So, this is something to express this person's opinion. So I thought "Oh, this is good".

Yeah and then this grammatical mistakes in the second paragraph. It caught my attention too. "Crops is being produced be chaging". I classify it as a grammatical error.

I think because of this, this prompted me to look at the grammatical mistakes of this person because this is something easy I think people can acquire. So, that's what I thought.

And then this also again caught my attention. "There is an experiment that has been conducted that GM has been affected".

I'm not sure about this writer, but this is typical errors among Chinese writers. So, that's what I thought here. Yeah.

And then again, grammatical errors. I think I focus on grammatical errors for this person. "This can make them to not easily being damaged" (third paragraph)

Yeah. Yeah. And then some spelling errors. "lea", "evironmental", "chaging", "ectixnt to those animlas" (third paragraph). "Side effecta". Maybe spelling errors.

Then I moved to the fourth paragraph here.

"strongly herbicides [t]o". More grammatical error then.

I thought the fourth paragraph was good in terms of the meaning conveyed by this paragraph. So meaning expressed in this paragraph.

Yeah. And then I, also, looked at this transition, uncommon transition. That's "further information" and "conclusion". So, this doesn't follow convention of writing. So, we don't say "conclusion" or we do not say "further information" either to move to the next paragraph.

And yeah, I thought that "you have freedom to make decision your health based on your hand and for your future generations". So in this sentence sounds good but I think there should be more sentences before this to substantiate this nice claim.

I think I reviewed the argument point and grammatical errors.

Also, I think I this time I focused more on grammar. Yeah. So, I especially paid attention to the last criteria for grammar in lexis where it says "This set contains many grammatical and technical errors which interfere with comprehensibility", which is the case of this text. And then I, also, noticed C and D criteria might occasional interfere with comprehensibility.

I just wondered between these, and then I decided to go for this because of the many grammatical errors.

You really couldn't understand this person's claim. Yeah. So, that's what I thought.

Another thing that I thought is that for grammar, this person was B.

But I thought that their arguments, I think in terms of this criteria, this person was good enough. C, D. Yeah. Definitely not pass. And yeah, "mostly developed". A little "more supporting on details are needed".

Maybe I looked at the conventions, but it was only about misspellings.

Yeah. And then I think I went back to text again. Oh, so fast!

I think I looked at the transition words after checking the transition words. And then I thought that this person's use of transition words was formulaic. So "first of all", "secondly", and then these are uncommon uses. So I think that's what I wanted to briefly check.

And then went back to the rubric and then I was just checking the criteria for transitional devices.

And then I was wondering if this is B or C for the essay as a whole. But for grammar, I thought this was definitely B. But because of the good argument and ... Not bad organization. I thought this person would go to a C. But I thought this very borderline between these two. That's why I'm wondering here.

I think so too. I gave this person a C.

Rater 5 – Essay 4

So the first thing that I thought is this text is long enough.

So looking around here, I thought in terms of the paragraphing, this is not good. So that's what I thought in here.

So I think, get back to here. I started reading it.

I don't think I thought something is special. I think as I have told you all the time, so I usually pay attention to this marker. "I think" to show this person's opinion. And I thought oh, this person agreed with the GM crops.

So "we must support the research for GM crops."

So this is good way of expressing this person's opinion.

And, but as I read in the subsequent sentences, it seems like this person seems to disagree with GM crops.

"Because of the monopoly of a few companies who can develop these GM Crops".

So, I thought I expected to read something that is positive about GM Crops, but in fact, this person listed something negative about this thing. So, that's what I thought. So there should be some coherence.

And then, "I thought the applied thing with GM technology should be in the hands of the government".

And this person proposes alternative idea about GM technology, but I thought that there is kind of weak connection between this person's statement here and here. I would say this is a weak connection between these, so if this person wants to say this thing in here, this should up here, up here. So for example, I think I must support the research for GM Crops with something related to this claim. So in other words, the initial claim is not sufficient cause I cannot anticipate this sentence, based on this.

So, even though there is a weak connection between this and this, the idea is really interesting. Like in the case of an emergency, companies should take the initiatives to help people. That idea was quite nice. I was engaged in reading this.

So this one is repetitive because I caught, this is here "Increasing exponentially". Even though in here I thought, "Ooh, this is good word." But, this person repeats this collocation "increase exponentially".

Don't remember. Yes. So here. I looked up "increasing exponentially", good evidence,

I focused on this word. "I think only the government should have the right to patent the GM technology".

I think this is good argument. I think that's what I thought, even though I looked at this particular word, I think I considered the whole idea expressing this paragraph. I mean, in this portion. "Government", "right", "exponentially", "patent", "GM technology", these things here.

I went back to the beginning of the paragraph.

Why? "Population." I guess that I'm checking the connection between this claim and this. The thesis claims "I think we must support the research for GM crops". And the last one, I mean, the idea expressed in here "So government should take the lead on the GM technology". Yeah.

So I think I'm reading the second paragraph.

In the second paragraph, I think I thought, "Oh this person is integrating source text". Yeah. I think that's what I thought. But this caught my attention because of the citing the source. I thought this information sounds a bit redundant, like "50 grams of the golden rice or for 60 percent of the daily vitamin A requirement". And maybe we can write something more general instead of getting in this detail.

And again checking the thesis statement with this example, and then I thought, "Oh, this is a good connection between this and this."

So I'm looking at the argument, and organization.

And then I, at first, I was looking for the word “paragraphing” because there were only two paragraph. And I couldn't find that term.

And next I evaluated whether or not the text was easy to follow.

So that' why I looked at, “require some efforts to follow may be hard to follow”.

And I thought the text was good enough. I didn't have any difficulty in reading their claims.

So organization wise this text was definitely C or pass.

I think I'm comparing this criterion and this criterion for organization B and CD. I mean, the first criterion. Like easy to follow or not.

And then I'm checking the pass criterion too for organization.

Yeah. And then I think, at this point of time. I was wondering between pass or a C. Definitely not B for organizations.

And I'm checking the pass criterion in the arguments. And then comparing this with this, Pass or CD?

For grammar, I thought this text was accurate. And I didn't have any difficulties in reading it. So it didn't interfere with my comprehensibility.

So I decided to grade this as C, D plus.

And what made me not to choose pass was that the supporting details should have been clearer. So that's what I thought. But probably most people would say this text is pass. But I thought just C and D plus.

I went back to organization. I think I focused on the paragraphing again.

And then after thinking about arguments and grammar, which were quite good, but in terms of the paragraphing, it made me to choose C and B plus.

Well, I thought that there were mostly correct spellings. Yeah. So I didn't care about conventions, basically.

Rater 5 – Essay 5

First thing that I thought, "This is not as long as the previous one."

as long as the previous one

My really rule of thumb tells that this would be either B or C.

I was evaluating the accuracy of this word. "Strange [?] about genetically modified food".
"Strange about"?

Then I go back to "they might have some bad influence". The thing is grammatical error about relative clause.

"Bad influencing the future in the result of food supply companies might finally control."

I couldn't process this sentence quite well.

"Supply companies." Maybe this subject prompted me to go back to this part "they might have been some bad."-

There shouldn't be some subject in here because of the syntax.

I was, I think, evaluating this syntactic structure for the first sentence.

And I decided that this is not good so it did influence the comprehensibility of this text, I mean, the first sentence.

I think on the one hand this claim is good, "As far as I'm concerned isn't doing much more benefit right now."

But on the other hand, I thought this person should have been more deterministic about his position. This is in between I agree or disagree. That's what I thought.

"Many people who cannot afford food." (paragraph 2).

Well, one thing I can remember is ... this is the collocation that this person used in here. "In result of." I thought ... this caught my attention because of repetition and I think I also paid attention to "People who cannot afford food". Yeah, this part, and this part was also a repetition. "People who cannot afford food." "People who cannot afford food." "There should be alternative ways.

"Global human 10%". Even though I didn't read the first source text at this point of time, I guessed this person pulled this information from the source text. The fact that this person did it.

I think I looked at this part, "the first article isn't fake" (paragraph 3), though I had difficulty in understanding this one.

"If we take our eyes farther." Yeah, I think this expression sounded weird to me. "Take our eyes farther."

I think I re-read that sentence. "Things cheaper which have the same quality."

Probably the quality of sentence, I mean meaning-wise. This is very generic.

"People always to buy things cheaper which have the same quality." I was not sure about the uniqueness or relevance of this sentence to this topic of GM food. We can talk about this thing for other topics, not GM food. That's what I thought.

"Which means most of common companies cannot afford it" in the sentence "because the GM technology's hi-tech nowadays, which means most of the common companies cannot use it".

I just wondered the accuracy about this certain clause.

I also wondered the necessity of this relative clause because this basically rephrases what this person said in the main clause.

And then given that this is a final sentence of this paragraph of this text, I expected this person to write more concluding remark.

So I read this sentence and I was thinking about the other sense of the concluding remark.

I then go back to the rubric. I started with arguments. I usually started with this two in arguments and grammar.

Especially that text I think, had many grammatical errors. I usually evaluate the comprehensibility influenced by those errors.

I looked at the arguments, and I compared the big criteria with CD criterion, and the same thing went with grammar. "The source text is integrated, though not skillful".

As far as I remember the text just said ... just "the second article said" or something like that in the beginning of the second paragraph. I thought that was not skillful even though the sources are there.

And Grammar, and even though this person attempted to complicate the structures of the relative clauses ... the use of relative clauses was not accurate enough. Or I couldn't follow sentence. That's what I paid attention to.

And then I went back to the text.

I was looking for the attempts to create complex sentences in the text.

This shows the check of ... check whether this person skillful integrated the text source in the second paragraph "Just as what the second article said". This is almost, only phrase that introduces this person's citing.

And I was briefly checking the accuracy of grammar. "Take our eyes father." Again now. "Provide for our people."

I was looking at argument, yeah. And looking at the C criterion that whether, or not this person skillfully integrated the source. It seems I focused on the skillfully and then grammar ... checking the grammar. Checking B, yeah I think I followed all of these three criterion for common lexis. and went back to the text checking the grammar. "Things cheaper" "isn't fake".

I simply stayed around this criterion and then I moved to arguments. First I paid attention to C for grammar, and then moved up to arguments for C level. Yeah, it seems I looked at this organization, "unity." I don't remember whether I considered this part, but at least my eyes are there. I came back to grammar

Then went back to text.

I don't remember.

Going back to the rubric. The last time that I looked at the rubric, I wondered if this text goes to B or CD. That's why I'm looking at the CD criterion and thereafter I re-looked at the text. I'm checking the criterion. I'm comparing the criterion. I'm comparing these two, B and C for grammar.

This time I decided on B plus.

Rater 5 – Essay 6

At first, I looked at the paragraphing in the conclusion part, really consists only of single sentence.

And then these transition words are really formulaic.

Then I started to read the text.

And then here probably this is typo but it caught my attention.

Then I thought this is redundant

"however I am not a scientist".

And this is run on sentence. So there should be a conjunction word which should be there.

So even though I thought this is good way to express this person's opinion,

I also thought about probably this is redundant

and then given some grammatical mistakes here

I anticipated that this text would be B. I got my rough idea of the level.

"So as we know genetically modified food is give..."

This is kind of crucial acquisition error, so I think this shows that this person's grammatical competence is not quite good. For example, if I look at the subjunctive hypothetical errors and hypothetical conditionals, I really do not care about that because it's higher end grammatical structures; but for this one, I thought this would be really basic grammar; so, violating this rule tells the grammatical competence of this person.

And then this person decided cited the source in the double quotation marks; so, I didn't really read this one. Once I got this, I think I skipped this sentence and I moved to here because I thought I didn't have to read this whole sentence because this is not part of this person's writing. So, I didn't read this sentence and I moved to here to here.

Then I thought this last sentence should have been elaborated more.

The last sentence for the second paragraph. "so African people will gain less money if genetically modified food is developed".

So to conclude this paragraph, I think this person should have got out of the example (African people)

and then should have expressed his opinion.

So as we will see later I thought that this person misinterpreted the source text and then this person says if he doesn't get too much vitamin A, it will be cause blind or death. And at this point in time I thought that the source text didn't say something like this. The source text said that because of the deficiency of vitamin A, there are a lot of people suffering from blindness; so that's what I thought.

And then I reread this interpretation and then I finished reading a second paragraph.

I think I am checking whether this interpretation is correct or not. I thought that claim was coming from the second text, so I moved to there.

So, I was looking for the word "blind" because that was the crucial word. And I couldn't find that. And I moved back to the first paragraph trying to find the word "blind", which is here. And then I got that and then I was reading the sentence before this "blind" and what this says is: because of deficiency of vitamin there are a lot of people suffering from going blind.

So I thought that the claim made by the test taker was incorrect.

I re-read the essay, especially this part. The blind or death part.

So if you don't get too much vitamin it will cause the blind or dead, which I don't think is correct based on the source text. So that's what I thought. I think I am checking the same thing.

And at this point in time I made my mind. I took notes about my grade-

Rater 5 – Essay 7

I first - I think I have been doing this quite often, but I brief, briefly looked at the first paragraphing.

I think I looked at the transition words at the beginning of the paragraphs. Even though there are no clear transition words, the way this person began his paragraph was not formulaic like “first”, “second”, “lastly”. So, the absence of transition words, I think, has one positive and one negative side. So, yeah. That's what I thought.

At the time I was not sure if it is bad. I was planning to decide later on.

I thought this (the first sentence) is really good way to start the text, introducing the background.

“Some people”, but “for others they think”, this person presents two perspectives.

And then this person concludes this paragraph by presenting his personal thought, which is quite good flow for the first paragraph.

And then, he's citing source in the second paragraph. I think there should have been one introductory sentence before the cited text, the first sentence, second paragraph. But, at this point of them, I thought that was fine.

And now, I think I was thinking about the quality of citing.

Even though this person also directly cites this sentence as the previous one (essay), whereas the previous text just to citing the sentence without elaborating on that,

this person seems to express his idea, not that much but at least elaborating on that. So I thought that was good.

So, I think at this point of time, I think I've already shifted to the third paragraph.

And what I liked about the third paragraph is that even though there are some counter arguments against the GM food, this person proposes alternative idea to solve the counter arguments. This person basically says “governments should take initiatives about GM food.” That's what I liked.

This person effectively uses "I think" for the problem, to show his opinion.

And then he presents his idea "governments should change their policy and give the biotech companies have a chance to give up GM products."

Yes. "May sounds a little". I think this is because of the grammatical mistake. "Sounds" after a modal verb.

But the claims are really easy to follow. I thought that was good, especially I remember that the rubric asks whether or not you had to make some efforts to follow the text.

Now, the last paragraph. I thought this was good concluding remark. Yeah. Yeah.

And then before getting back to the rubric, I thought this text would be C D plus or pass.

And then I was reading, leaning more toward pass. Maybe I didn't look the B criterion that often. C and pass. I can briefly checking each component. Oh, that's it.

And then go back to text.

I think I was checking the transitional words after checking the rubric. I really didn't see many transitional words, I think, in this text. That's what I checked in here. So, because I was looking for a transitional words, I think I was paying attention to the beginning of sentence. But, for "vitamin C every day", those things, I really didn't see many transitional words.

Mainly, I'm checking the criterion about the pass students and there was clear focus.

I felt some unity in ideas in the text.

Yeah, I'm not sure whether the arguments were fully elaborated but, I could see some elaboration there.

I think I'm wondering between C and D and pass.

I think I didn't look at convention because for this text, conventions do not distinguish between C and D and pass because this text was, I think, perfect in terms of the conventions. So, I didn't really care about that.

And then - maybe I'm checking the whether or not this person cited the source text skillfully.

On this one, it seems I'm checking the comprehensibility of the claim because of the grammatical errors. As I said, the text was really easy to follow because of few grammatical errors.

I'm coming back to organization. I think I was checking whether I felt a clear focus in text.

So, I think at the time I decided on pass.

Rater 5 – Essay 8

First I looked at, briefly, the whole text.

And then one thing that I caught attention this time is "I am originally from the Dominican Republic" and I thought all this person was talking about very personal thing.

And I was wondering how this person substantiates his or her claim by saying this. But I didn't read carefully about this at this point of time.

I started to read the first paragraph and then I think I'm re-reading the first sentence "shouldn't be supported because it takes the natural components like vitamins, fibers, carbon, hydrates, and many other elements that provide good food" blah blah blah.

And I was just wondering the meaning of "take" in this context. Is it something like "takes out" or is it something that "gives" or "it is not there" or "it is there"?

And then I was wondering, "the independent citizens", the meaning of that.

And then I noticed this person included this paragraph by defining this term, which I didn't think is good. There should be more sentences to this paragraph. Yeah. I think I was just caring about "independent citizens".

And in second paragraph, I think as I expected, this person is talking about his own personal thing.

To me, this person seems to be familiar with the plans of Jim Booth.

But at the same time this sentence, "money is power and suddenly control people's acts if they do not have the ethical standards," and then I thought this sentence was not quite relevant to GM food argument.

Again I thought, again this person is talking about his or her self.

Now I am in the third paragraph. "A member of the church similar [inaudible 00:02:42]". "The probabilities to improve our health".

Maybe I was evaluating the accuracy of "increases the probabilities to improve our health". Maybe I thought there should be a better way to express this, even though it makes sense to me.

So this is what we already said. I'm re-reading it (the first sentence of the third paragraph again), trying to understand this text. Then go to the third paragraph.

I'm not sure if I thought at this point of time, but I thought that this person tended to say, "it is ... that ... It is blah, blah, blah." It's something negative. "It is clearly observed ... On the other hand". Yeah, so I think "it is" caught my attention. Actually, this person uses "it is" in here too. So, you think it really caught my attention. I thought that was kind of limited way of expressing this person's opinion. And then I think here too in text "it's observed that" and then this quote goes back to here, "clearly observed that". And then here, okay, "it is completely disappointing", so "it is" bah, bah, bah.

I think as far I remember I think I thought "it is completely disappointing finding people that blah, blah, blah" is really strong argument, so probably this person may need to use some hedging in here.

I think I'm re-reading these two paragraphs quickly [two body paragraphs] to refresh my memory before going back on the writing rubric.

So first, I thought this wouldn't be going to pass. Definitely not. And I wondered B or a C, D.

And at first, I thought I evaluated whether or not I made some effort to follow the text. It did require some effort on my part.

I thought at this point in time less likely to be B, more likely to be C.

Even though the claims were really personal, I could follow his or her argument.

The second thing that I thought, the relevancy of the claims. So as I said before, this person's argument sounded like a bit off topic.

So, I was looking for the criterion regarding that. So I think I looked at the word "relevant" in the rubric. "Most likely clear and relevant" (level C/D), and "most likely clear and relevant, but needs more explanation" level B). So, I was just comparing these two (levels), but actually both are talking about "it is relevant", right? They did not say, "Their arguments are irrelevant." So I really couldn't evaluate the text just based on this proposition. So, both descriptors in the rubric assume that the essay is relevant. This doesn't distinguish between B and C.

But I think grammar, I thought was good enough. Yes. And then I was staring at word "repetition" in the rubric because, as I said, I caught many of "It is. It is. It is." So I thought that was repetition.

And then later on in the text ... I'm rechecking the relevancy of these two body paragraphs because these two body paragraphs are mostly about his/her argument coming from his or her personal experiences.

Yeah. Another thing that I thought of is ... This statement at the beginning of the last paragraph "It is observed that GM Food is just a way to control people" sounded strong too. Sounded strong, like very assertive. So that's what I thought.

And then this assertiveness appeared to be here too, as I said "it is completely disappointing finding people that support". So this person, I think, is likely to express his or her certainty about her or his claim. So that's what I thought.

I spent a little bit more time on the last paragraph rechecking the grammatical criteria. I think I thought, grammar wise, the text was good enough.

And then I also checked the comprehensibility, which I thought was good.

And then I think I made my mind at this point in time.

Rater 5 – Essay 9

So, first, as always, I briefly looked at the whole paper and then I expected this person not to write the concluding paragraph because this is incomplete sentence and there should be one paragraph at last.

So, this person said "everything is affected by technology and the benefits from the state of the art based on imperfect light"

The first thing that I thought is that, this is quite strong again. So, is it true? So that's what I thought.

So first thing that I thought when I was reading the second paragraph was that I couldn't process this sentence, "Farms which are genetically modified two groups to draw more attention of biotech companies". I couldn't process this sentence.

Because of the grammatical violation.

What does it say? "More interested in modifying genes". Don't remember what I was thinking about.

And then I also had difficulty in processing the last sentence at the first glance. So, I think I was wondering the meaning of "American process". So I really couldn't understand this part.

But as I reread this sentence, I think I could understand that this person is talking about. I could understand this person's argument.

When I usually read some cited text, sources cited, I usually judge without knowing this person simply cite the information or elaborate something on that.

So far, the first sentence is simply about the text that they're reading.

I don't know. I pay attention to these particular words "go blind each year"... "vitamin A deficiency". I guess I am evaluating the accuracy of information in the first sentence of the third paragraph.

Usually I check the source text depending on to what extent the citing is accurate in line with the source text and then even though it is not perfectly accurate, as long as this person is understanding the source text, it's fine.

But for text 6, that was completely opposite, or not accurate.

That made me judge that text (essay) as B.

I think I'm reading this particular sentence but this is also coming from the source text. I think I'm just following that.

“should be allocated to improve...”

After the last sentence I thought there should be an analysis sentence to conclude this paragraph. Even if this person has finished this first sentence, I think there should have been another sentence that follow this sentence.

I think at this point of time, before going back to rubric, I think I was trying to read the text again to refresh my memory about the previous paragraph.

For example, I paid attention to this good phrase.

So, this person presents actually positive side and negative side; so there are two sides of GM food.

This person ends up with presenting his position but at the same time, this person doesn't specify his ideas in the thesis statement. So that caught my attention too.

I think I was trying to understand this last sentence of the second paragraph.

I think I thought this text would go to the C level and that is the most likely option for me at this point of time.

And then I'm going back to the rubric. So basically, when I was reading the rubric, my attention was primarily on C, D criteria. So looking at the organization and compare this to the criteria for B and then went to the arguments and details. “Arguments are mostly developed although there should be more supporting ideas”. I was just briefly looking at either pass or B to compare these to the C, D criteria. And then went to grammar and lexis.

I thought about that person attempted to create complex sentences but there were also grammatical errors

so I think this descriptor for C, D was in line with this person's proponent.

And then went back to text. I don't think I thought anything at this point of time.

Yeah. I think that by this time I was already determined to grade C.

I think I was looking at “comprehensibility” on the rubric (even though tracking doesn't clearly show up).

I think I marked down the grade.

Rater 5 – Essay 10

So I think for this text, I don't know why, but I didn't look at the whole paper at first. But I was just directly got into the text.

So I think the beginning sentence was really good, nicely introduces the background of the GM food.

And then I think the second sentence is quite nice in terms of the syntax. So even though It's complicated, but it's accurate enough.

So what I thought about the first paragraph is that this person doesn't take up clear position, so instead this person...

I anticipated this person to discuss both sides of both aspects of GM food. Pros and cons.

So the reason why I asked you to show the writing prompt is whether or not this person, the test-takers were required to take a clear position or not.

So if we look at the writing prompt this should, indicating whether you think GM food should be supported or not. So I thought that the test takers were expected to take a clear position instead of like their pros and cons. You know what I mean.

So while I was reading the second paragraph

I thought the use of vocabulary, I think the range of vocabulary was limited, but I think "subsidized rate" is quite nice. And that caught my attention.

And then when I was reading the third paragraph

This person integrates the text source. The thing is as this person says, in this last sentence of the first paragraph. This person keeps his position, like, in between. So this person neither agrees or disagree. With the GM food. It's a bad thing, I mean, in terms of the test response.

That "proponents". Yeah, I think I was just focusing on this contrasting statement "that proponents". So "proponents says" blah blah blah. "Opponents say" blah blah blah.

Which I didn't like. I expected test takers to take a clear position with agree or disagree. This person seems to be in between those two stances. And then even though this person says, 'In my view.' To express his opinion or ideas. He's still isn't taking his position, like, it is very good, but at the same time it has to be insured that blah, blah, blah. So even at the end of the text. He is just presenting pros and cons of GM food, which is quite in line with the thesis statement. But I really expected test takers to take either position.

"in golden rice"

So I think I thought that, so this" it is found that golden rice is the most [inaudible 00:03:50] blah blah blah".

Is something that should end up here, at the beginning of the paragraph? Because this is some example. So I thought there should be some sentences before this talking about the main topic of this paragraph particularly.

(last paragraph) I think I was wondering if I should take into account the fact that this person didn't take a clear position. Yeah, I was not quite sure about that. I think I focusing on this ambivalent position, taken by him.

So last sentence of the third paragraph. So contrasting statements. "Proponents" say blah blah blah. "Opponents" say blah blah blah blah.

So I was wondering if I should penalize, given by this unclear position.

And then I think, again, before I moved to writing scale. I think I had already determined to grade this as C.

Mainly because I didn't find this text difficult to follow, so it was quite easy to follow this argument. But at the same time, as I said. I think I decided to penalize this person by his not taking a clear position.

So, yeah, I think wondered if this person can pass. But I think I decided to wonder between B and C. Or because I think even then I still wondered if I should penalize this person.

I mean I think I was thinking about grammatical accuracy. And I thought that was good. I think sentences were accurate enough for me to process.

So yeah ... The arguments and details, I think I wished this person had had more supporting details.

And also I wanted this person to take a clear position.

So yeah in comparison with past says criteria.

I thought this text would go to C D.

Maybe I was just briefly looking at the text. Make sure. Then I was checking.

And then I think I marked down.

I think so.

RATER 6 (10/26/2018)

Rater 6 – Essay 1

I was reviewing the rubric to, like have a general idea, remind myself what was ... because the organization part and the grammar part have the greatest portion. I was, I think, looking those. To see what is the pass exactly and then, what's in the middle, like CD.

So first of all, I noticed that it ... so there are only three short paragraphs here.

All of them starts with “genetically modified food”, and it's repeated in the start of each paragraph.

And then, so I was a little bit confused because I didn't know how to start it. But, I looked at the introduction, I decide to read the first three sentences of the first paragraph. I kind of read those and then, yeah I was reading them. I looked-

I think I wanted to see how it's written. First of all, because it's really short. I wanted, I think, to know how it's written grammatically, and the structure.

And, then, I read the thesis statement. I think, yeah and I have to go back to the first sentence again to understand it, I think.

Then, I found that there's not a very clear thesis statement there.

Then, I tell myself, "Okay, I should look at the topic sentences."

And I looked at the second paragraph down. It seems like I spent a little time on this chiefly because it was not clear totally if he or she is actually agree or disagree. Yeah, it was not clear.

I was looking for something like the thesis statement.

Yeah, second paragraph, reading the first sentence.

And I was trying to figure out if this example is like a positive sentence, means he or she agrees with that. And then yeah, I think I didn't read the whole paragraph.

And then I went to the last paragraph.

So it's very short.

I just ... and I went back again.

I notice again that all of them start with the same phrases. I think I was thinking of that.

And then I notice that the start of paragraph two and paragraph three are like the positive reaction to the topic.

And then I found that this is not the concluding paragraph even, understood how they're trying to finish it.

So it's not a complete essay. There was just, like, comparing these sentences, that's why it just goes off and then.

I think that it's kinda fluent, the writing, it's kinda fluent. And there are some good vocabs here,

But I think it's not exactly C/D level, it should be like B plus maybe. So he or she can write, but it's not what we want them in C/D level. So it should be in B plus maybe.

Rater 6 – Essay 2

Yeah.

So, yeah, again, there are three short paragraphs here.

So I looked at every first sentence “first of all” (second paragraph), and then I went back to the first paragraph and decided to read the whole short sentences to see how the introduction is.

And I liked it (introduction) because it says- it starts with, "there are people who say that GM food production should not be supported", and also, "but there are some people who agree with that", “should be supported”, “should not be supported”.

And then the last sentence, very clearly, it starts with the word, "However," and "In my opinion."

And the correct use of punctuation.

And then it very clearly states that “it should be supported, actually”.

And then it also says the reasons because it's cost effective ... yeah one more- two reasons, actually.

And then, yeah, I went to the second paragraph. I think. Maybe I re-read it again to be sure.

It starts with, "First of all," which we usually say students start your topic sentences with such phrases.

And punctuation, that's good.

And then the topic sentence is actually good because it is exactly one of the points from the thesis.

so I went back to read (the thesis statement)

to sure that it's one of the points. In the introduction, it was- It says, like, "cheaper compared to other foods" and here (topic sentence of second paragraph), it says, "cost effective."

Yeah. And then ... yeah. Right there. Second and third sentences, and see? I saw that there is an example there which supports the topic sentence.

And then the last sentence he wants or she wants to say something, I think in contrast. He says- It starts with "However."

And the punctuation. The right punctuation. Yeah.

And then I went to the third paragraph.

And then, again, it start with those chunks, like, "on the other hand"

and then the topic is exactly one of those... I went back in introduction to see it. It's the same. And it was the same.

And then he or she writes some support sentences, and again use the word "however". And yeah.

But the essay is not completed,

and is, like, only two body paragraphs.

There is no concluding paragraph. [crosstalk 00:02:55]

And then, yeah, I decided that- okay I should- I want to say, I didn't notice any specific or multiple mistake. It's, again, fluent.

but I think it's C D. It can't be B

because there is a clear to the statement with the points. The topic sentences are good using those chunks at the beginning are very good. There are examples to support the topic sentences. Those are all good.

But it can't be pass

because it's not completed and, actually, the body paragraphs are very short.

So it's C D. Or maybe C D plus? I don't know.

If he had time, maybe he could write longer.

Rater 6 – Essay 3

Again, a very general look of the layout

and then I decided to read the first paragraph, which looks like an introduction paragraph.

And I read every sentence there.

I wanted to see if again he or she clearly stated what points he's going to talk about in body paragraphs.

And I read it carefully actually. I read it slowly. Yeah and I went back again. Actually I read twice. I read the introduction twice.

It was not clear for me what he wants to talk about exactly.

And then, I saw that there isn't a thesis.

“I don't think genetically modified crops should be supported because they are lots of side effects on it.”

But what we teach in the C level, we say that in thesis statement you should actually talk about two or three points and then you should expand them in your body paragraphs.

He did a good job, but I don't think it's clear that he agrees or disagrees.

But then in the topic sentence in the next paragraphs, he says another thing. He wrote something else. They are not included in the thesis.

So then in the class, what we say them is that include what you wrote in topic into your thesis statement. So, I was looking at those. While I was reading I was thinking of this.

Yes, I read it twice. So once I read the introduction paragraph very slowly I then tried to read again the second line and the third line.

And then I was deciding that “okay he's not talking about any specific points there.”

And then I went to the second paragraph. I read very quickly the topic sentence.

Then I just decided to go back again to the last two sentences from the introduction.

Then I went back again to the second paragraph.

It starts with a good chunk like "first of all". I looked at every phrase starting the body paragraphs and then the conclusion. And I saw yes there is "like first of all", "secondly", "third", “further information”, “conclusion”. These are not completely correct. You cannot just say "conclusion". "In conclusion". And "to conclude" or something like that. And even this one is not 100% okay but it's not bad actually.

And then I went to read the second paragraph.

And I noticed that actually, it's a very good topic sentence.

This point could be included in the introduction in the thesis.

So, I went back to the introduction again

To see if the topic sentences are one of the points from the thesis but it was not. But it's not. It's not included in the thesis statement.

Actually, this is what we teach in 101C.

I think it's maybe the other type of writing. You know in different countries or maybe in different majors they write differently.

Because most of these students when they come to 101C classes they write like this. They just say something, not the exact points that they want write in their topic sentences.

And then yes, I read the paragraph. Again, continued reading the second paragraph to understand better.

And I found a mistake already here, "it's really remain"; that was a minor grammatical mistake. And I noticed that one. When I noticed that, I found that “okay, because of this error, maybe

there are other kinds of these minor grammatical mistakes that I didn't notice." I decided to come back.

I read the introduction and the second paragraph again. I read the third paragraph, I read it more carefully.

To see the minor mistakes. Because of this one that I noticed. "It's really remain."

And now the third paragraph. Yeah, I read the topic sentence.

This is good. It starts with "secondly".

And then I found this mistake. Maybe typo or something, spelling mistake.

And then here, so something is missing there too. So, it's grammatical mistake.

I was moving to the fourth paragraph. So yeah I read in the fourth paragraph

And I found mistakes there. Grammatical mistakes, like in the third sentence.

Actually, there's a little typo. "Can'y", it's supposed to end in T instead of Y.

Oh yeah and then I saw something else. This is very minor, but they shouldn't write for example "don't have", the contractions. They should have write "do not have". So, it's the informal way of writing.

I found some mistakes, "in connect people". That's wrong.

And then I moved to the last part, which is the conclusion.

And then yeah, I was confused here because of this error.

"GM food are not supposed to be suppoeted".

It should be "supported" or "supposed" maybe. So, this is wrong spelling. "effecta". So that's some like spelling.

I'm not sure maybe, it's possible because of the keyboard. Sometimes they cannot find the right key. Or they just touch it by mistake.

I went back to the introduction. I just looked over thing to decide. I looked at again the topic sentences and the conclusion. And then I went to that "support" and "suppose".

Not sure 100% that it's good enough for a Pass level or not. And then I decide again, he's good but it's not like a Pass one. C plus maybe.

He could be one of the good students in my 101C class.

Rater 6 – Essay 4

First of all, I just got surprised by the organization, because it was a very long paragraph and then a short paragraph after that. So, it was not divided. So that was interesting. I just looked at the overall outline first of all.

And then said, "Okay, so I need to read every sentence then."

And then I read those, just read all of the sentences.

And then while I was reading, I was trying to see if the sentences are related to the topic or the sentences actually state the student's view or something like that.

"I think we must support the research for GM crops, but ..."

And then in the sentence, it was good that he used, he cited actually, he wrote the name of the author, Tom.

Not every student, test taker does that.

Yeah, again I read the first sentence to understand.

I saw "for example" on line six. That was good.

And the word "I" most of the times, is in small letter, which is not a good mistake at all.

He didn't write "genetically modified" in full but used the abbreviation "GM". He just said, "GM, GM", but we don't know what "GM" is.

And I saw many "I think"s

and then there was one more example, at the end of the paragraph.

Here, in the middle I saw like, "therefore".

But I noticed that some sentences are very short, like this one in line six, "This will make many people suffer."

There is actually a grammatical thing, sentence fragment or something. Suffer from what? So it was not quite a completed sentence.

But it was good that he used such words like "therefore" and then in the next sentence he says, "Next I think..." and then it gives an example. That was good, the flow I mean.

I was looking at "political power" here. Actually, the student used good word, good vocabulary. "political", "Companies, political", I think, have not bad knowledge of lexis. I like this one too, vocabulary, "supplies" and "exponentially increase". That's actually a good phrase.

And here I was looking at "case scenario", the words that are good choices actually.

And then the last sentence, I just wanted to say, it's, "all in all", conclusion of this paragraph.

We also teach this part in 101C classes too. Yeah, what do I do there ...

And here, yes, I was moving to the last paragraph

and then I noticed that, okay it seems like a student is doing summary of those two texts or something, which is not task 2.

Yeah, because this one, he's talking about Tom, the article, and in the second one he's talking about the other author. And he mentioned that. So I noticed that.

So maybe he's not aware of maybe the question that you provided or what kind of essay he should write. I think he misunderstood the task maybe.

The first one is a long paragraph. He doesn't any sub-paragraphs. There is only two paragraphs there.

And then, I read all of those sentences there

And I noticed that again, the small I's, which are not capitalized.

Here, I noticed that he's citing two people, so he's summarizing the source texts.

“The golden rice is a very good example.” Maybe I was processing the paragraph.

Yeah, and here, if I focused here right, because the grammatical things, "As studies show that is cost effective and also full of nutrition." It's not a complete sentence. There's a mistake there. Yeah.

And then in the last sentence, he has the closing sentence. “So, I think development of GM crops should be supported”, which is good.

But here, maybe he misunderstood the task or something

It's not an argumentative essay.

The organization actually needs more work.

And then, it's again, CD.

Rater 6 – Essay 5

Okay, so first of all, I again looked at the overall sentences and the organization. And yes, I just quickly looked and then found out, “okay, this is not the organization that we are looking for actually”. So actually there is no conclusion paragraph or the body paragraphs.

And then I went to the introduction to read this.

But before I forgot, I noticed many grammatical mistakes there in this essay while I was reading it, so yeah.

then I read that, it was not very clear or actually I didn't understand it.

But I was reading it (first sentence of the introduction) slowly. Yeah, that's why I went back and read it again to make it clear really for myself.

And then, yeah I read the second sentence. Not for any specific reason.

And then I read the last sentence. "As far as I'm concerned, GM Technology does do much more benefits right now, which is ... has problem, grammar problem there.

Yeah. So, there's no thesis. Nothing, but he states he's concerned, right? It has benefit but is grammatically not correct. And it's not a thesis statement that we are looking for.

Yeah, I was thinking that this is not, 'does do much more' is not correct. Yeah, I had to spend time there.

Then I went to the second paragraph. I read that.

"Many people cannot afford food with [inaudible 00:02:20]". Yeah, I read the topic sentence.

Actually, it's not bad.

Then I read the next sentences. "Produce more", yeah.

I know something like "that is to say", or those phrases. Not very academic choices.

Yeah, and then ... and actually here, I think there are some mistakes. "Nobody cannot afford food before. Before could eat more, and have more chance". I think it's not clear here.

Then, it start with, "There are many people nowadays." And "needn't" is not correct here. So there are, it seems like this student has grammar problems. So, that was the first thing that I noticed. Yeah.

And also again it was kind of summarizing those texts. Yeah, not writing this type of essay, it was like summarizing those.

Then I went to the last paragraph.

Yeah, I noticed that's a very short sentence at the end. "It's also benefit to hunger people", the last sentence in the second paragraph.

And I read that last paragraph. I don't know why I'm doing this.

Maybe grammatical.

"As time goes by, companies who cannot use GM Technology will break down. Because the GM Technology's high-tech nowadays, which means more stuff come and companies cannot use it."

You know the sentences like this are kind of good, but ... they are not the best way that it could be written.

You know, that's why it's not very clear.

Actually, in the introduction, it was the same.

You cannot say it's wrong, the sentence is wrong, but it was not the best way that he could write. So yeah. For example in this sentence, "Because the GM Technology is high-tech nowadays, which means more stuff come and companies cannot use it."

And then comma, "Finally, most food would supplies by just few companies," without any period at the end.

So, the word order should be changed I think here. Start maybe with 'nowadays,' and then yeah. The grammar and maybe words choices.

Yeah. I was deciding what level he should goes. It's B.

Rater 6 – Essay 6

Yeah again, general, I was looking at the general organization to see how many paragraphs there are, or what are the paragraphs.

Then I started reading the introduction.

It was not easy to follow again, not very fluent.

"However, I'm not a scientist. I also have my own opinion for genetically modified food".

So, there is a problem, I think, in this sentence. I think the function of "however" is different here. So, that was a problem.

"In my perspective, the genetically modified food should not be eaten."

But, it was not bad. It's better than having no statement. So, the student has something. He states he agrees or disagree with that.

But, again, he doesn't talk about the points that he provides in the body paragraphs.

Then, I was just reading it more carefully, trying to understand it again what's the student's idea, something like that.

And, then I went to the second paragraph, read the topic sentence, and I read it again, I think.

"Genetically modified food will hurt the farmer's benefits."

It has problem, too. "Will hurt the farmer's benefits"...benefits of what?

And, I read the rest. But, as you see, I'm reading it slowly because it's not very fluent.

Yeah, the grammar error, "although it do not change".

And, oh here maybe it was not clear, it says "In the text 1 the author mention that" you know, "the text 1", who is the author? What particular author are you talking about? What is text 1? We don't know that.

And then "it makes cotton cheaper to grow for highly..."

I like that, that at least it's in the quotation. He knows the plagiarism thing, or rules. That looks good, yeah, I was noticing that.

And then... yeah next paragraph, "Second", okay no punctuation there.

"The genetically modified food" and here, I think, the students could use the other variation like "GMO food", but he's repeating that phrase in all the paragraphs like 1, 2, 3. So maybe he doesn't have that knowledge, use other variations instead of repeating the same noun phrase.

"Influence the health of human"

And yeah. Yeah usually, there are some grammar mistakes. I was, you know, focusing on some parts or try to see if the sentences correct for grammar. That's what I was doing here.

"So, the GMO food will make a bad influence on human bodies."

So, I like it this part. That is like a closing sentence, something like that, yeah.

And then, the last paragraph seems like a concluding paragraph.

But it says, "according to my two reasons", not bad, but it could use a different transition word.

"I do not think that genetically modified food is a good thing for human"

So, 'thing' is not good to use in academic writing.

And then, actually, it says "it will inference" with a comma. Actually, it needs something to add like conjunction or something.

"It will inference on so many ways."

"on many ways", not very clear. What ways? The ways that you mentioned or maybe other ways that you didn't mentioned yet.

I went back to the second paragraph. Yeah, and I did it a lot, spent some minutes on just looking at everything.

I was looking at how many mistakes there are, the grammatical mistakes.

Because the organization is not bad, but his grammar needs to be improved.

And here, again the way he said, "As the text 2 said". The text doesn't say. The text "explains".

Looking again at the introduction paragraph, how it starts, how the paragraph starts.

And then I thought it could be C/D, but not a very strong student in C/D.

Rater 6 – Essay 7

Yeah, again, I looked at the overall look, layout of the essay.

And then I read the introduction. I was reading it again

because, again, the first sentence was not that clear.

Ok, and then the last sentence of introduction, it says: "As a personal thought about this argument, I think that genetically modified products should be highly supported and be widespread on earth".

That's good. It seems like a thesis.

And the first and second sentence in the introduction are not bad, because it talks about the both side, "that some people support that", "GM food should be supported", "but for others, they think that the genetically modified foods should be banned". It's good. Good background, and then the thesis.

And then I went to the, I think, next paragraph.

Yeah. And the next paragraph, I read the topic, and kind of went back to, again, refresh my mind, it was not in the last sentence of the introduction. It's a new sentence there. Here, this sentence talks about, okay, "the news report has reported a ten percent of people that eat rice has a vitamin A deficiency, and it can cause many" ...

And it's a good use of comma

and "and" here, to combine the two long sentences. That was good too.

And here, I also was thinking of, "Okay, does the student need to cite what text he's referring to or not?" I just spend very short time there, thinking that.

And then, I read the rest ... "But no genetically modified has been given ..."

I also thought about these words in the parenthesis. I was not sure if it's necessary to be in the parenthesis, or the student should have the skill to include it in the sentence, somehow. But it's in the parenthesis. I think it's easier to put it in the parenthesis, than combining or mixing it with the previous sentence. But I just did it very fast, focused on that part. I mean while I was reading, I noticed that.

"It's a great movement for food".

What is a great movement, first of all? What does "it" refer to?

And it's a very short sentence.

And then the question, "Why is that?" It's not clear at all.

Or, actually, he could combine these two sentences, and rewrite them, and then that would become a good question maybe. But not like this.

Yeah ... Again we have words in parenthesis there. "Roughly two ounces."

Yeah, it's good again, it's there in quotation marks. He knows that. Actually, he said the name of that article. American Journal of Clinical Nutrition. And "has just showed that". All of these are good.

I remember I read the third paragraph, and I read it twice, actually, I think, to understand it. "For some biotech companies, they do not have the power to develop the genetically modified food in some poor countries."

"I think for this problem, the government should change the policy, have a chance to develop if they do not have enough money ..."

Ah, yeah. This sentence. In the third paragraph, third sentence. "Then they can AS the people to collect." Something is missing in this part. "Ask", maybe.

"And then it sounds a little bit difficult to them, but to care about their people's life ..."

And you know, using too much of these pronouns: them, they ... We are not completely sure who is talking. I mean, the student's talking about whom, exactly? The people, government, you know? So one of them should be the real subject, not the pronouns, actually.

Oh, here there are commas here. Like it says, "It may sounds a little bit difficult to them," comma, "but to care about their people's life," comma, "I think this plan is needed for them." Here, should be period. Just, period.

"I think this plan is needed for them," again, for whom? So the pronoun is so far away from the real subject over there.

So, and then, "It's helpful for poor people also it is a better way..."

Yeah, and the last sentence has problem too. It is need punctuation, or divide the sentences, or start with "also," and then the rest of the sentence. So it needs revising, kind of.

Yeah, then I read the concluding paragraph.

It seems like concluding paragraph, because it starts with "at last." He means "finally," something like that.

yes, a problem here, "health", not "healthy"

and then comma in between. That's not correct.

"And is the most important thing for family ..." "Now the GM product has been developed, people should use the chance ..."

Actually, there are some good use of grammar, like "has been developed," it's a good one.

Oh, I was done here, and then went there. I was just very overly look at the sentences, to see if I can catch anything more wrong or right.

And then, I think it's, again, C/D.

Rater 6 – Essay 8

Again. To look at the essay organization overall.

And then I went to the introduction paragraph to read that. I actually read this paragraph twice to understand how the student start writing or his or her grammar, or the type of words that they use. Something like that.

Yeah, it (introduction) was not clear again, or maybe I couldn't understand it, so I read it twice I think.

And then here "As read in text one". "I read", so it's missing the subject.

I stopped there because it said "As read in text one, GM food can be controlled by the government." Yeah, I needed to read it again to understand what the student really means.

"Which creates a dependency that reduces the probabilities of having independent citizen."

Actually, this is good, here like the choice of the word, like "dependency and reduces the probabilities of having independent citizens", these are good.

But there is not a clear thesis here.

And then I went to the second paragraph, and while I was reading, I noticed that it's ... I don't know to say, like narrative, it says or something. Because it says, "I originally from where I have known about companies."

So, the essay includes the student's experiences, or observations, those things. It's not wrong, totally, but actually there is no topic sentence from the thesis, the organization has problem, actually.

Yeah, in the second sentence, here it says, "They are not punished as they should by the country justice systems."

"They are not punished as they should", I think, is not right, right? I think, yeah. It's not okay. "Money is power..." Oh, so the important thing here is that it's not talking about GM food, it's not focused on GM food, but it talks about the money something, and then "they should be punished".

Oh okay, so this means "Because GM food is hurting our communities, and because they pay enough money to government, they are not punished."

I think this part mostly focused on money part than what GM food does, or something like that. It's kind of off topic, but not totally, because it says GM food.

Then the last sentence says, "That being said, cheap does not mean quality and GM food does not mean healthy food."

So, yeah. Maybe it's not wrong, but I didn't like that because it's not totally talking about the GM food.

It's like you're talking about money, and "anything cheap doesn't mean good". And then so GM food if it's cheap, it doesn't mean it's good. It says something like this.

Yes. And I then went to the other paragraph, the last sentence of the second paragraph.

And again, I was just thinking about what it said. That it's like about money over GM food, so what's the relationship?

And then the third paragraph ... it starts with, "It's important to mention that I am member of a church that encourage to have food store."

So instead of this, the student could change it to a good topic sentence. So it could leave this part "that I'm member of church, I'm encouraging what..." This part could be one of the examples in the body paragraph actually. So you should have a topic sentence, and then you can use your

personal pronoun or your experience or background as an example to support that topic sentence. So while I was reading, I was thinking of that actually.

And it was good here, that the student started with "having our own garden. It's like sentence variety

It's what we teach at university.

"Consuming and can also help to build a solid ..." then I went to the last part.

And I don't know, it's like a body paragraph or a concluding paragraph, but it starts like, "Based on the text, it's clearly observed that GM food is just a way to control people..."

And there is nothing more powerful to do it" To do what? "... then by food and money". So, the second part of the sentence has problem after the comma actually.

"On the other hand, it's not only the facts of having a cheaper food, it's about finding an easy way out in an economy that is constantly changing."

So, it says "a economy", so wrong article

It seems like a concluding paragraph, kind of, but I'm not sure.

Actually, he's not adding any new information here. He's again talking about food and money. So I think he's not adding, he's still talking about money, cheaper food, and those things. And he says, "As I mentioned before, cheap does not necessarily mean good or excellent."

So that's like concluding paragraph. So if the student is adding new information, then it shouldn't be the concluding paragraph.

But then it says in text two, new information from text two, but again, it's the same concept, the money. And the cost, yeah.

So, organization has problem.

"It's observed that investing in supplementation... "fortification programs would cost more"

These are actually good choice of words.

And then there are some grammatical, actually like sentence fragments, not very clear sentences, there are such problems.

I went back to the introduction and just looking again at the paragraphs, very quickly.

I think, it's a C/D plus.

I just said plus because of the choice of some of the words is okay. And I think the students only maybe have problems with the organization and telling them that you shouldn't write it like a narrative. Instead of talking about yourself, you can use those parts as an example, something like that.

Rater 6 – Essay 9

First of all I noticed that there is no concluding paragraph, so again, talking about organization.

And then I went to the introduction paragraph, and again I read it slowly, to see how the student write that.

Yeah, I remember that. For example, "the cons and pros of taking this ability", that was good to write actually. It means that's he's talking about both sides, kind of.

And it says "however, I believe the negativity overweight", it's good its kind of clear, what he wants to talk about.

First sentence, yeah I read it twice. And then I moved to the second paragraph, "in the first place, farms which are genetically modified, to grow crops and draw more attention to biotech companies in order to apply..."

Yeah, not bad. So again like the other things that I said, it's not included in the topic sentence, but it's a good sentence.

And then, somehow the next sentences are related to the topic. So, the other sentences after the topic sentence it should be related to the first sentence of the paragraph. You should support it, it should be related to that. Which is this one, here, it is related.

That's the other important skills that they (students) should have.

And again the use of words like "however" in the second paragraph, and a comma.

Yeah, and now I was thinking of, actually, all of these together. Are they related?

What are the grammar mistakes, if any? Which, I think I didn't find any grammar mistakes.

Moving to the the last paragraph, yeah. So- So, it start with, "on the other hand billions of people rely on rice as their main course food". And talks about the vitamin A and those things.

And with the numbers, and I was thinking if again, the student needs to cite the article, the name of the author, or where he read that, or found this information. Because, there is like big numbers, percentages so maybe, something with plagiarism.

So he needs to know that too.

And then "among them, half lose their life within a year".

Oh and actually, "on the other hand", I think it's function is different here. Because the second paragraph it talks about what, draws more attention of [biotech] companies. I think "on the other hand", it could be used in another way. Because, he's not talking about the opposite things. It's adding, like "in addition billions of people rely on rice". It's adding more information. But at least it's good that the student knows that he can start with these phrases, fixed phrases, at the beginning of each paragraph. That's good, yeah.

And again, yeah, the sentences are related to the topic.

There are not specific grammatical mistakes.

Yes, and the other thing that I noticed for example, some of these sentences in the last paragraph are like giving examples or make it clearer so he can use the phrase of like "for example" or "in other words", it should be like this, it should be like that.

He needs to know that.

And then, I think I read it again, kind of, the whole essay. Very fast, because I was not sure. I think I'm thinking, to see if I can finish it or read it, read the whole thing again. And I think I read the essay again very, very fast.

Yeah, to read it again just to be sure, I think, of the level. Like that C, D level.

Because the sentence structures are good, but he needs some improvement in like what I said, okay these are good sentences but they could use "as an example" or "on the other hand" for example here. Knowing the function of these words, in what sentence they can use it. Yeah.

And actually, maybe, the student is not completely informed that a body paragraph needs a topic sentence, and then the sentences should be related. And then some supporting sentences, examples ...

Yeah, these are the things that he can learn in 101C, a class I teach.

So he has the general idea, and it's good; but he needs some improvements.

And actually, it's (the essay) not done. The last, the very last sentence says, "other budgets can be applied to improvement of agriculture facilities, so that better earth can be"... And it's not done, yeah.

I think because of the time or something.

Rater 6 – Essay 10

Again, I looked at the overall paragraphs first, and then I moved to the introduction part to read it.

And then "some people are favoring the GM foods as they are getting benefits".

I can't say. It was good, the sentences are good.

But it was not written kind of the same way as the other student wrote.

Because it says, for example, if you put "are favoring the GM foods as they are getting better", you know the language is kind of different. In a good way, actually. "getting benefits by having control over the entire food chain", so I like that.

And I actually didn't get it at first reading, so I read it again, I think.

"They do not have access of such technologies of producing GM foods nor they have enough capital".

So, they DO NOT have any access of such technologies NOR this. This is a good structure.

And then, at the end, it says "this essay will describe both of them and finally come to a conclusion".

So, totally different than the other essays that I read.

It's not what we exactly say to do in one of 101C classes,

but actually it can be a good thesis. So, actually it says that I'm going to talk about both sides, and then there's a conclusion.

Yeah. And then I moved to the next paragraph, I think, and read that.

So, the first sentences in every paragraph doesn't start with any fixed chunks, it directly goes to the sentence like "GM technology are heavily controlled by the biotech companies as they fund the development of the technology and keep the control of them through patenting". Yeah.

The language is kind of different than the other essays, kind of in the wording, you know, the grammar type.

And I have such students in my classes too.

It's not very easy to follow, but it's not wrong to.

You know, maybe it's... they're translating from the language or from how they write it in one, and then I interpret it like this. Maybe I'm wrong, but... so this is like that.

"GM technology are heavily controlled by the biotech companies as they found"

We didn't see any of these kinds of structure in other essays.

And it's good, actually. It's more complicated, not that much, but it's not as straightforward as the other sentences in other essays.

And then, it says "Rich nations like US are getting benefits of such technology as they are providing it to their farmers".

And then, "as a consequence", the liking word.

Again, I didn't see this in the other essays.

"The American producers are giving tough competition by providing these cheap..."

But it's short

and doesn't include any examples, or it doesn't start the way that we tell them to start - the topic sentence. And then actually the topic sentence is not very clear. Like this paragraph is going to talk about what? You know. It's kind of allowing to... if you read that again "GM technologies are heavily controlled by biotech companies as they found the development of the technology and keep control..." there's not any key words in this topic.

So, that's also the other thing that we teach them in 101C that there should be a key word, okay it's talking about this. This paragraph is talking about this, this key word.

And then I think I moved to the third paragraph, I guess. "Golden rice which is very effective in eliminating Vitamin A deficiency".

Again, the topic is kind of better than the second paragraph, because... Actually, kind of, you know, that's talking about Vitamin A deficiency. It's kind of clear.

"And it can be produced with a much cheaper cost"...

So the topics sentences are not very straightforward to find what the paragraph is going to talk about, I think.

And then, yeah, it was also interesting that a study of this, then the year "Lomborg (2013) found that..." That's really good in terms of citation.

This is something that we teach in 101C about citing, and the reporting words. And this is good.

And, oh here, by the way, these two writers in the third paragraph, second line. These two writers Silva and ... Where is it taken from? I'm not sure if it's in the original text or it's something that the students knew already. Anyway, it should be cited in a way. I don't know.

So, "is against of such foods, saying that it is creating the problem". And then, "the proponents of such GM food".

Again, we didn't have such structures in the previous essays.

"The provenance of GM foods." The word proponents, I think, is good to use and other students didn't use that.

So, "they are rich source of Vitamin A while opponents opposite GM Foods says they are not sustainable."

It's good, actually, the ending. The closing sentence is also good. And, actually, there is a closing sentence because, many students don't do that. They don't have the closing sentence.

And then I moved to the last part.

And actually, here... the interesting part was that in the introduction, the last sentence says, "we finally come to conclusion", but this paragraph doesn't seem like a conclusion.

They say "however, it's found that golden rice is the most cost effective source of Vitamin A."

Oh, okay, maybe "however" is misleading me here. So because when they say "however", I'm expecting to read something else. But, it talks about Vitamin A, which was taken from the third paragraph. So, it could be a concluding paragraph, but the starting word is not right, "however" doesn't belong here, I think.

"Therefore, in my view, it's very good to have such technology."

That's the conclusion.

And I read the last sentence again because I was not sure if I fully caught everything in the first reading.

The word "affluence", if they were not taking from the original text, those are good, actually.

I'm not decided yet about the level. It's very close to pass.

I'm thinking about the topic sentences and the thesis statement

and I'm not sure if, in 150, they teach these again or not. I'm not sure. So, if they review all of these again, then it's good to go for 150.

I can say this is a Pass.

But, I'm hoping that in 150 they're doing the same, maybe, kind of a review, or overview of all of 101C things. Not 101C, but the general things like the topic sentences, thesis.

RATER 7 (02/01/2018)

Rater 7 – Essay 1

First I looked at the rating scale before I move on to the text. I tried to remember especially the expressions involved in the scale. So, I just tried to remember what was the difference between "B" and "C" and "pass" in terms of organization and special arguments.

To me the first two parts, arguments and organization, are more important than the grammar and convention part.

I just read it from the beginning to the end without stopping.

While I'm reading it, I've tried to find the thesis statement and a topic sentence. Tried to catch the main idea, the argument. I don't think I could find one. I mean the thesis statement especially.

That's why I went back to the first part a couple of times. So, I was just reading. Then I read the first sentence of the second and the third paragraph.

At this point, I didn't feel like this is going to be a kind of "pass" essay

because I couldn't understand the argument. But while I was reading I was like "what was the argument here?". So, it was not clear to me.

But the grammar was okay. The grammar problems didn't distract me at all. That's why I could read to the end without stopping.

Then, I moved back to find the thesis statement in the first paragraph, yeah. And then I think I moved to the topic sentence of the second paragraph. I couldn't find, or, they didn't satisfy me.

So that's why I decided this is not a Pass definitely. And since it doesn't have any grammar issues, it's not a "B" either. So, it was a good fit for "C" and "D".

I looked at the scale just to make sure. I was just clarifying myself, or, confirming myself. This was just kind of, confirmation. I just focused on the C/D part because that was my decision.

So, the argument was vague. It fits very well but the examples to me ... I mean, they were clear, but the argument was vague.

Sources were integrated

so then I decided it was not a "B".

I was looking at grammar and lexis for C/D. Just for confirmation.

I was looking at the essay again. I don't know what I was thinking there. I just confirmed myself.

So definitely this is a C/D.

Rater 7 – Essay 2

I decided not to look at the rubric in this one because I wanted to read the text first.

So, I just tried to find the ... I did the same thing. I tried to read from the beginning to the end.

I kept in mind that I might catch the thesis statement and the topic sentences.

The thesis statement for this one is not clear to me.

So, I read the thesis sentence a couple of times to make sure that that's the thesis sentence. So, I did it a couple of times. //

So, I found it (the thesis), but it consists of two sentences. So, the thesis is there, and the controlling ideas were in separate sentences. And it was not good for me. So I was not really happy. But I could find the main idea of these.

I was going to the next paragraph. I read the topic sentence twice and moved on, I guess.

In that one, I read the topic sentence twice to make sure the main idea of that paragraph.

And then while I was reading the examples, I was thinking about, "how are they related to that main idea."

And I was thinking "I like the examples here." They are very clear and support the topic sentence.

And I liked the grammar. Although there was a couple of grammar issues they didn't distract me.

So, the examples were OK, and they were from the sources, the reading texts. So, to me, it was integrated successfully.

Not perfectly though, like how they should cite, but for that paragraph, I was happy.

I decided this is not a B definitely

because he knows the organization of an essay, but in terms of thesis, topic, and the introduction of examples and everything, I was trying to decide whether this is a C, D, or pass.

And I decided to read more and to the end.

And the last paragraph was not complete, which was not a big issue though.

But I didn't like the topic sentence of the third paragraph. It sounded like a summary to me.

That's why I read the topic sentence of the second paragraph again

to compare or to check if it was OK or not. Yeah, I read the thesis and the topic sentence of the first paragraph a couple of times to compare because the last paragraph was a little bit.

So, I decided this was not a B definitely.

But I was checking the area in between C/D and Pass. I was trying to remember what was in Pass to make sure that if it fits there or not. So, I was going between C, D, and Pass all the time.

At this point, I was a little bit towards C D because of the last paragraph.

And then, for argument, it was OK. Not fully but mostly everything ... It fits C D very well.

In terms of organization, I like the organization because ... The flow was OK. Although the major sentences were a little bit problematic.

So that's why I was, "OK it's not exactly C D, and it's not exactly pass because pass it is well organized and everything." So, I was in between, definitely.

I was going back to the text again to find a couple of grammar problems if there was any and if the grammar problems were serious.

I was remembering a couple of grammar issues. Yeah, the second paragraph, the last sentence. So, "which means" something. So, there was a really simple grammar error.

Then I think I decided this is not a pass.

So, a pass student should know that a sentence shouldn't start with "which means" clause.

So that's why I said it towards C/D, but I didn't want to punish him just because of that error, so I decided this is a C/D+. Not exactly C/D.

Rater 7 – Essay 3

I started from beginning 'till the end, focusing on the major sentence.

From the very first sentence, I felt like this is gonna be a good essay.

I think because of word choice, like "alter",

and the grammar was great.

In the second sentence and the third, there were words like "matting". I don't know what's that. "chaging". I felt like this is a good example for the spelling problem.

Yeah, so there were many spelling errors. Like three words in a row. I was like, "This is too many."

So, I was like going back, reading again trying to understand. They did distract me.

But still, the intro was good. The idea was good. And I liked the thesis. It was simple, but it was clear, and the argument was there.

Then, I moved on the next paragraph, I read the topic sentence I guess twice.

Again, there's a spelling error, there so I focused on it too much

And also, there's a grammar problem there. It is one of the major sentences; so, it should be accurate. It bothered me a little bit, yeah. It bothered me.

But still the idea development was good to me.

At this point I haven't decided yet, but it was definitely not B. It was between C, D, and toward pass

because of the idea development and maybe word choice I would say.

But I decided to read the rest.

The second topic sentence was better than the first one definitely. So, I like that, and I like this paragraph better than the first one.

The examples were great.

But there were again, a lot of spelling errors, like missing letters. For most of them I could understand the words still, although the letters are like mixed up, I could understand it; so, it didn't really affect my understanding of the text.

At that point I was thinking, I was thinking, "Okay this guy wrote a long essay, compared to the other ones."

So, I guessed he tried to type fast. That's why he did spelling errors. I didn't feel like he doesn't know how to spell the word, but I felt like, "This is coming typing fast," rather than lack of

knowledge. Maybe it's the testing experience, how he was trying to write fast because of time pressure.

I didn't like the transition of the third paragraph ... fourth one, "For your information," I didn't like that ... To be honest. It starts a bit like, "First, second and then ... for your information," and it doesn't sound right to me.

But the sentence itself was okay, I mean although it has grammar errors, it was one of the main ideas. I was happy, kind of. The examples were good here too.

The word choice, they were not simple, and I like this paragraph too.

But again, there's a lot of spelling errors

Something like grammar errors. So, like "consumed". That part got me thinking.

So, the conclusion was okay. It was just a summary and it was a kind of finish; so, I was happy to see a completed paragraph.

I didn't see this in the previous essays.

But still, at that point I haven't decided my grade yet. So, I wanted to check the rubric, especially the grammar and the spelling part.

At the beginning of the session I didn't look at that, but for this one I wanted to see if there's any difference between C, D, and pass in terms of special spelling. Interestingly I didn't see any difference. I looked at both a couple of times. There was no difference in terms of spelling, which I didn't notice before. The phrasing was exact the same.

So, "spelling's mostly correct," which is not.

And "spelling errors are minor; not interfering," and then I was thinking, "Did they interfere?" Not much to be honest. I could still understand the text and everything.

And then I decided if spelling is not an issue, the rest of the criteria were to me is towards Pass.

Then I moved to argument, but I was thinking this was towards pass. I just tried to confirm my decision towards pass, so I read the descriptors.

Yeah, at this point, I decided on the Pass. I felt the student doesn't need to take C class or maybe D because this is what we teach there.

If that person knows how to write the thesis statement, how to write the topic sentence in just forty minutes, I think there's no need to take one semester to learn how to do this in like one week or two weeks, because in the assignments they do it in two weeks, this person definitely does it in forty minutes. If he is supposed to write an essay at home without a time pressure, there would have been no spelling errors. The idea development was good here and it would be better if weren't time pressure.

I was reading just major sentences like topic sentences and language again.

Rater 7 – Essay 4

At first, I just looked at the whole text trying to see how many paragraphs are there, and they were just two.

I was not happy with that to be honest, because the first paragraph was really long and the second was like short and they were just two paragraphs, so I wasn't happy.

Starting from the first sentence, I felt like this student is good at the sentence level, and there are not many problems, almost none, so I was happy with the grammar level.

But since the organization as I said, did not look at first okay, I was thinking, "Okay, I should find the thesis statement."

But at the very beginning of the paragraph, I saw a reference from the source text. I was like, okay, what is happening? Is this intro? Is this the body paragraph? I couldn't decide it. So, I was a little bit confused in terms of the organization especially.

Still, the word choice was great.

So, I was like, "okay, this can't be a B"

because of the sentence level, because he doesn't need any grammar or grammar instruction, but he should know how to write an intro paragraph, how to write a thesis statement.

I felt like this is not the pass either,

because still I couldn't find the thesis statement, and the argument was not clear to me. So, is he just summarizing the text, what is the argument there? The flow was not okay.

I remember going back and forth within the paragraph trying to catch the idea. I focused on that sentence "therefore, I think..." too much.

I felt it sounds like kind of thesis statement, because there is this "therefore".

So, I felt like, "okay, the previous part was kind of providing background, so this sentence is kind of support to the thesis. So, this should be a kind of good thesis statement of an opinion paragraph. I said okay, that's the thesis, kind of."

But still I was thinking "I should read the rest, because normally if that's the thesis, he should move on to the next paragraph."

Yeah, I went back to see reading what I think is the thesis again...

Because if that's the thesis statement, then the first part should be kind of a background information, and I was checking if that's the background information or not.

But he was using the information from the text and also, he was building a kind of argument. He was talking about his own experience as well.

I was like, "it didn't sound like background information". I was, again, confused.

So, I decided to read the rest of the essay.

And then I saw the word "next" there. I was like, "what's happening?"

I think I went back to see the previous parts, this and the other parts.

Yeah, I went back to see if there is any listing of ideas. So, if there is this "next" there, so there should be others transitional words for listing elsewhere before this right?

In that part, I was like, “okay, that's the last part of the first paragraph”. Therefore, I think this sentence is not the thesis statement, the last sentence here should be the thesis statement, because that's the last sentence of that first paragraph. I was still trying to identify the thesis statement.

Then, I saw “all in all” there, a transition word. But generally, we use it to conclude the essay or if this person is writing a paragraph, then that can be concluding sentence. But I saw there is another paragraph there. I was like, “again, what is happening?”

And then, “I think the government should have the right...”

I felt that sentence sounds like a thesis statement; so that was the perfect main idea to me.

But the transition word confused me a little bit because we don't use that transition word before a thesis statement normally. I was like, “did he misuse that transition word, or is this is a concluding sentence?”

Still, I haven't decided, so I read the next paragraph.

And then I saw he's referring to the other text, the second text.

I felt like, “okay, I think he talked about only the first text in the first paragraph, he summarized it, and this “all in all” sentence was the concluding sentence actually. He basically summarized the first ... or he summarized and commented on the first text, and then he noted on the second text.” That's what I felt at that point when I read the first sentence of the second paragraph.

Again, since the grammar was okay, the grammar didn't distract me at all,

I decided to just focus on the idea.

I think I went back to the rubric to see, especially for the organization argument, what is the difference between B and C/D.

So, I was like really in between B and C/D here.

At the sentence level, it doesn't sound like B, because he has many complex sentences.

So I felt he was more towards C/D.

Then, I was like, he has a lot of organization problems.

So, I just double-checked the difference between C/D, and B. I focused on the keywords like “somewhat organized” in B, and “mostly well-organized” in C/D.

So, I said, "Okay, there is not a mostly well-organized definitely."

So, I was like more towards B.

For the argument, definitely the argument was vague, because I couldn't find ... I really struggled hard.

So I was just double-checking on the scale. I saw in C/D, there is this “mostly developed”

Then, I said, "No, it's not most developed."

I was like, again, more towards B.

Yeah, at that point I decided that this should be a B+.

Yeah, because, again, that's my way of doing. I went back to confirm my decision.

Again, I was trying to find the argument, because that affected my decision. I was like thinking, "This is a summary or an argumentative essay?" Then I saw this reference to the text, so it was a mixed, both summarizing and building an argument, which is good actually.

But he was only talking about the source text in that paragraph. I confirmed that he just refers to the first text. So not enough. He should have mentioned both.

So, I was reading the rubric to confirm my decision. For the grammar part, if I would see something like convention problem, like the spelling problem that I had with the previous essay, like the same expressions in terms of grammar between C and B, I would be confused. So, I was just checking if the grammar in B part should be problematic.

So, I was like, "okay, this is not a B."

So in the end, I decided this is a B+.

Rater 7 – Essay 5

First, I just had a look at the whole essay to see how many paragraphs are there to get just the general idea.

Starting from the first sentence, the first sentence sounded a little bit weird because there was a run-on sentence there. It was a long sentence consisting of three different sentences.

So, I had to read it a couple of times. I read it again and again because I tried to understand what he's trying to mean.

Then again, I did the same thing, I tried to find the thesis statement and I was thinking "okay this is the introduction." I could catch the thesis statement, the main idea.

I was like, in terms of organization, this person knows how to write an essay.

So, the next step was to find the other major sentence, like the first topic and the second topic sentence.

So, the first sentence, the topic sentence. I didn't like the opening expression "just as what the second article said." I didn't like that much, but still it was a topic sentence.

So, I haven't decided if this is a pass or CD, but I was like, this is not a B. Because after reading the thesis statement and after seeing the topic sentence in the next paragraph, I was like, okay this is not going to be a B.

I was thinking "just read the rest."

The second paragraph, in terms of the examples it was okay.

There was a reference to the text and the combination of the text and his or her own idea. So, I liked that.

And there were some nice transition words, so the flow was good in the second paragraph.

I was like, "okay the grammar is bad, it sounded okay,

So, I should focus on the other part, the organization, the argument, development."

After I am done with the second paragraph, I read first sentence to make sure this is the topic sentence of the third paragraph.

To be honest, I liked the expression in the first paragraph. “The worry of the first article isn't fake.” I don't know, I kind of like that.

But I was like, he talks about the second article first and then he moved onto the first article, so that kind of confused me a little bit but then I said, “this is not a summary, so he does not have to go in order.” So, he kind of rearranged the text based on his own argument. So, then I said that's okay. The third paragraph was okay

until the last sentence because there is this “finally” there and it sounded like the run-on sentence and I didn't like that.

And then I said, it was this conclusion paragraph? So I got confused.

I tried to read it again to make sure that, I'm reading it again to make sure that this is another body paragraph not the conclusion paragraph.

So, I went back to the thesis. I went back to the topic sentence of the first paragraph and second paragraph to make sure the organization is okay.

After this, I decided that “ok, the conclusion paragraph is missing.” He has the intro to the paragraph and the conclusion is missing.

I didn't decide yet at that point, but I decided this is not a B. I was like more towards CD rather than pass.

I didn't look at B. I was just focusing on CD and Pass. Just for the organization argument and grammar. I'm not sure about that yet.

Yeah, So I was making sure ... was the grammar in the past? A wide range of grammar structures? Not a wide range. But the grammar was okay, so I was more towards CD.

I didn't look at the conventions because it was okay.

At this point, I decided this is not a pass, but I was not sure if it's a CD or CD+.

So, I'm reading this again, just skimming to make sure that I'm CD plus or CD. So here I'm just confirming myself skimming again. I'm just looking at the major sentences and everything.

So, since the conclusion paragraph is missing and there was some run-on sentence and everything, but still this person knows the general structure of an essay.

So, I was like, if this person took 101C, he would be talking about these things anyway and he knows already.

But I couldn't say this was a pass either. So that's why I said this is a CD plus.

Rater 7 – Essay 6

Again, I just didn't look at the rubric, I just moved through the text.

The text was okay when I just looked at the whole thing, the paragraph numbers and everything but it was not long enough, I guess.

The first couple of sentences sound like this is not going to be a Pass, I guess. That's my first impression.

The word choice and the sentence structure were kind of simple to me.

I haven't decided yet, but I was like, "okay this can't be a pass, actually."

Then I saw this, "however, I am not a scientist, I also have my opinion."

So, I think there is a misuse of the connector there.

It confirmed my decision that this is not a pass, but I was like "is it a B or C/D?"

Here, I was trying to identify the main idea and the organization.

There was a thesis statement, like, "In my perspective GM food shouldn't be eaten."

And I'm just reading to make sure that the word choice and the vocabulary is not that high level.

Then, I moved on to see the topic sentence. I was going back and forth to compare the thesis and the topic in the second paragraph to see if the topic sentences explain the thesis. Again, the topic sentence was really simple to me, just a simple sentence, but still it was a topic sentence.

Again, I couldn't decide whether C or B; so, I wanted to read the rest.

I thought I liked the second paragraph, to be honest more than the other paragraphs because the place of the citation was okay. He just put a quotation there. I think it makes sense in terms of the connection between the previous idea. I felt like, "okay, he knows how to cite."

Actually... there was a kind of confusing sentence at the end of the second paragraph. I was like, "ok, he is aware of the organization" but I was just making sure with the specific details.

I just skipped the quotation. I didn't read the quotation because it was just a quote, so I thought "it is a cut and paste. It's not his own language."

I am reading the second topic sentence in third paragraph, second topic sentence. I read it a couple of times.

But it was very simple. Then I thought about the first topic sentence, which was again, this simple sentence structure. I was like, "in terms of sentence variety, there is not much variety there".

But still, that was another topic sentence to me. I felt like, "okay again, he is aware of the overall organization of an essay."

Now, I was just making sure with the specific details and connecting it to the main idea of that paragraph.

Again, he cited. He referred to the text and to me it was effective. He just attempted. It was not perfect, but it was good. I felt like, "okay he knows how to combine the necessarily information from the text into his essay,"

which is not something a B student. It affected me, to be honest.

The last paragraph is just one sentence. It was not okay.

But since it is just a summary sentence, I said, "okay, this is a finish so still, he knows how to finish an essay, although it is not a fully developed paragraph." It gives me an idea that he knows how to write an essay.

But the content made me think that this is not a pass definitely. I was more toward C/D.

I just looked at the rubric to make sure if it is B, or C/D. I am just comparing almost everything.

So, the arguments were not B definitely. They were mostly developed, so I was like, “okay, that's not B. It fits well with the C, D.”

I skipped the organization part because in B, it says, "somewhat organized," which is not the case for this essay.

I think I read the grammar part more. So, in B it says, "simple grammar structures," and, "attempting more complex." I was like moving back and forth from B to C/D in the grammar section.

I was confused at this point a little bit because there were a lot of simple sentence structures. The grammar at that point was really difficult for me to make the decision. There were not many errors. When I came to that point, I said “okay, there are not many grammar errors.” They were just simple sentence.

For the other parts, I could easily decide, but for grammar, it didn't fit well there. But it was more towards B because of the sentence structure. So, I said, “okay it's not a B, it is a kind of C/D. I think at that point I made sure that that's a C/D. I said, “okay, yeah, that's C/D.”

I think I just skimmed through the text to make sure this is a C/D essay.

I was actually looking at the grammar because I was confident with the argument and the organization. I was just trying to find any grammar errors to remember, to make sure that there were not many so that I can say, okay this is a C, D. I thought there were not many grammar errors.

They were just simple. They were basic sentences. I couldn't see any grammar error, except for that point, the run-on sentence there.

Again, I was just looking at the grammar again.

I was like, was it more B, or C, D? There were some attempts for complex sentences and I think I sat the first two parts with more than the grammar part. There was a run-on sentence there and there were a lot of simple sentences.

So in the end, I decided that this is a C/D essay.

Rater 7 – Essay 7

I just looked at the text directly. I just looked at the whole text, scroll down to see how many paragraphs are there. It looked fine. There were four, so it looked like an essay to me.

And then I read the first paragraph.

So basically, he is talking about one group who agrees and then another group who disagrees, and then he moves on to his own thesis.

So, I like that flow very much.

And in terms of the sentence level, the first paragraph sounded really great.

And then I was like "okay, this is gonna be a kind of pass student."

But I felt like I needed more because the word level and the flow was great, so I read the rest.

I read the thesis a couple of times, I guess, to make sure I got the idea ... And I'm just reading the thesis again, a couple of times.

So, the first sentence of the second paragraph was a little confusing to me in terms of ... Because I was, like, expecting what is this paragraph going to be about the advantages of GM food. But what is the topic here? Instead he is, like, referring to a news reporter, so it didn't sound right to me.

So, I felt like "okay it was a pass, but there was this problem. So, I needed to rest more, read more ..."

So again, there was no problem about the grammar.

At that point, I saw "rice has become a more serious problem" the second sentence, I was like, "Okay, is this the topic sentence here?" Yeah, so I felt like, okay, it doesn't have to be the first sentence all the time. And then it sounded right to me, and then I said, "Okay, so this is gonna about rice."

So, I said, "Okay, let's read more." Because if that's the case, then it would make sense to me.

So, I am just reading it again and again to make sure if it is a topic sentence or not.

So, I am just reading the rest. To connect it to that sentence, to rice, specifically.

Then, I saw the citation in this sentence I said, "Okay this is a citation?". "Just 50 grams of something, something." So that is from the source, so I said, "Okay this is good." And it's about rice, so it felt good because it is related to the topic that I think is the topic sentence.

So, I think at the end of this paragraph I said, "It's about, again, golden rice." And so I said "It makes sense."

And then, I just read the topic sentence again to see if this paragraph is okay.

I'm looking at the first sentence again to see, okay, this is a kind of an intro sentence and then the second sentence is a topic sentence, so I'm just making sure about the flow of the ideas. The first sentence is not the topic sentence, and the second sentence of the second paragraph is the topic sentence.

Now I was more towards pass.

So, in the third paragraph, I'm just making sure the example is related to the main idea, that I think is the main idea. But I am just trying to connect the idea to the topic of that paragraph. That paragraph felt okay.

So, it's about government strategies and everything.

I was not focusing on the grammar per se there, because the grammar was ... it sounded okay, but I found a couple of grammar errors that I think I looked at a couple of times.

So, this last paragraph, I was trying to make sure if it is the conclusion paragraph or if it is not another point ... another body paragraph. So I read the intro sentence again and then I said "okay this is the conclusion paragraph", and like the intro, it sounded good.

At this point, I couldn't say this is a pass. I couldn't say this is a CD either because he knows the overall essay organization.

Yeah, so I said "okay but I need to check the rubric because I wasn't really convinced that this is ... like confident that this is a pass."

I'm just skimming the essay again. So I am just skimming it again to make sure if I would change my mind.

As I was reading the second paragraph, I said, "am I pushing myself to think that this is the topic, or is this what the writer was thinking about?" It's still not clear to me.

But I was more towards pass. But I was like, "that is not a CD definitely because of the organization. He knows it. I don't think he would need any extra semester." That is what I was thinking. So, I was like "is this gonna be a CD plus or a pass?"

And then I decided to go to the rubric.

So, I am not looking at the B definitely. I'm looking at Pass and then I'm comparing it more with CD. Mostly, I'm looking at everything

except for the connections because the organization was okay. I just skipped that. Because the into and the conclusion was okay, so that's why I just skipped organization.

For the argument since the body paragraphs was the ones that struggled me a lot, I am reading that part.

So, did he skillfully integrate the sources? Yeah, for the sources, it was definitely not Pass because they were not mostly integrated skillfully. It was not like "not skillfully" as in CD. But it was not perfect either, so that's why I'm trying to see if it can be in CD plus or not.

For the grammar, it was okay, so I just didn't look at grammar much.

And I didn't look at convention either because there was nothing, like no spelling mistakes or anything.

I was confident with convention, grammar, and organization. Only the argument part was something I was not sure, so I am going back to text again to make sure.

I'm looking at the argument here in the essay again. I'm reading the thesis and just looking at the argument there.

So, I'm just skimming to see if the argument is okay.

So I've decided that this is a pass.

Rater 7 – Essay 8

I just looked at the text to see how many paragraphs are there. It was okay. It was four paragraphs.

The length seemed okay.

And then I started reading.

The sentence, I mean, the word choice and the grammar were great, to me.

So, I was like, "okay I'm reading a good essay," after I read the first sentence.

And then when I came to the thesis statement position, I was not sure this was a good essay because the thesis was not there, so there was no main idea there.

So, I felt like, "okay, I need to read the full opening paragraph."

But I'm like, the intro, at first, it felt good, but the thesis statement was not clear to me. So, basically, he's explaining what he means by independent students, which is not related to any argument, so I said, "there's nothing there."

So, then I moved on to the next paragraph to see if there are any points he is discussing. And then I read this, "I am from Dominican Republic,"

So, he was talking about himself. He was introducing him, which was not good, I mean in terms of organization, this is really problematic as this is not relevant to what he has been talking about.

And I said, "no, this is not a pass, definitely."

Okay, and then he's talking about GM food in general in that paragraph and the full in paragraph.

But he always introduces the paragraph, he talks about himself first, which is redundant, like not relevant too what he is discussing.

And also, he should kind of signal what is this paragraph about, which is missing there, so I didn't like the second paragraph.

And then, I moved on the third one, and then I saw that "I am a member of a church and..." blah, blah, blah...

I mean, he didn't relate this to any topic. So, I didn't see any purpose of having a separate paragraph for that information. So, I said, "he really has problems with argument development." There is no argument being built, so I didn't like the third paragraph either.

I just moved on to the fourth paragraph.

There was, again the sentence level was good. He was using different sentence structures.

And vocabulary is good.

But I was asking myself, "what is this paragraph about? Am I reading a conclusion paragraph now or did he move to another topic?" So, I was confused there. After I was not still sure because if this is a conclusion paragraph, what is that specific detail doing there. Why is there a specific detail, if it is not, what is this paragraph about and where is the conclusion paragraph? This told me that this person doesn't know organization.

Before I looked at the rubric, I decided that this not a pass, definitely. And, this can be a perfect C/D student because he definitely has a good command on grammar, and his vocabulary is good but he doesn't know how to organize an essay, which is a perfect student for C/D. However, he can be a good student for 101 B classes because he would learn how to organize a paragraph, which he doesn't know either.

So, I was like, "okay, he has perfect grammar, not perfect, he has very good grammar and vocabulary, but he doesn't know how to organize ideas.

So, I needed to look at the rating scale.

So, I was not looking at pass. I didn't look at CD organization much. I was more focusing on argument.

Again, I was like quite sure that he has problems with argument as well.

So, I was like, "how can I put this paper in terms of grammar?" His grammar was good.

I was looking at grammar and then I was looking at argument and organization. So, I thought, "okay, this is 30% and 25% of organization/argument, he's more like B and for grammar and connection he's more like C/D. So, I needed to decide which way to go.

I thought he has serious issues with organization and argument. But then I said, for the other part he's definitely not there. So, I was really, literally in the middle. So I said, "okay this is definitely B+ because I couldn't decide."

Rater 7 – Essay 9

I just looked at the text. So I just looked at the whole thing. I felt like it was missing something, maybe a concluding paragraph?

Then I just started reading it.

His grammar and his word choices is great.

The thesis statement is not perfect, but there is one and it might be fixed, but still I got the point so he is explicitly saying his thesis, so I like that.

But at this point, I didn't decide yet, but I was really sure that this is not going to be a B, definitely.

So, I'm just reading the thesis again. I'm reading the topic sentence. Reading the thesis.

The topic sentence sounded good. I like that.

And again, after reading the first part, since I felt like the grammar is okay.

So I'm not looking at the grammar because it didn't strike me at all and there is no grammar error almost

so I'm more focusing on how he builds his argument.

In the second paragraph, I'm just looking at the specific examples he is giving to support his point there, which I liked. And the examples sounded okay, but after reading the second paragraph I felt like he needed to support it more a little bit. He would have given more examples there.

The last paragraph, the topic sentence, again, I didn't see any problem with that except for that he was giving a lot of specific details there. Normally, we don't do it in the topic sentence, which I felt like is not good.

I read the rest of the paragraph

and the examples are better than the second paragraph.

So, in this paragraph, the example and the writing is okay, except for the topic sentence. In the second paragraph, though, the topic sentence was okay, but the examples were not well developed.

I mean he has a variety of sentence structure, so I like that.

So, in terms of grammar and vocabulary, definitely pass.

But in terms of organization, after reading this last paragraph, I said, "The conclusion is missing." So, a little bit problematic.

And the idea hasn't developed fully.

So in terms of organization and argument, more C/D. That was what I was thinking

So, I was like in between Pass and C/D. I was literally in between there.

Then, I looked at the thesis statement again. Just skimming again, because I was not satisfied, actually with the development. The idea development. I felt like I needed to read it again.

To be honest, I said, "This person doesn't need any C class, C or D classes." I mean he has the major sentence, he knows what he is doing in terms of how to write an essay.

The reason his conclusion paragraph is missing is not because he doesn't know how to put one there. He was just interrupted, so he couldn't submit the essay. Most probably so. If there were not, I'm pretty sure he would write one.

I decided to just ignore that problem.

So, in terms of organization, I said, "He's good."

I read the last paragraph very fast and then I focused on the second paragraph more. So I read it slowly to convince myself in terms of his argument and supporting sentence and everything.

Again, here I said, "He or she might include more examples here a little bit." If I had seen more details, I would definitely say this is a pass.

But I said "no". I think I said, "Not a pass."

After this moment, I said, "Okay I need to look at the rubric." I'm going to rubric.

I'm just reading pass and C/D. Not looking at B. Especially, I'm focusing on argument because I think he has problem with argument. I'm just reading to remember the expressions there. So, I'm moving back and forward from pass and C-D to make sure that he fits somewhere there.

And I looked at the grammar in C/D because it says, "Some complex ones, some grammar errors might occasionally interfere comprehensibility"

which is not the case. So, this is not a C-D definitely.

I'm again going back to argument because that part really ... the problematic one. I'm reading it again and again.

And then I think I said, "okay this is in between definitely ... definitely not a C-D." Because he has problems with argument, he fits there in C-D.

To me, argument is a little bit more important than grammar

So I couldn't say pass, I said C/D+.

Rater 7 – Essay 10

So, there were four paragraphs, which was good.

I just started reading.

The second sentence sounded ... It was weird.

And then, because it was too long

I just read it a couple of times to make sure that it is not a run-on sentence, to make sure that his grammar's okay.

So, and then I said, "Okay, this is good. Sure, he is good at the sentence level, if he attempts to do something." So, he's not relying on one simple sentence structure, so I like that.

And in terms of thesis statement, I don't like this kind of thesis statements. So this, I say we will talk about kind of thing, but there's no argument there. There's just a kind of announcement that he's gonna discuss something. But he doesn't say it exactly, so I don't like that kind of thing. I couldn't understand what his main idea about by looking at the thesis.

I think here, I'm just looking at that sentence again. The second sentence, the long sentence to kind of sense something in terms of his argument. I'm just reading it again.

So, I couldn't feel anything of the kind of story. I said, "Okay, I need to read the body paragraphs." So, since the introduction didn't give me anything, I said the body paragraphs are important. So I said, "Okay, I should move on."

So, I moved on to the body paragraphs.

I thought he has a good command on grammar and vocabulary.

So I decided not to look at the sentence levels but look at arguments.

The topic sentence was not clear there. It was too long, so I couldn't make and decide about argument. I couldn't even decide if he is arguing something.

At first, it feels like he's just talking about some facts, coming from the text. His thesis says he's summarizing, so I was looking for his argument, his own opinions about something. So, I couldn't find it in the first topic sentence, and I'm not sure if I found it in the second paragraph at all. I felt like he's just summarizing.

Yeah. So, when I finished the second paragraph, I was not convinced again in terms of the argument. Since the idea that open is problematic, I was really in between.

The third paragraph, especially the topic sentence, I read it a couple of times to make sure that it includes some idea.

So, he's defining the term (GMO) in topic sentence, which we don't do in topic sentences.

And he's citing from the text, but I didn't see any idea that comes from there either. I mean, the student himself or herself. They're all just paraphrased versions of the text, so there's no argument of their own there. So, all this information there is coming from texts. So, what is the purpose of putting it, if you don't relate it to anything? So, I didn't like the third paragraph, because he's totally summarizing the source texts.

And the last paragraph, I like that paragraph in terms of grammar.

This paragraph has a function there. He basically kind of summarized everything. So I felt like, "okay he knows how to write an essay."

And he has a summary sentence in conclusion paragraph, which I like.

So, I think at that point, I was not still convinced that this is a Pass. This is not a B definitely. I could put a CD, if there not CD plus in this grading scale, definitely.

So, I said I should read the body paragraphs again.

I'm reading those two body paragraphs again, a couple of times to make sure this is not a B or Pass, to see which way to go.

Now I'm just between CD and CD plus.

I focused on CD mostly and especially in terms of argument and I looked at Pass because there's nothing in CD plus. So, I'm looking at it to find a way between these two grades.

I was looking at organization just to make sure he is not a CD.

His organization was okay to me. So, for example, like "simple transition word" and everything. I think he has more than that. In terms of paragraph unity, each paragraph had distinct in itself. There was no redundancy definitely, but the whole essay might be a little bit ... Yeah, if you look at the whole essay, he has not any problem with organization to me. But at the thesis statement can be fixed. It's not a big deal.

I think I have decided. Yeah, I think ... I decided that this is a CD plus.

I was just looking at the argument. I'm trying to find the flaw actually, "does he have a purpose of writing that second paragraph" and "does he have a purpose for moving to the third paragraph?"

Why is this a separate paragraph? Then I thought, "yeah, that's different. That paragraph to me sounded different than the third paragraph". So, I said, "Okay he has a purpose of separating those two paragraphs." So, I said, "Okay, he knows what he's doing."

RATER 8 (09/29/2018)

Rater 8 – Essay 1

Yeah, first of all, I had to review the rubrics and descriptors for the bench-marking criteria, because the source should be integrated into the writing.

So, I decided to pay close attention to whether the source materials are integrated, or they are paraphrased well into the essay.

So, I think the source material should support the examinee's arguments. So, I had to just look at the descriptors for the conventions, and also the argument's detail, and organizations. The grammar and the lexis things, because it just stand out.

So, if I just look through the entire essay, I can see, easily, whether the examinees are good or bad in the complex grammar structure on vocab. It really stands out. I mean, in terms of the grammar and lexis.

So, I just wanted to play close attention to the source materials and whether it is well paraphrased, and whether it supports the arguments and details and how it can fit into the entire organization. So, that's my focus of my evaluation.

I just read the instructions for the task 2, because the conditions to be met for the task 2, is that examinees should provide four paragraph long essay for the task 2.

The reason why I am looking at the pass descriptor for the organization is that, because the instructions for the task is like that one. And also, I just want to see if there is any mentions about

the existence or presence or the absence of thesis statements or topic sentences. But I don't see the existence of the topics sentences and thesis statements.

So, but still I think that is really important aspects of the organization. Of the essay. So that's what I was thinking about when I was looking at the descriptor for the Pass. For the marking criteria organization. Yeah.

Yeah, I just look through the pass column across the marking criteria because I just wanted to see how I can just differentiate between the CD descriptors and the pass descriptors. Because that is really important borderlines, right? Because these is really high stake decisions. Because the examinees, I think that it doesn't matter between the B or ... The less important than the B and CD borderlines. Because the borderline between CD and pass is really important for the examinees. That would make the life in the first semester for the examinees change.

So, I just wanted to see how I can just differentiate between the CD essays and the pass essays. And seems to me that the big difference between them lies in the grammar structures and the modifiers. Because the bold phrase descriptors stand out. So they grab my attention. So could control of a simple and some complex ones, and I see the past over here, but I do not see the ...

I was expecting to see the many grammar, and complex grammar structures any many complex ones over here. But somehow I do have a three items for the CD for the grammar and lexis. But I do have just one item for the pass. So I was already confused at the time. Because we do not have matching numbers between the C and D and the pass. So I just want to deal with and address this kinds of address. But it is obvious to me. Because it says it is a little bit of complex ones. Well not too many. And if I just run in to some grammatical and lexical errors frequently, more than often, then that might indicate, okay this essay should be classified as a CD or something like that. That's why I was just focusing on-

8:Oh, yeah. The first sentence just indicates the topic of the essay. So the genetically modify the food grasp my attention. And I realized it.

Okay, the topic for this essay is the same as the topics for the essays that they are rated about three days ago, or four days ago. And okay, so I just said to me, thought to myself that I am familiar with the topics itself.

And then, I read the first one and the second.

And then I was looking for the source materials actually.

And in the first paragraph, it says a little bit weird to me because the second and the most of the sentences start the same. Start with the same subjects. And the first one is "genetically modified food". The second one is "some". The next next one is "some". And the next, next, next one is a "some".

So I see that I don't have a thesis statement.

That means... because I was working as the instructor for the 101C,

if I do not see the thesis statements, this means it wouldn't meet the requirements for the major assignments. And that's what I thought to myself at the time.

And then, I was just looking over the second paragraphs

to look for how the source materials are integrated to support the examinee's topic sentences.

And I just, there is the authors of the newspapers over there, [Lomberg 00:07:58]. And then, I was looking for the other author's name through the second paragraph, but I couldn't see that. And also, it's weird to me but the most of the sentences start with the same subject. Genetically modified food, some, and then bananas and bananas again. And the third paragraph start with genetically modified food again. So I see that's the evidence for the lack of control over the complex grammar structure of the examinees.

Oh, accident actually. Because I was just go, I was meant to just to go to read the source material, yeah.

Yeah, basically I was reading through the source text because the examinees mention the Golden Rice.

I thought okay, that is the evidence for the advantages of the genetically modified food. But I was not sure about that. I just wanted to see whether the examinees selected the right material in the right place. Because I need evidence, okay? That's why I was looking for the information upon the Golden Rice. I couldn't see it here.

So that's why I was just moving on to the next source materials, which is written by these people. I was just making sure that...

Because this is the really important aspects, because the only evidence that I got to decide whether the examinees incorporated the source materials to his writing in the right manner. Because the success lies in whether the examinee is paraphrased the source materials well.

So I need to just make sure, I need to just make comparisons between the source material materials and to the paraphrased ones in the examinees essays.

That's why I spend some time on reading through the entire source text, not entire, but about this area.

Yeah. Because I read already the newspaper, so skimmed through the newspaper, I think would be enough to refresh my memories about what the Golden Rice is.

Okay, and how can the existence of Golden Rice support the proponents of the GMO, and that's what I was looking for... Looking for the Golden Rice. And I found it over there.

And then, I just wanted to have more information about what Golden Rice is, and so I just look through the entire paragraphs. Yeah. Yeah, I just read through every paragraph. And every details of it. Because this is the only source materials were incorporated in the examinees essay. Yeah.

So now back to the essay again.

This is the part [second paragraph] where the source materials are incorporated into this one. So I think, okay, that's fair enough. So, this is about the Golden Rice.

“Going to provide enough vitamin A to poor people in India with vitamin A deficiency. And, it is a lot more affordable than the other programs in terms of the cost”.

But, I think that is not enough, actually. And that is not enough.

And some of the information is not correct. I mean, somehow, it's correct. Not entirely accurate.

And also, I was not sure that the original source was paraphrased well enough to convey his ideas to the audience.

So, and I thought to myself that okay, where is the other source material. Alright? I don't see any evidence of the whether the second source materials. I mean the other ones. The other ones was incorporated in examinees essays or not.

So, at that time I decided that, alright, this is not pass. Okay?

And, when I had a look at the last one, briefly over here. Okay, it's stopped over there, "So much power". And that's it. It's not a complete essay. I don't see anything over there.

So, I guess, I'm sure that I went back to the previous, yeah over here. So I had a look at the rubric organizations. And I had a look at the okay, "lack of focus, and unclear". And I look over here [column CD], and "adequately organized"

because I already decided about the convention parts, over here. Because the second source materials was not incorporated, reflected in the examinees essay at all. So I already decided in these parts, and also I decided in these parts too [grammar and lexis], because all the sentences start with the same subjects. Either genetically modified food or some and bananas. And that's it. Okay. It's really monotonous. I don't see any variety of the structures over there. So I decided in these parts [convention] and these parts [grammar and lexis] together. So that's why I was just focusing and looking around this one [organization and arguments].

Because it's really only two paragraphs.

I don't see any clear arguments over there because I don't see any thesis statements.

So that's why I decided okay, this essay is B-

Rater 8– Essay 2

This is the same rubric and the rating scale. So, okay, I am good with it. So I just ... Let me just read the essay. That is my thinking at the same time.

And then I read through your first paragraphs.

And then I found out, okay, there is a sentence looks like a thesis statement at the end of the introduction section. That is a good sign.

And then, at the time, I just decided to go back to the previous one the rubric.

So, okay, in terms of the organization and arguments, I think it's better than the previous one (essay).

I thought to myself that, okay, this examinee was somehow capable of organizing his thinking at the end of the introduction paragraph. And, okay, that is a good sign.

So whether ... because it is not perfect, but at least this essay has a potential to be classified C or D.

Then I just had a look at the second paragraphs.

Okay, I see the ... Okay, this is familiar stuff because I had a look at source text 2, and then look through the ones over here, the author text 2 and the information about golden rice. The second paragraph looks familiar to me.

All right. And I especially had a look at the "compare" over here. I think that that's the grammatical structures, and then that is the evidence for the lack of control over this complex grammar structure.

And I had a look at the author of the text 2 (in paragraph 2). It's about the golden rice again. The newspapers talk about the golden rice, and it says something about the cost-effective approach to saving people from vitamin A deficiency. But, again, I'm not sure how those... I mean, that's okay with the supporting details for the other advantages of a genetically modified food.

But somehow, some of the steps are not satisfactorily paraphrased to be a good supporting detail for the first argument of the examinees.

And also, the two things grab my attention, especially for the "compare-to" and the "saved **form** vitamin A deficiency".

And also, the last one, "which means all 500,000..." because they should have some comma things, okay? It could be a punctuation problem or the examinees might have had some sentence boundaries issues as well

because when I was teaching 101C, a lot of students are having sentence boundary issues such as run-on sentence or the comma splice.

So whenever I run into the essays with the sentence boundary issues, I wouldn't give him or her the pass grade because they need some help by taking the 101C or other courses.

So that means, okay, thanks to the "compare" and the "save form" and the "which means", okay, in terms of the grammar and the lexis category, okay, he or she wouldn't get a pass grade for that category.

And then I was expecting to see the counterarguments to the first one in the last paragraph because he says "on the other hand." That is a logical connector.

And then "it is true that", okay, that is good structure, and that is good phrase. And "GM industry is controlled".

HK3: But I was really, was not sure about "there are chances that it will get better over time". What does "it" refer to in this context? It's really ambiguous.

When I was reading a newspaper, the author did not mention anything about some bright future of the situations. As far as I remember, he didn't mention anything about "it will get better over time." So, okay, is this kind of weird?

And it says something different from the previous and the right next sentence. So somehow, this sentence "However, there are chance that it will get better over time." Logically, it doesn't fit into entire paragraph.

I was looking for the information whether the examinees, just he came up with his or her own ideas instead of the paraphrasing the original sources. In that case, do I need to penalize him or not? That's what I'm looking for, okay? Because I think, at that time, I was half sure that that idea does not belong to the original authors but seems to belong to the author's idea. So I was looking for the descriptors, whether the fake source material. But I do not see any descriptor for the ... How would I say? Off-topic materials or extra materials.

But the examinees pretend as if he just recited materials from the source materials, but actually, it was not actually. It was from the author's own ideas, right?

Yeah. In the last part of the essay. I just read through the last two ones again because I think there should be something after this one, because I think I read this one (paragraph) again, especially about the "There are chance that it will get better over time."

So I was thinking to myself that, okay, this sentence does not fit into this paragraph.

And then I was sure that "There are chance that it will get better over time" might not belong to the author's original ideas.

So I just move on to the original authors- The first one. And I read through the entire paragraph, looking for the comment about the bright future, the bright future of, "things will get better soon." But I did not find any information about that. At least that's my impression.

So after reading through the entire newspaper, I just decided that, all right, that sentence does not belong to the Tom Chiver's idea. I mean the newspaper's. But it seems like, to me, that the examinees came up with his or her own ideas.

Yeah. I think ... Yeah. This one around here. So, because I felt myself that there is evidence of the integral citations "as the author of text two" ...

And I think this should have been text one. It should be "the author from text one" or something like that.

And I do see some types of the reporting verbs, such as "mentions" and such like that.

So I was decide, okay the capability of examinees to make citations, I think the citation materials are sometimes ... I mean, they're somewhat mixed with this own ideas. But I got the impression that the examinees just pretends as if all the materials are actually paraphrased from the original author. That is really like ... To me, it's like evidence of violating the plagiarism, committing of plagiarism issues, okay? It's incorrectly cited original author's ideas. And then ... Yeah. So I think, okay, this is not good.

Then, I looked at the introduction again.

because according to the thesis statement, "it is cost effective", "because it will get better over time". And I do the same phrase in the last paragraph, "It will get over time." So I think it's like I ... It's not a original author's arguments but the examinee's arguments about the benefits of GMO.

Rater 8 – Essay 3

So my first impression of this is "Wow, this examinee wrote a lot".

And I wanted to see the first paragraphs.

Okay, what grabs my attention is the "orater". I think this should have been R-T-E-R, not the T-E-R.

This examinee has a rich vocabulary but has some spelling errors. That's my first impression.

And then, after that, "which stable or additional nutrients", "which" modifies over here. What is that? That is not right.

I don't understand, this entire structure of the first sentence interferes my understanding of what he or she is talking about.

Then I had to look at the second sentence again.

This is serious sentence boundaries she is over doing. That is a run on sentence.

"These crops are not being produced in natural ways, indeed, they are being produced by scientific way".

"Scienteeffic way". A misspelling error.

And the run-on sentence.

and I do see the colons over there. So usually a colon after I was expecting to see some kinds of list or controlling ideas actually.

What does mean the "mating", what does it mean, "meting" or "mating"? And "chaging the code of DNA"?

It's not clear to me.

And "I don't think generally modified crops should be supported because there ary are lots of side effects".

So, I realize the examinee attitude toward the GM is negative, because I do see the side effects.

Overall, the first paragraph I do see a lot of grammar issues, and spelling errors, and sentence structure issues although he or she attempts to try the complex grammar structures.

Then I had to look at the "first of all" and read through the entire text. "First of all, the GM crops is being produced be chaging their DNA code to increased number of crops".

And the first sentence itself does not make sense to me.

"Actually those genetically changes will effect".

Okay, I think "effect" should be changed into "affect", I mean the A words.

"And its will remian the toxiness in our bodies".

At that time I realized that in the first arguments, as long as we got in the first paragraph, it's entirely the examinee's ideas, because I don't see any source materials over there. Alright?

"There is an experiment..." and also, I see the relative clause problems over here. And this is sentence fragments.

"the GM has been affected the pregnancy mother and the side effect has been affected the infant".

I notice here the long grammar structure, which is evidence that the examinee cannot have master over the passive voice sentence, or something like that.

And "secondly, CM crops are highly tolerant to herbicides".

I do not see any mention about the herbicides in the two source materials.

"This can make them not to easily being damaged by the natural disaster incident".

This is the wrong grammar structures again.

This topic sentence does not make sense to me either.

And "this will lea"; so "this" is like an unsupported summary now. So I am not sure what this differs back to the previous sentence. It is really quite unclear.

"Hurting the environmental", another misspelling.

"because farmer need to use herbicides again"

I don't see any mention of the herbicides again in the source text?

The examinee didn't bother to mention any citations either.

And again I see "this is very harmful". Again, unsupported "this".

Then, "plant wil losing thier places", and this is the misspelling again.

All those things together interfere with my understanding of the essay.

I know I would need to look at the next paragraph for information.

"GM crops will hurting".

This is not good. "Were hurting", this is impossible structure in English.

"GM crops are usually conducted by those highly technology countries because needs lots of experiments".

The entire third body paragraph does not incorporate the original authors' ideas.

It mentions something about advanced countries or developed country, "because it needs a lot of man power or money to conduct some experiments", but I do not see any kinds of information from the source texts about that.

Because I wasn't sure; that is why I did not just move onto the source materials.

I see a lot of "can'y", this is another misspelling,

and another evidence of this unsupported reference "they".

And it's very hard for me to follow his words in the sentence.

Even for the last logical connector, "conclusion". That should be indicative of the last paragraphs. It is wrong, right? It should be "in conclusion".

And here "GM crops are not supposed to be suppoeted because of those side effect impacts lots of human beings and countries".

It is another sentence fragment, right? Here in "Even they have benefits in other ways."

"You have freedom to make decision and your health is on your hand and for your future generations".

I do not think the last comments is fitting to his entire argument.

That is why I did not bother to look at the source materials. At the time, because I needed to give some grade to this examinee, so I decided to go back to the rating rubric.

I was looking at the grammar stuff over here. I look at this one, conventions.

Because I see a lot of spelling errors. I do see some misspellings in here, because misspelling frequently interfere with my comprehensive ability.

So conventions should be B, okay?

And then, grammar and lexis, I already made up my mind about the grammar and lexis, this should be B.

But then later on, I decided to make comparison between the descriptors B and C and D for convention and this area too. I went to the organization argument detail

because I do see the obvious structure of the essay. I do see the thesis statements and 2 paragraphs. And the concluding paragraphs as well and I see a lot of logical connectors. So I see some evidence of the capability of organizing stuff, so that he or she could enhance the coherence, cohesion in some way.

So, the grammar and lexis and conventions, it's a decent essay. So organization arguments stuff, he or she might have some basic knowledge of organization structure. I just reconcile the, synthesize the grammar and lexis things and convention things and organization arguments detail.

I just made up my mind. This should be B+

Rater 8 – Essay 4

In the first phase, I was confused because I was expecting to see the four distinct paragraphs, but I do see only two paragraphs over there.

So, that is indicative of the lack of organization.

And I was looking for the sentence boundary between the topic sentences and the previous paragraph.

Then I decided to look at the first sentence.

Yeah. What grabs my attention is that, all right, “progress is always good as the population of the world is increasing exponentially”. Okay, that is a really good word.

And also, I just see the comma splice over there, two sentences are conjoined without putting any proper punctuations. We do see only comma for there, so I was expecting to see some conjunctions over there, but I do not see any. So that is another evidence of a sentence boundary issue.

"I think we must support the research for GM crops and plants".

Okay, I think it looks like a thesis statement to me, but right after that...

"as we can see in paragraphs by Tom Chivers, the corporate companies have gained almost full control over this technology, and that is detrimental" ...

Okay, that is a good word on top of the "exponentially", "detrimental", so that is quite unexpected for this level of language learners, okay.

And "welfare of the common people and the world as a whole".

And I followed the materials cited here, I think that's the second one, the newspapers by Chivers. "And this will make many people suffer". I think basically the source materials are somehow incorporated to support the ideas. I mean, support the arguments but somehow it's really contrary to the thesis statements, right?

Because according to the examinee's arguments, we should support the research for GM crops, right? But he just here explained a lot about these are the advantages, potentially the side effects of the GMO for this much.

So, yes, that looks really weird to me.

Okay, maybe you could say the examinees might be running out of time so that he or she might just fail to address the balance between the paragraphs in the essay.

So I think he or she should pay more attention to the second paragraphs over here because this paragraph is about supporting the thesis arguments.

So, I read through the entire parts over here and then that move on to the next parts, and now I see the Lormberg's newspaper in the second paragraphs, and then, yeah, it's about the use of the GMO, right?

I went back to the thesis. I just compared this one (the first sentence last paragraph) and the thesis because this should be the direct supporting ideas for the examinees' thesis statement in the introduction.

Yeah, but I was wondering about the studies over there (last paragraph). "Studies show that" and "the studies show that". Where do they come from? Because "studies" means a lot of studies, right? But I do see only two source materials, right? So where do they come from? I was wondering about that. So, because "the studies", but I do not see the authors of the studies and so on.

That's why I just move on to the source materials over there. Yeah, I do not see any things over here

So, I looked at text two. I just read through to decide where "studies" came from.

So I just read through the entire parts over there and there and I just read into the studies over here. "Two recent studies in the American Journal of Clinical ..."

Okay, I realized, "Okay, the examinee just refers to these two recent studies in her or his essay. Okay, I got it, I got it".

But somehow, okay, because it's the indirect citations, right? Again, because within the source materials, the original author made the citations of the external source again, so it's like embedded citations, right? So, it's really hard even for us to just make the embedded citations, right? So that's why I was confused, right? That the two studies within the source materials and the two studies are actually embedded source materials.

But yes, somehow the examinees attempted to make citations of the embedded sources, but I'm not sure whether that is done in the correct way.

So, I looked at the essay again after source text two.

Yeah. Just to make sure that these are the two studies that this student cited in the essay.

And then it says, "As we can see ...", "I think developments of GM crops should be supported".

So the last sentence aligns with examinee's thesis statements in the very beginning of the paragraphs. Okay, that's a good sign.

I looked at text two again. It says here "50 grams of golden rice by 60% of the recommended a daily intake of vitamin A".

So, that matches with the information in last paragraph of the essay.

I was trying to just make sure that, okay, these parts belong to the original author's idea. But since these parts belongs to these two studies, maybe the examinee should have made a different set of citations to make citations of the embedded citations in the source materials because these parts belongs to the two studies published in the American Council or something like that, not the newspapers. Yeah, so that's what I was looking for.

I was looking through the middle columns on the rubric over there intuitively.

Okay, this should be C/D one.

Because although I do not see the clear distinctions between the paragraphs, but I do see some hidden boundaries between the paragraphs, right? It's just the convention things, okay?

Also, although I do see some kinds of sentence boundary issues. Okay, he got a good control of simple grammar structures and some complex ones.

So grammar and lexis should be in C/D.

Also, although I do not see the distinct boundaries between the paragraphs, but I do see the clear arguments, see the evidence of thesis statements. Over the entire essay it's adequately organized. I think that it is easy to follow.

And also, he or she might need from instruction.

Okay, I just decided to give a C or D to the examinees.

Rater 8 – Essay 5

So, I had a look at the first paragraph to look for the thesis statements again, and then, I think that there is, yeah, thesis statements.

"As far as I'm concerned, GM technologies does do a bunch more benefit right now".

It sounds a little bit strange to me, but it's okay.

I was looking at the first sentence and I think there's another comma splice

and I also was really, find it hard to understand the first sentence because "thanks to the [integers] of cheaper food, supply companies might finally", okay, it's the wrong logical connectors, okay. It wouldn't connect the previous parts and the later parts.

"Other sort of view that GM technology would do benefit, for those who cannot afford the food".

Did I see summation about the food? I was not sure about that, okay?

And then I had a look at the second paragraph, it says a lot.

So, just as what the second article said, it's about the Golden rice. And many people cannot afford food again. But the examinee, on the examinee somehow kept mentioning the food, again and again.

And also I see the integration of GM technology and this is not a correct logical connector. "And GM technology is much cheaper because, it grows faster, purchase more for [ends clear]". What is that? I was really confused over that one. That is to say, okay, this is not a logical connector.

"People who could not afford the food before could eat more and have more change to survive".

Alright, let's see. Did I see these kinds of things in the second newspaper, I was not sure at the time.

Okay, so let me just keep reading (the essay).

"Though many people nowadays needn't worry about this question, we can not forget 10% of global human by hunger".

"10% of a global population", okay, I think, I said to myself, I see some kinds of the same figure in the second newspaper (source text). But, I'm not sure whether it is precisely the same as the source material at the time. In the second paragraph, it's about the hunger and the food.

And "the GM food could provide the requirement for food nutrition such as vitamin A, with a cheap price".

The examinee just focused upon the benefits of GM food because it's cheap price. And just focus upon the hunger problems. So, that's why I was just wondering about that one.

And I just move on to the next one (the last paragraph),

and then I see it again, again. The cheaper food. Yeah, so, because I do see the same food mentioned over again in the second paragraph, okay, cheaper food in the last paragraph, and then I do not see any... because it's basically the same materials, the same materials as the second paragraph.

So, where do I see the first one? Yeah, I see, "the companies who cannot use the GM technology will break down, means go bankrupt". Okay, "going bankruptcy".

I do not see any mention about the bankruptcy, okay, because it means, that companies who are not capable of harnessing GM technology are rightfully to go bankrupt. That's my understanding. So I didn't see any kind of that information in the first source material.

So that's why I just go to this one, text two, text two, and I just read it through the food mentions, okay. Because it's about the vitamin A deficiency. This entire paragraph is not about the hunger, alright, yeah. The entire paragraph is about the vitamin A deficiency or do I see some "creating hunger and malnutrition", but that does not have a direct relationship between the golden rice over there. So, yeah ...

Yeah, essay again.

Okay, so I do not see anything over here, I mean the food things in the origin of source. And I decided, okay, although the examinee attempted to make a citation of the origin of source, but the order of supporting ideas are kind of mixed with his own ideas. Okay, because, alright, the examinee just picked up some word, which is "hunger", which happens to be hunger, and he just make an argument, or he or she just make an argument, that okay, the GM food would be beneficial to people because it would save us from hunger problem.

But actually, specifically, the second paragraph is about the vitamin A deficiency. Not specifically about the hunger.

Yeah, to make sure that the entire paragraph is about the vitamin A deficiency, not about the hunger especially.

because the second paragraph in the examinee says is about the hunger problem, and I just make a comparison.

And also I had to look at this one. "Companies who cannot use the GM technology will break down". "Because the GM technology is a high tech nowadays which means most of common companies cannot use it".

I do not see any kinds of information about that, okay. Yeah, of course the companies with the GM technologies might control over the entire food chain. That is the concern of the first author, but I do not see any kinds of information those companies who cannot implement GM technology will lag behind the other companies, or might just go bankrupt. And this kind of information. So, okay, this one, although the examinees pretend that this one is from the source material, not it's not actually. It's the examinees own ideas. Or their imagination.

Yeah, so that's why I just look at the, come back to the the EPT writing scale especially I had to look at the argument details, looking for a citation in the arguments over there.

So "arguments are vague and underdeveloped". Yeah, I think this is somewhat developed okay, because I do see a clear thesis statements.

And also I had to look at the organization. Okay, I think this is somewhat adequately organized, or something like that.

But when I had to look at the conventions over there, "sources might not be cited".

Okay, although the examinees mention the author's name, but it does not match with the original materials. What the examinees talked about in his essays, did not match with were the original materials in the source materials.

So I think, okay, that's the evidence that the examinees might not be capable of the reading, and picking up the gist of the text, and incorporating them into his or her own essays with success. He or she cannot do that. So, that's why I think, okay, the examinee should understand this way, although the organization and arguments are in this way.

So I decided, okay, this is the source based material, source based writing, so I need to be, he or she should demonstrate his or her skills in incorporating the source material somehow.

But, I think this is a clear failure, right here. So this is why I just decided to D+ to this essay. But D+, really close to a C.

Rater 8 – Essay 6

I started right away reading the essay, and I looked through the first paragraphs

And what grabs my attention is that ... I see a lot of "genetically modified food" again.

And I see a "scientist" over there. Usually I do see the thesis statements after the "however", but it says, "I am not a scientist" .

And also, another comma splice, "I also have my own opinion for the genetically modified food," okay?

I think that kinds of sentence, I mean, quite redundant actually.

And also, it is expected to see the thesis statements right after the "however"
but I do not see the thesis statements after the "however"

And quite redundant sentences are there.

And in the last part of introduction section, okay, it's like... Oh, there was a giant dot over here
"should not be eaten".

Okay. So it's like a opposition to the GMO. That is here position of the examinee. Okay, I got it.
So and then I move onto your next paragraph (paragraph 2), "first of all... farmers' benefits... is
give" ...

There is the lack of control over the verb over there.

"And using from other species" or "it does", or so there is subject-verb agreement problems
again - "It does not change", okay?

And "it do not change and give them new property and so it is very easy to add the gene in one
species to improve species production".

Okay. It's obvious that, that kind of ideas are directly from the examinee, right?

"In the text one, the author mention". Again, I do see another kind of subject-verb agreement
problem, the subject verb agreement.

And the most problematic parts in the second paragraph is, I do see the quotations mark over
there. I don't like that error because the original material should be paraphrased somehow, right?
And so that is the wrong conventions, if you want to just makes direct quotation, then you need
to know how to properly give credit to the original authors, right, by identifying the author's
name, publication year, and the page number, okay. It just makes me uncomfortable with the
entire paragraph.

"So in the African people we gain less money is developed so widely".

So where does the African people come from? So, It's really halfheartedly created ideas all over
the place in these second paragraph.

(paragraph 3) "Second, genetically modified food will influence the health of human, we can
actually know what will happen". "So as to text two", okay. Text two is, you know, the second
newspapers, right?

So, "golden rice have more vitamin A than the normal rice. So if human get too much vitamin, it
will cause to blind or dead".

That is entirely wrong information, "if human cannot get enough sufficient vitamin A, that might
cause blindness or death". Quite opposite situations, right? That means, the examinees didn't
understand text at all.

I mean just it could be the situation that the examinee just looking through the text to and pick up
some superficial information. And then treat it as if that is true.

So, but I was just wanting to make sure. But anyway I just wanted to keep you reading the entire
paragraph to the end.

Alright, "and I think according to my two reasons", okay, I think that is not a proper logical entrance for the concluding paragraphs.

So again, "I don't think GM food I a good thing for human. Too simple words.

"It would influence us on many ways". And that is a really unclear mention again.

It wouldn't summarize these main ideas of the entire paragraphs either.

And I just got to the second text about the vitamin C, how it is associated with the blindness and the death. That's why I just look through the entire paragraphs, and over and over again, looking for some connection between the lack of vitamin A and the blindness over there.

And I see that, okay, alright, so because of the vitamin A deficiency and not the too much intake of the vitamin A, but the vitamin A deficiency would cause blindness and death, this is quite the opposite (with the idea in the essay), right? So okay, so, that is the evidence that the examinee did understand what the main idea or the details of the text two. Okay.

Alright so, I just went back to the essay over there, okay.

Yeah, to check whether the paraphrase materials align with the materials in the original source.

I went back to text two.

Yeah, again, just to make sure what is in there, so what is the connection between the blindness and the death and the vitamin A deficiency or something like that.

And I just kept thinking about that connection between them and I concluded that okay, too much vitamin A wouldn't cause the blindness or death. And that's what I'm thinking about at the time, yeah.

So I just went back to the essay again. I don't know why I went back to the essay again.

And maybe I was thinking about the direct quotations again to make sure the quality of the essay before finalizing my decision.

So at the time I was thinking, this one shouldn't be C or D , something like that, but I need to justify my decisions, okay.

So I went back to the rating scale over there.

Yes, I did. So in here, I just was looking at the arguments and details or argument and details which is the obvious.

And saw how you had to look at the grammar and lexis because I need to just decide on how the sentence boundary issues affect the grade that I would give to this essay and the conventions.

I also had to look at the sources, sources might not cited, not paraphrased at all because there are direct quotation marks and the first five paragraphs again.

And but in terms of organization, I mean I just see kinds of organization stuff, okay because I see the distinct paragraphs, and I just see the topics sentence for each sentence, I mean for each paragraph again.

But because this is really, so it's source-based writing, so I think that is not acceptable...in my opinion, okay. Direct quotations, not paraphrasing.

I was looking at the third paragraphs again. Why did I...

Oh yeah, to just finalize my decisions okay because this does not match with the original stuff
And this one is direct quotation, okay.

I just define my decisions about the quality of this essay. B plus.

Rater 8 – Essay 7

I decided to read the first paragraph. I just followed the train of thoughts of the examinees
and I thought, "Okay, the first sentence is good".

And I do see the clear contrast between the second and the third sentences, "some people think",
"but for the others, they think." Okay, there's good contrast between them.

And also, when I was having a look at the last sentence as a personal thought about GM food, "I
think GMO products should be highly supported and be widespread on earth". Okay, I like this.

And then, I had a look at the second paragraph.

"The news reporter". Okay, which news reporter? So the examinee didn't identify the source of
his arguments, right?

"10% percent of that, it was as a vitamin A deficiency, and also it can cause many children to go
blind in their early age."

So that is the wrong grammar structure.

And "So rice has become is a serious problem that people were facing."

Okay, I was really hard time to understand this structure,

So, I think that that is a sentence fragment or something like that.

So "people were facing, for who eat rice as their daily food. But now the genetically modified
food has being developed".

Why is that? The examinee should explain why.

So now I see the source of the original, the citation, okay. "The American Journal of Clinical
Nutrition has showed that". But that is also the embedded citation in the second one. I'm not sure
if this is how he or she should cite it.

"Just 50 grams of golden rice can provide 60%..."

I do see again the direct quotation, directly from the second newspaper. I do not like this one
because that means somehow the examinees cannot paraphrase the original source.

"So it means the problem about the vitamin A deficiency has been fixed"... "a little bit of golden
rice, then they will have the recommended daily intake of what they need."

"Intake", okay, that's a good word, okay.

And then I moved on to the next one. "For some biotech companies, they do not have the power
to develop genetically modified food in some poor countries."

Okay, because it mentions the biotech companies, obviously the next paragraph talks about the
source material in the first newspaper which was written by Tom Chivers. But the paraphrased
information is not correct. That is incorrect information.

So I knew that at the time, but later on, I would need to just make sure that that information was correct or not. But I just decided to keep reading until the end of the essay.

And then I just read through the essay. "If they do not have enough money to do..."

And I decided that at the time that most of the materials are not precisely correct.

So also, "it may sounds a little bit difficult to them, but to care about the people's life, I think this plant is needed for them."

All right, that is good, the passive.

"It is helpful for poor people also it is a better way to stop the high rate of vitamin A deficiency death.

That is the also another run-on sentence, because improper run-on sentence.

So "At last, I think people should really care about their health. Healthy is the most important thing for family".

That is a word form problem. "Healthy".

And now the end, "the GM product has been developed, people should use this chance to rebuild their countries and help the poor get enough vitamin A intake as they needed. Genetically modified food is a great movement for all of the countries."

Okay, so I am done with my reading. So I just decided to review the stuff in the newspapers.

So, first of all, Tom Chivers, I was reading through the newspaper

to see if information on the essay matches the information in the original source, especially about the biotech companies, whether the biotech companies cannot develop some GMO products in poor countries. But I did not find any kinds of information that's in here.

The information about the unprecedented power of organizations or GMO companies that can exercise enormous power over the people or something like that. That is the main idea of this newspaper, right?

It did not mention anything about the poorer countries, the GM technology cannot be developed in poor countries or something like that. Okay, I just double-checked whether that everything aligns with everything in the source materials.

And then I just went back to the third paragraph because the third paragraph has something to do, closely related to the newspaper article written by Tom Chivers. And I just still engaging in the decision-making process whether this matches with the original stuff.

And I decided, okay, somehow the paraphrased materials are biased, OK?

And then it's mixed with the examinee's own ideas. I cannot just differentiate between the examinee's original ideas and what is paraphrased.

Oh yeah, I went back to the second paragraph because thinking about the incorrect paraphrasing materials, I just went back to look at the direct quotation marks again to see if it's correct.

So I was kind of deciding whether I can give him or her a C or D or something like that, yeah.

And then I went back to the rating scale.

I intuitively thought to myself that this one should be like this one, C or D.

Because although there are some sentence boundary issues, but most of the grammars are, I mean the sentences are easy to understand, and also the choice of the lexis is quite good, I mean, satisfactory to me. And also I see the organization is not too bad. And the arguments, I see the clear arguments of the examinee's too.

So, the organization, the arguments, and grammar lexis established somewhere between the C or D or CD+.

But I cannot justify giving full credit to the conventions over there thanks to the direct quotation mark and the biased paraphrase. Because somehow the examinee distorted the original information. The examinee did not paraphrase the original materials as faithfully as possible. All right? I cannot just distinguish between what is original, the examinee's original ideas and what is paraphrased over there. So that's what I'm concerned about, the conventions.

So other than this one, I would have given this examinee CD+. But thanks to the direct quotation marks and the ... what is that, the distorted, biased paraphrased materials, I would just give him or her a C or D.

So again, accuracy of information is really important to me, yeah. Because source-based material ... The writing assesses the reading skills and the writing skills at the same time, right? That's my understanding. And also it just assessed the examinee's ability to decide the degree of importance of the information. Right? So, the reading skill is really important, but biased or distorted paraphrased materials or information in the examinee essay, that is indicative of the lack of this ability.

Rater 8 – Essay 8

My first impression was about these examinees, "Okay really good lexis".

components like "vitamins" and "fibers" and "carbohydrates" and those vocab, I didn't see them in the original text. And also, I see the composition of the body and I see the "compromises" or "minimizes". They are the good evidence of the lexis.

And also, I did see "as read in text one", okay, that is the location elements. We can see these kinds of elements in the research paper. So "as read in text one" as indicated by someone of higher level.

So the search materials are really skillfully incorporated into the essay too. "As read in text one" and GM food also this supports his argument, even in the introduction part "which creates dependency and reduces the probability of having independent citizens." So, it's the evidence of the examinee's paraphrasing ability.

So "I mean by independence citizens, people that can plan their own lifestyle without being conditioned by a second or third person".

I think that sounds a little bit redundant.

The thing is that I do not see the clear thesis statements in the introduction parts, okay?

I got the impression that he or she seems to disagree with the idea of using GM food.

But let me just have a look at the second and third one.

When I have a look at the second and the first body paragraphs, all of those things are actually from the examinee's experience. Okay?

"I'm originally from the Dominican Republic"

That is not a good start, especially as the beginning of the paragraphs. I'm expecting to see something like the topic sentence but that is not, obviously not the topic sentence.

So "I have known about companies that utilized poor quality GM food hurting our communities and because they pay..."

All the supporting details are good to me.

And "money is power and totally control people do not have a high ethical standard"

And that is good words too.

And "that being said", okay there is a high grammatical structure.

Okay I think there should be some other one, yeah, grammatical error yeah.

But that didn't interfere with my understanding. This was just tiny mistake I think.

And also, for the third paragraph "it is important to mention that I am a member of a church".

It's about the examinee, his experience again. Not really relevant here.

"that encourages to have food stored for emergencies and to grow their own fruit if possible"

and also here I do have a gerund as the subject for the entire sentence "having your own garden"; so that is good evidence for the grammatical complexity.

"Increases the probability to improve our health because we know what we are consuming and can also help to build a solid financial life for us".

Okay I do not see any source materials as supporting ideas for these topic sentences.

And the topic sentences are up in the air. I do not see the clear topic sentences for the second and third one.

In the last paragraph I see some things are actually from the original source, are incorporated into the last paragraph. So "based on the texts". We have two text. So I'm not sure which one is referred to as in here.

"So the GM food is just a way to control people"

It is quite unclear because the one text is in favor of the use of the GMO, whereas the other text is against the GMO. So, you cannot say that "based on two texts", okay? I think the examinee should have identified which text is being referred to in here.

"There is nothing more powerful to do it then by food and money. On the other hand...", and so "as I mentioned before, cheap does not necessarily means good or excellent."

I think there are some grammatical mistake

but they are not critical ones.

"In text two is observed that investing supplementation will cost more than just consuming a natural well grown healthy golden rice..."

I think some grammatical structures errors over there again.

That is a good long sentence, I meant that just "consuming of natural well grown and healthy golden rice". So, I do see a lot of putting pre-modifiers. I mean that is a quite complex, long phrase okay? Well written long phrase over there.

So "it is completely disappointing finding people that support such type of food that will end up hurting not only them but also their posterity."

But ... I don't see the clear topic sentence for here okay, for the last paragraph.

And then how do I say, it ends in an abrupt way. I do not see any concluding paragraph.

And the reason is I think, okay the last paragraph might function as the concluding paragraphs, but I do not think that is the case.

I just went back to the rating. So I just look at the organization for the pass level.

Because this wouldn't meet the pass requirements for the organizations

because I do not see the clear topic sentences, and thesis statements.

And I do see the arguments of the examinee, but they are not fully elaborated. Not clear. Some of the sources are skillfully incorporated, but it's not enough. But the grammar and lexis things, they are really good. Okay? Some of the mistakes I ran into, but they are not the critical ones.

So, I will do give the pass grade to the grammar and lexis and conventions.

Even I could possibly give the pass grade to the arguments in detail, but the thing is that the four paragraphs seems like a separate paragraph, okay? I do not see the clear connection between them. I do not see any some kind of logical connectors that would connect them in a coherent manner. I think that that's because of the absence of the topic sentences and the thesis statements.

Because organization matters a lot, especially this is the argumentative essay.

So that's why I just gave the C, D plus to this essay.

Then, I went back to the essay again.

I just wanted to make sure of my final decisions. To see whether they are the thesis statements or the topic sentence or something like that.

Rater 8 – Essay 9

All right. So, I had a look at the essay over there.

What grabs my attention is "state-of-the-art ways of improving life". Okay, that is a good way of expressing something.

And also "agriculture industry is not there an exception".

All right, that is good word, too.

And also "manipulating the genes". Okay. This is a new word I never seen in the source materials here, so, which means that at least the examinee knows these words.

"There are both cons and pros of taking this ability of modifying genes."

This is a good sentence, grammar over there.

So I see the variety of words, "manipulating and modifying genes". That is good sign of the vocab.

However, I do see some punctuation errors over there. Maybe there should be some semicolon and the comma between 'into work' and 'however'.

But that is not a critical one.

I think that is the heart of the entire essay, okay? That is the heart of the argument of the examinee as he or she says "I believe the negativity overweight".

So, I thought, okay, this should be revised as follows maybe, "disadvantages of modifying genes OUTWEIGHS, not "overweight", the advantages of modifying genes" or something like that. Obviously, there's some confusion about the use of the words

but obviously I knew what the examinee was trying to do in this part.

"In the first place, farms which are ... go across and draw more attention of biotech companies..."

So I just had a look at the topic sentence over here in the second paragraph and just read through the entire paragraph, okay?

And I checked whether the source materials are incorporated into the second paragraphs or not.

And then I move onto the next one. So, the second paragraph is about negativity, disadvantages or maybe the arguments about those who are against the use of the GMO. That's the second paragraph over in there.

That matches with the "cons" in the thesis statement, okay? The "cons" were mentioned in the first. And then the examinees move onto the next part which is what the "pros" of taking or the modifying genes, right? 'Billions of people rely on rice', okay?

It talks about the vitamin A deficiency.

And when I had a look at the British medical study, okay, where does it come from? I was not sure about that.

So, I decided I would look over the source materials again about British medical study estimates in total or something like that.

So but anyway, I kept reading the entire essay until the end of the one and I looked at the end of this sentence.

And it stopped in an abrupt manner. It just stopped suddenly.

And then I moved onto the source materials over there very briefly.

And then, move onto the next one. Yeah, here, "A British medical study estimates that in total vitamin A deficiency kills..."

But somehow "in total" over there and it kills exactly the same number, 'under the age of the five each year', and I just went back to the original essays over there. And here, 'under the age of the five year in total'. Exactly the same sentence is over there in the essay. It is not paraphrased at all. All right? So I found that. All right.

So, I just looked over the starting from the first sentence and then I looked for the topic sentences again. So this is our ... this is the statements, okay? Okay, this is about the cons and this is about the pros, okay?

But the entire essay is not complete. That is the main weaknesses of the essay.

And I found out, okay, the vocab and the grammar stuff are really good, and examinee seems to have a really good command of the complex grammar structures, too. But, the thing is that this is not complete and that is a really big weaknesses in this paper.

Oh, I looked back at the second part. I just double-checked whether my decision is right. And I was just trying to solely define my decisions about the quality of this essay.

And then I went back to the rating scale, yeah, especially about the thesis and the topic, topic sentence.

Thesis, cons and pros. But I was expecting to see after the pros paragraph, maybe there should be one more paragraph that shows the obvious attitude toward the ... I mean, examinee's obvious attitude toward use of the GMO. Okay? Maybe, okay, "I'm really in favor of the use of the GMO because this and this and this" or something like that. So it's like indecisive, actually.

I know that examinee was trying to strike the balance between the paragraphs

but I think the examinee should have more attention to the first parts because that is about the examinee's arguments, okay?

Yeah, looking at the ... especially the Pass columns, okay? I think just compare this one and I think, C/D and Pass.

Because I thought this examinee is somewhere between C/D+ and pass.

but the organization stuff, okay, that is great, but the arguments in here, not fully elaborated. All right? That is the main weaknesses of this one. Yeah, details and examples are somewhat clear and I like that, and also the grammar and the lexis is okay. And also, the convention things, okay, that is okay with me, too. But this one is not complete, okay? Here paraphrased appropriately, but there was the verbatim copy of the original sentence, all right?

So, that's why I couldn't give the pass grade to this examinee, but instead of giving pass I just decided to keep it C/D+.

Rater 8 – Essay 10

Yeah, I see a lot of good command of the grammar structures.

"...GM food as they are getting benefits out of it by having control over the entire food chain, while others are finding deep trouble out of it as neither they do not have ...".

But this sentence is a little bit too long, so I was kind of having difficulty understanding. I kind of understood what the exam is trying to say, but I thought somehow the entire sentence, "some people are favoring" all the way down to "the funds of such research", should have been separated into two or three sentences to make his message clear.

And I was kind of curious about the thesis statement.

"This essay will describe both of them and finally come to a conclusion".

So what is the purpose of this essay? So, the examinee is going to describe both of the texts and how he or she can come to a conclusion.

Let me see, so let's have a look at the second paragraph.

And obviously the first paragraph is about the danger of monopoly of a biotech company.

It just paraphrased a lot of main ideas of the original source. And I see that is really clearly, somehow satisfactorily paraphrasing, but I cannot see the evidence of citations. Okay. The examinees didn't bother to identify this one from this source text or something like that.

And then I just moved on to the next paragraph

which is about the second article. It's about the vitamin A deficiency again.

And I was really impressed by this sequence over the Lomborg publication here. "Lomborg (2013) found that..." 2013 in parenthesis along with the imported verb and that- clause. Okay, that's really quite impressive. All right.

And then I'm not sure about this one, not sure about the Indian names. It's also probably the source material again. thought I saw this name in the source materials too, in the same study.

And the examinees just went on to describe what's in there, in the source material. Which source of vitamin A, and the proponents argument is this one and opponents argument is this one.

Okay? And that's it. And so, this is about the second and third paragraphs about the description of the two source materials. No arguments from the examinee, okay? So, this is not effective source citation.

And the conclusion part is the most important parts of the entire essay.

And "However it is found that golden rice is the most cost-effective resource of vitamin A. Therefore, in my view, it is very good to have such technology but at the same time it has to be ensured that the benefits are reaching to all the people, irrespective of their country of origin or affluence."

I think this should be the thesis statement. And that this sentence should be placed right in this place.

So that's why I was just looking over this one over and over again.

And actually, this is an argumentative essay, the sources should be incorporated into the argumentative essay. But it looks really weird.

The appearance of this essay is really weird.

Vocab part and grammar part, and conventions is okay with me.

But the organization-wise, it's peculiar to me because thesis statement is over there, in the last paragraph. And the thesis statement in the first paragraph is like, "okay I will describe both of them and then I will talk about my opinion."

I'm not sure how these two would support his arguments in here. I wouldn't see the clear connections between these conclusions with both of them.

Because to me, descriptions do not mean any evaluation, right?

So I got the impression that examinees have the neutral attitude or he tends to argue both of the contents in the original material. So, I was really confused at this time. So that's why I just thought to myself, "this is really tough". Okay?

Organization. I see a clear organization but it's really unconventional organization, right? Thesis statements in the last one, it's all the body paragraphs it's just descriptions or the summary of the two source texts, right?

It's not an essay actually. It's just a summary task. It's really similar to the first task.

So, I thought to myself, and kind of curious about, can I just look at his or her first task. Maybe it could be possible that this examinee just copied from task one and pasted the same stuff into the second and third paragraph in this essay. That could be possible. But because I could not have access to the task one, I couldn't verify my hypothesis. But it could be possible.

Yeah. So as long as there are some topic sentences or the clear thesis statements in the first introduction, I would have given this examinee a pass grade because I like the grammar and lexis things and convention things. Argument is okay.

But the organization stuff is really peculiar to me. So, I cannot just give a pass grade in terms of the organization

and part of the argument and details because there are no details or arguments in that second and third paragraphs.

So that's why I just gave the C/D Plus to this examinee.

Well I looked at the essays again to double check my decision again.

To verify if the second and third one is actually the summary. Okay, so I just wanted to see some tiny elements of these argument or her arguments in this second or third one. But I couldn't see elements of arguments. It's just pure 100% summary. Yeah.

RATER 9 (02/08/2018)

Rater 9 – Essay 1

I was reading through the text from the first sentence. For this essay, I think it was okay to read through. I followed the student's storyline.

I didn't find any, that many serious grammatical errors here. I thought it was okay introduction.

And then I was thinking whether this student can pass or not, but I thought it's higher than C. That's what I had when I went through the first paragraph.

And then I slowed down a little bit when I saw the author's name. That was one of the cues. I was thinking whether this student was quoting or citing the original text appropriately.

And then I found that some additional, I think, interpretation of the test taker because I didn't think Vitamin A in the original text was explained in detail. But I got some additional information from this essay. I'm not quite sure whether I just missed the part in the original text. That's what I felt. So, I thought, okay, so this student understood the original text. That's what I thought.

So, I thought, "okay, so this is quite a good essay. That's what I assumed, so far."

And here, in the banana part, he introduced another example here. I knew it was not mentioned in the original text. But the thing was that, it was good to introduce an additional example so that they can help the reader to understand what he or she was talking about.

But the thing was that all the sentences were not really well-connected. They are just chopped apart.

I thought that this is not absolute pass as his or her writing skills wasn't that good enough.

And then I read through the third paragraph

and I found that the expression at the end of the first sentence, "that teaches but to an extent," and then I thought that, "okay, he or she's writing the way he speaks, he or she speaks." So I thought, "it's not that good. That's not a good sign." That's what I thought.

And then he or she didn't finish the sentence over there.

So, if the student had enough time, probably he or she could have C+ or D+, but he or she spent a lot time thinking about these sentences

so, in terms of writing quality, I would give C+ to this one.

But he didn't finish this, the full essay here. But if it was the decision in the placement test, I would give C.

But I thought this student, if he or she's assigned to C, I think he or she will be mad.

Rater 9 – Essay 2

Now I started reading through the first sentence. "There are people who say it should not be supported, also..."

So, I found that the first comma there.

And then, by connecting these two sentences, the test taker used "also". It's not the proper conjunction there.

So, I noticed that, "oh, he's not a good writer."

Let me just call he, I don't know if it's he or she, but...

And then, I went through the second sentence.

I thought even though the essay was short

but it's a good start and introduction. I thought, "okay, and yeah, it was a good start."

And then I moved on to the second paragraph and in the first sentence, I thought that he's a good writer at that point, so far.

And then in the second sentence, he introduced some examples there with the golden rice and all those numbers.

And then, the sentences were well and neatly written there; so, I didn't stop there.

But I thought to myself whether he just copied the sentences from the original text. That's what I was a little bit suspicious at that time.

But I just went through to the end of the whole essay.

And then I decided that I would check it later.

Okay, so I moved on to the third paragraph, which is the last one for this student.

Here he said, "the author from text two." Even though it is not academically proper way of quoting something, but I just assumed that it's okay in this essay EPT test.

And then here also, the sentences were not bad.

And when we reached the next part of the sentence here, "The only goal for any company is to control ...".

Here, it's not his own argument. He borrowed that idea from the original text.

And then here he says that, it's like it's already accepted truth for the GMO companies doing these things. It's one of the arguments in one of the texts, so I expected that he's using some hedging expressions, "it seems", or at least he can say "I think", but here he just used, "The goal is to control," just like accepted truth, but it's not.

So, I thought that the expressions are not that academically acceptable.

And then he couldn't finish the whole essay as well.

So, at this point, I thought, "Is it okay to give a C?" That's what I thought.

And then I moved on to check the original text. I was just skimming through text 1.

I was looking for the numbers for the first essay. I was looking for the numbers whether he cited those numbers from the first text. But they were not there.

So, I moved on to the second text.

And then, the second text I found numbers running, so my eyes are on the numbers there and then I was looking for the expressions which was used in the original text. Here, "\$100 for every life saved from Vitamin A." And then I found that one.

And then, I moved back to the student's essay, and then I compared whether those sentences are identical or not.

And then I found that those sentences over there in the second paragraph, those expressions like "for each life saved from Vitamin A deficiency" were exactly the same.

So, I deducted one... I gave one level lower than what I originally thought, which was a C. So, I gave it a B+.

Rater 9 – Essay 3

First I had look at the overall structure of the writing. Not the details, but I wanted to see how many paragraphs this student wrote.

And then I read through the first paragraph.

And then I thought the first sentence was okay.

And then I found a little bit of grammatical or errors there.

And then still I had something in mind. I had the length of the essay in mind, "Oh this student might be good at, I mean, in terms of the fluency." That's what I assumed.

But I was keep finding some grammatical errors. I don't know what words they were.

And then there I found that, last sentence of the first paragraph. "Because they are lots of side effects on it"

and then I thought that "there are lots of"

so probably this is typo. I was struggling over there. Is that a typo or not?

But it wasn't really awkward.

So, I thought probably this is a C plus.

And then I moved on to the second paragraph. And then I read through.

And then I found the word 'effect' and 'affect' again, spelling issue.

And then I was getting suspicious that at this point, possibly this is a C maybe because there are some grammar errors. That's what I thought.

And then, in the third paragraph I read through.

I thought there were some expressions that was not very natural. "easily being damaged by natural disaster." I think I just stopped there.

There's a typo. "This will lea to" - it's just missing D. As you can see, there are several typos.

And then at this point, I thought, "is that really typed in by this student? That's what I was thinking, because there are many typos which is ridiculous.

And then, to the fourth paragraph, I went through.

I think it was okay at first.

But when I read through the fourth paragraph, I was a little bit confused. In terms of the organization, I was wondering whether he is doing okay.

Then I was struggling whether it's going to be C or C plus.

And then, as you can see here, my eyes moved back to the first paragraph. So, I wanted to check the thesis statement of this essay.

And then I came back to the fourth paragraph, and then I read through.

And then I thought some errors were there, the sentence linking with the connections were not that natural.

I was thinking that whether I am reading a written text or transcription of auto text. That's what I thought over there. I mean, it's just like the way he speaks. He wrote it the way that is like that.

And also, here, there are several typos, and also the verb 'affect' 'effect' - the typo.

Almost the end part of the essay, I found weird sentences. That one - 'hardly technology country. It's a weird expression.

I was thinking to myself, "Is that 'hardly advanced, technologically advanced country?'" I was just looking for the right expression for that one. That's why I slowed down.

And also 'monopolized' - that was different spelling.

I read through the conclusion.

“Not supposed to be suppoeted”? “Supposed” or “supported”? Yeah. Spelling.

"Even they have benefits in other way. You have freedom to make decision and your health is on your own."

Yeah, so sentence fragment over there.

And some expressions are just strange.

And then finally I decide that, "Okay this student is not that good." So, I finally decide that it's a C.

Rater 9 – Essay 4

So, I read through the first sentence in the first paragraph. I felt like it was generally well written.

And then I read through to understand.

And I found a little typo, upper case, lower case “Genetically modified,”

but it was nothing to me.

I was just following what he wrote over there

and then I was looking for any expression that was just simply copied from the original paragraph. But I couldn't find anything; so it was okay.

And then when I went through the third or fifth line of the first paragraph, I noticed that, "Oh, this guy doesn't have any paragraph division. It's all written in one single paragraph."

But still, the flow was okay.

And then I found that "I" was lower case

but I thought that it was okay as well. Generally, it was okay.

“They can hoard supplies and topple the government itself.”

I noticed that he was using those low frequency words, like "hoard". Is that a word? And "topple"? Those expressions I am not even familiar with. So probably he wanted to show some low frequency word? So, I don't know whether we can use that in that context, but that's what I found. And then probably he may have a good vocabulary. That's what I thought.

At the end of the first paragraph, I got the impression that he's writing the way he speaks. Spoken language. It's just like speaking.

He didn't use the, what is that, dependent clause. It looked like it was all connected with "and, and, and".

I consider they are more complicated expressions and they are used more in written than spoken.

And then the last part, it was just summaries.

I think it was almost equivalent to the first paragraph. (do not code)

I mean the level of writing for the second paragraph was ... It's just like spoken. Sophisticated expressions, but as I said it just felt like spoken.

And the quality of the writing wasn't ... It wasn't that bad

but I think the organization of this writing wasn't that good.

And then he tried to, I don't know whether he was trying to disguise the grader by putting the low frequency words sometimes because he knows the tricks probably.

And also, I got the impression that probably the student's from India or some area. That's what I got the impression. I think that's what I've seen in my class a lot. The students from that area write this way.

So, I decided this was C+, not a full pass. Conditional pass because it wasn't well organized, but the sentence level was okay.

And the second last sentence, "As we can see, just the 50 grams of the [inaudible 00:05:05] rice offers 60% of the daily vitamin A requirement and when compared to supplementation and fortification, is way more cost effective".

So here, I felt like it's just spoken expressions. It's "way more cost effective". I would say, "Highly cost effective" or some similar expressions.

Right. And it's just connected with and, and, and. So that's what I felt.

Yeah.

Rater 9 – Essay 5

First, I noticed that this was short essay.

And then I read through from the first paragraph.

And then I thought that general sentence structure was okay. It was easier for me to follow. It was generally okay, but I felt like it was not decent sentence. It's not that well written.

And then in the first paragraph I just followed the logic of the story line and moved onto the second paragraph.

I found the expression "in result of". He kept using "in result of", I don't think it was a good expression because he said in his statement that "GM technology does do much more benefit right now". He was kind of saying that "Oh this is going to be a good story" then he kept using "in result of." I think I would have rather like one using "thanks to", or something positive words.

And "GM technology is much cheaper..."

And then around this part, in the second line of the second paragraph, I felt like he is also speaking style, not academic.

Then, the conjunction wasn't that good enough.

So, I thought to myself, "okay this guy does not really well-trained, he's not well-trained with academic writing style. He's just writing the way he speaks."

And then the last paragraph, I didn't find any huge grammatical mistakes there,

but I was not satisfied with his writing in terms of the quality like vocabulary use. Use appropriate vocabulary than he did here.

He should use more logical connectors, I think that's one of the important things.

And also, this is very short essay, 3 paragraphs.

And then I was thinking, I gave the final score C; but I was thinking this could be C+, but I was in between, but close to C. That's why I gave a C.

After finish, I wanted to check whether my final decision was okay. If I found anything tricky, probably I would go back. But for this essay there were no such things, but I wanted to just recheck, because I was in between C or C+.

So, I just briefly went through this the writing again.

But I think it's close, leaning towards C.

Rater 9 – Essay 6

I read the first sentence, "genetically modified food is more popular such as the rice, plant, vegetable, even meat".

And I found that the use of "such as" wasn't correct I think in this context. I don't think its correct.

And also, I found a missing word in the second sentence. "Find the" what?

And also, the structure of the sentence is very simple. Every end of sentence almost every sentence in the first paragraph is quite simple.

The ideas were all chopped apart.

So, I thought that "okay so, this is level C or lower."

The second paragraph, and I thought that it was... let me see... "First of all". Then I looked at the other paragraphs to see the transitions. So "second", "third". Very simple.

Also, I found this word "one specie" or "species," this thing. That's not plural, is it?

And he copied this whole sentence without rephrasing or doing anything else to copy the sentence from the original text. So, he's making his sentence by using ... by copying the sentence from the source text.

At least once you copy, you have to give some comment, or you have to give a very short quotation or if you have a long quotation, probably you need to have some explanation, comment on that one. But here, he just ... in one sentence, he just copied the whole sentence, long sentence and period that's it.

He has "the author mentioned that" at the front, but so what? I don't personally like it.

If I were in his situation, probably I'd just copy the keywords and put quotation marks. And then rephrase the other part of the sentence. That's more natural.

And the last sentence in the second paragraph is also very... I mean the idea is not sophisticated. Too simple.

And then the third paragraph, again, "So as text 2 said". He did not comment on this information at all.

And there were plural-singular forms of the words, which was not correct.

Yeah, I think the idea and also the sentence structure, both of them are quite simple.

“Some of them we cannot found” and then “make a bad influence for human body”. Is that possible or not? Yeah, and also the last sentence he says, “it will influence us on so many ways”.

I don't think it's natural. It's like just speaking

And then mostly, I was thinking whether it is just B because to writing is quite simple and ideas ... idea itself wasn't that good. But I think this level could be B plus.

So, I came back to the evaluation rubric again and I checked the words there in the evaluation rubric. So, I compared B and C; so I was moving, looking around. I mostly focused on grammar and the organization.

Then I decide that okay, this could be somewhere in between. B+.

Rater 9 – Essay 7

I was reading from the very beginning.

The first several sentences were okay to me even though the second and the third sentence could be combined because one is a positive the other is a negative one, but it just chopped it off into two sentences, but it was still okay.

At first, I thought that this essay was well written. So, in general in this essay, I thought that probably it started with a good organization.

And also, good sentence structures.

I thought this could be a pass essay.

And then, in the fourth sentence, “as a personal thought about this argument”, I thought is a redundant expression because it says, “I think that genetically,” so it's good enough.

So, I was a little bit suspicious, now I'm going to lower the expectation here. Now I change my mind, but I was still looking for whether it was C or C+.

And then here in the second sentence in the second paragraph, “rice has become is a serious problem”.

“Rice has become is,” did he miss the WH? What rice has become? Or did he just use two multiple verbs there as an error. So, I figured out that that was a verb error, yes.

Yeah, I stayed there quite a long time. As you can see, I stopped there quite a long time because I was trying to figure out what he was talking about.

And then I found the first weird expression, “has being developed.” It's not a correct expression, right? And then I found another “has being developed”. And then I found here another expression, “vitamin A deficiency has being fixed.” That was another error.

Yeah, now at this point it was a C level essay.

Oh okay, “everyday, people just needs to eat... then they will have blah blah blah”.

He's connecting these two parts of sentences by using “then”, T-H-E-N. I don't think that was logical to use that one; so, I think I was looking at the first part of the sentence and the second part of the sentence. Looking for whether he mistakenly used “then”.

At first, I thought that was he wanted to use T-H-A-N; so that I was looking for any superlative form in the first part and then I couldn't find it.

And then I figured out “okay, so he was going to use T-H-E-N, but just by connecting these two parts of the sentences.

And then I moved to the third paragraph. The third paragraph I read through.

Overall, I thought it was okay.

In the third line of the third paragraph, I thought that without proper conjunction, he was just linking sentences.

That's just the way when we speak, which is not academic writing.

And then at the last sentence ... Yeah here I think I had some issues with the expression, "they are health is the most important thing for family." I could not understand this.

In the last paragraph. Again, "has being developed", wrong grammar.

I thought he spent his whole energy for the first part and then the remaining parts were not good enough.

Now, I wasn't quite sure whether it's could be C or pass.

So I decided to look at the evaluation rubric.

I looked at the part that uses the mostly Pass descriptors and then I just envisioned whether this essay was good enough to have the pass category.

But it wasn't good enough, so I finally decided as a C+.

Rater 9 – Essay 8

Okay, I read the first sentence.

I thought it was well written and then it was easy for me to follow.

And then here, in the last sentence of the first paragraph, "I mean by the independent citizens, people that..."

Here, I noticed that he is writing like speaking “I mean by the independent citizens”, which is not right. That is what I found.

But still he is a fluent writer at this point.

And then he introduced himself around this part in the second paragraph in the second sentence and because the sentence is all connected without a period

so, I was a little bit lost on what he was talking about.

And also, here I found the grammatical expression "that being said". Even though those are high frequency words but that expression is a little bit difficult for novice writers. So, I thought, “okay, he is using the expressions correctly, and then some difficult expressions as well.”

So, in the second paragraph, in terms of his idea, I didn't really like it. I don't know, he's writing something that's not satisfactory in terms of logic.

But I think the way he wrote was okay, I mean the sentence structures.

But, his writing style itself was just weird.

And the third paragraph, I was reading but I couldn't follow his idea.

I didn't like how the logic of this paragraph. It doesn't make sense to me.

I thought syntactically, these sentences are well-written and I could tell that the sentences themselves were okay.

They are also connected well.

But semantically, I thought there are something broken links. I could say supporting details are not well provided.

In the last paragraph, "it is completely disappointing finding people that support..."

Here I also thought that this is also speaking, because in written language, it doesn't make sense. "It is completely disappointing finding people that support..." I'd rather it says "It is completely disappointing to find that people"

So, I thought that he's a fluent writer

but I found something that is not natural to me.

So, I reconsidered whether I could give this a Pass. Well, I decide to give a Pass, but I was not really satisfied. I would go with a low Pass.

Rater 9 – Essay 9

So I started reading the first sentences.

In general, I think that the sentence level was okay.

So, I thought that I could give a pass after reading the first two sentences exactly.

And then he said that "there are both cons and pros..."

Usually we say "pros and cons", but he used "cons and pros".

And then he added, "however, I believe the negativity overweights".

So, I thought that negativity is the theme of this essay and the third sentence was the thesis statement of this essay. That's what I expected, even though there are two sides, but he sided with the negative part.

So, at first, I thought it was a pass.

In the second paragraph, I read through. I wanted to have the flow of his idea, and then once I started the second paragraph, I read through and tried to understand what he was talking about.

It took me a while to process what he was talking about because of the organization.

Once I finished the second paragraph, I moved on to the third paragraph, and then I read through the third paragraph.

In terms of the sentence structure, it was okay. It was easy for me to follow.

Once I finished the last sentence of the last paragraph, the third paragraph, he couldn't finish the last part.

But I was confused what he was trying to say.

And then I went back to the first part of the third paragraph again because I was confused with this third paragraph. And then I came back to the second paragraph, and then I read it again.

What was that? I read again the second paragraph and I don't know whether I looked at the thesis statement again

And then I just figured out that he used the wrong transition over there. In the second paragraph it was about negative things, but he started the third paragraph with "on the other hand", and then it was another negative thing. So, I was confused.

And then, then I think I checked the thesis statement again.

So, in terms of his sentence, he is okay with his sentences but some errors here and there. But we can easily fix them. I think we just pinpoint, we give some feedback, I think this student could easily fix that. It's not a serious problem.

But I was not happy with his paper because those key connectors made me confused.

So, I gave him a C+.

Rater 9 – Essay 10

I started reading the essay.

"Neither they do not have any access of such technology producing GM food, nor do they have ..."

So, I was checking whether that expression is grammatically correct. I wasn't quite sure, so I was looking at it and then I was thinking, "Is this correct?".

And then, I just decided that "it's okay. I'm not quite sure, but I could understand the meaning. So, I'll just move on." That's what I decided.

I thought in terms of the sentence level they were okay.

And also, here the student used "while", a conjunction to combine two different point of view which I thought is smooth.

But this thesis statement was not clear. It's more like description of the following two paragraphs, saying he'll describe both of them and finally come to a conclusion. I could not see the viewpoint of this writer.

And then in the second paragraph, I just read through

and then I felt like "he was definitely a graduate student because of the writing style."

And then here there was a little small spelling error or typo. "A cheap produces".

But I thought that's a minor error.

And then the third paragraph, yeah I think that was easy to read through.

And then there is good citation here, the use of the reporting verb "found". Those things are the tips that I can tell this student is a graduate student.

And then I wasn't quite sure of the expression "against of something"

but it was okay for me to understand what he was trying to say.

In the last paragraph, it says that "it is found that". So, these reporting verbs, the expressions were more like journal research article. And then "it was warranted".

So, I thought that this is good essay. So, I could pass it.

So, the last part, I just went back to the second paragraph and third paragraph to review whether my decision was correct

just for flow of the sentences. Were they just naturally smooth?

Then I decided to pass this.

APPENDIX F: INTERVIEWS CODING SCHEME

Category	Class	Code	Example
Descriptor Clarity (Clarity of the description of criteria for each score band)	Positive	dc_pos	It's good, detailed. Very useful. Actually, once you get used to it, it's very good. When you're teaching those levels, it's understandable. It's very clear. You know what the rubric says.
	Mixed	dc_mix	It's straight forward. But source integration is the one thing that I'm not sure about. I usually looked at if the students cite the sources correctly. In this case I only found one text which misinterpreted the source texts. In that case, I usually penalize these people. But as long as test takers say something about the source texts, I will consider it integration about the source texts. But I don't know how integrative it is; so, that's my issue as a rater.
	Negative	dc_neg	I have a hard time with the "many" and "some", "many", "some", "mostly". But if I try to visualize it as a certain number it can help sometimes. But that's totally on my part how I interpret "many".
Score Band Number (Number of score bands available on the rubric)	Positive	sb_pos	I actually did find the categories (B+/C+) very useful. One thing that I found when I was doing the grading, I felt like I had more options. So those essays I was like "I don't know if this is a C or if it's a Pass, it's a C+" and similar with "I don't know if this is a B or if it's a C, it's a B+". And so, in that case it was very nice.... because they were also essay where I was like "nope this is a B, nope this is a C" but for others where I was like "I don't know" it was in the in between it was nice having that category to utilize.
	Mixed	sb_mix	I think the distinction of a B and a B+ isn't that useful. The position that I find myself in is the CD and the CD+. It's not quite a pass but it's not really a CD. That's the situation that I find myself in most of the time. B and B+, I don't really know how that would be useful, especially if they're going to go into the same B class. A CD and a CD+ might be a little but more important because sometimes you just can't give a person a pass. Like no no no. But they're not also that bad, I feel like.
	Negative	sb_neg	NA

Appendix F Continued

Category	Class	Code	Example
Category Weighting	Positive	cw_pos	I think that the numbers (%) are useful because without numbers, I think it suggests that all of these should be weight equally but clearly, they're not.... I don't think I would change the weight because while I prioritize organization in an essay... when we're thinking about how we're setting our students to be successful if we don't prioritize grammar as much as organization, I think we're setting our students for failure.
	Mixed	cw_mix	Maybe the weight for each category [needs improving]... I was always like, "what if this student fits very well in these two categories and doesn't for two other categories. So, should I be considering these rates? 50%? Should I be counting them?".... I don't look at that because in my mind I know conventions are not really big issue and to me arguments and detail are more important than organization. So that's why I'm not really looking at the rates much... Maybe remove the numbers. It doesn't affect my decision. I just ignore them.
	Negative	cw_neg	I think the ability to integrate source texts is very important, especially in task 2 because I feel it's very clear in the prompt that it wants you to summarize and incorporate your experience or any background knowledge not just rely on the summarizing part.... So perhaps we should have more weight for it or maybe we should put it in a separate category.
Criteria Relevance (Relevance of the rating criteria to the construct of source-based academic writing)	Positive	cr_pos	I think it is very well-constructed. I think it reflects the distinct components of the writing ability of the examinees. I think it follows the previous studies in the multifaceted facets of writing ability.
	Mixed	cr_mix	I can't decide when one other paper when the spelling was important or not. So, the spelling descriptor was exactly the same for C and D task... Normally I don't look at spelling much but in one of the papers the spelling error was really distracting, and it was too much. So, I had to look at it maybe for the first time in my life. And then I noticed that it is exactly the same thing, so it really didn't help me. Other than that, it looks fine.
	Negative	cr_neg	... somehow I think the first three categories, organization, grammar and lexis, and arguments should be separated from convention because convention is a distinct quality; so if we have more man power or time, how about rating the first three categories first, and then move to convention. It's just my opinion. By convention I mean source use. It should be a different category.

Appendix F Continued

Category	Class	Code	Example
Score Band Labeling (Labels used for the score bands)	Positive	bl_pos	NA
	Mixed	bl_mix	It's not exactly the same thing but for somebody who is familiar with the classes that students will take potentially based on these ratings, maybe I think too much about the classes they are going to take. Maybe it's good, maybe it's bad, I'm not sure. But somebody who's more impartial maybe thinking the class placement is not helpful.
	Negative	bl_neg	Probably that would make knowing what course they're going to put into irrelevant right. Make it more like, less burden thinking process and constantly second guessing your choices. Cause if feel like I always do that. Am I being too lenient or am I being too harsh? So, I think it would be... probably a number would be better. Like the EPT speaking they use a number.

APPENDIX G: FINAL LIST OF WRITING FEATURES RATERS ATTENDED TO

Category	Code	Description	Examples
Arguments & Organization	1	Reasoning, logic, or topic development	<ul style="list-style-type: none"> • But it already had problems in the way that they organized this introductory paragraph. • The writer did not elaborate on the thesis sentence. • It's a lot of assumptions here that hasn't, haven't been defended.
	2	Relevance	<ul style="list-style-type: none"> • So, then I was already confused. Okay, so the student starts talking about golden rice, but what is the point of this paragraph? I did not see a clear topic sentence. • It's kind of off topic, but not totally, because it says GM food.
	3	Coherence	<ul style="list-style-type: none"> • So, there's a lot of jumping between ideas so I don't really see it's not a transition between ideas, it's not very clear. • But still, the flow was okay.
	4	Use of transitions	<ul style="list-style-type: none"> • I feel like you need to transition with something like, "Based on what I ..." To still acknowledge. • I realized that they're using transitions, but they're not using it correctly. "First of all, secondly" is okay, but "further information" ... You don't really say "further information", in addition, something like that.
	5	Text organization	<ul style="list-style-type: none"> • I was reading this to see how it was for a conclusion. And my thought was, basically what do we have for a summary or a closing statement. Some sort of a true conclusion. • I'm happy to see the student was attempting to structure their essay in this five paragraph structure. • Basically, the whole thing guides the reader quite well, finishes off with a conclusion.
	6	Writer stance	<ul style="list-style-type: none"> • They didn't really state what their position was at the end. • and I was happy that I found where the position, the test taker's position is. • There was ... The topic sentence was not clear there. It was too long, so I couldn't make and decide about argument. I couldn't even decide if he is arguing something.
	7	Ideas or rhetoric	<ul style="list-style-type: none"> • It's not rhetorically very sophisticated. • the first reason is not very illuminating, the second one is... • Then also noting extra information that was not from the source material, but from the writer. • But I should say that I was happy to see that the student was trying to come up with an example.

Appendix G Continued

Category	Code	Description	Examples
Grammar & Lexis	8	Comprehensibility	<ul style="list-style-type: none"> • But you know, the text doesn't only need to have good grammar, good vocabulary, but it also needs to make sense because that's the point of academic writing, and this didn't make sense to me. • I really didn't like this text. It was very hard to follow. • then I read that, it was not very clear or actually I didn't understand it. • I couldn't process this sentence quite well.
	9	Word use	<ul style="list-style-type: none"> • But I do recognize in their first paragraph, they use the same vocabulary over and over again like, "Should not be supported, should be supported, should be supported," Right? • So I did a quick glance of some word choice which I liked.
	10	Grammatical complexity and accuracy	<ul style="list-style-type: none"> • In the introduction, they did say, "This essay will describe both of them." I didn't really know what "them" is referring to. • This is very ungrammatical.
	11	Overall language	<ul style="list-style-type: none"> • This person has a good control of language. • Uh language use is not strong.
	12	Spelling or punctuation	<ul style="list-style-type: none"> • There's a lot of spelling errors. • but still there's a lot of misspellings.
Styles & Conventions	13	Style, register, discourse functions, or genre	<ul style="list-style-type: none"> • It's got this bookish version of English, formulaic way of writing compositions. • It says 'IMAGINE THIS' at the beginning of the second paragraph and this style is not, to me, it does not look appropriate to the situation of this writing. • He or she's writing the way he speaks, he or she speaks. So, I thought, it's not that good. • So again, that doesn't sound how a native English speaker would speak.
	14	Source use convention	<ul style="list-style-type: none"> • What I'm talking about is when they were mentioning bananas. Okay, where did that information come from? In this third paragraph, they're just not exactly copy and pasting but just putting a lot of facts in.

Appendix G Continued

Category	Code	Description	• Examples
Discourse Synthesis	15	Understanding of source text	<ul style="list-style-type: none"> • I noticed that there's an attempt to cite, but what I noticed is that what the person is writing based on the reading text, there's a misunderstanding of the content of the reading text. Right?
	16	Citation quality	<ul style="list-style-type: none"> • There's an attempt to cite sources, but the integration is not done very well. • They're trying to summarize the information (in the source texts), but they're just not including their own opinion. • but you don't really know the purpose of those facts, right? What are you trying to say using these facts, right? • As far as I remember the text just said ... just “the second article said” or something like that in the beginning of the second paragraph. I thought that was not skillful even though the sources are there.
Others	17	Task completion	<ul style="list-style-type: none"> • Yeah. In the third paragraph it's clear that it's not done. Overall, this essay is very incomplete. and the way that they're summarizing it ... They're trying to summarize the information, but they're just not including their own opinion. So, it didn't feel like a task two.
	18	Text length	<ul style="list-style-type: none"> • Because, you can clearly see that it's not the ... The task says that you need to have four paragraphs, and this is two paragraphs • First glance, four paragraphs, finally. • So, it's (can be paragraph) very short.

APPENDIX H: WALD STATISTICS COMPARING PAIRS OF RATERS

1st Rater	2nd Rater	1st Mean	1st SD	2nd Mean	2nd SD	Wald statistic	df	p-value	Effect size	
27 (n = 90)	11 (n = 182)	0.14	1.14	0.18	1.08	0.28	270	> .05		
13 (n = 110)	11 (n = 182)	0.01	1.05	0.18	1.08	1.33	290	> .05		
13 (n = 110)	27 (n = 90)	0.01	1.05	0.14	1.14	0.83	198	> .05		
28 (n = 100)	11 (n = 182)	-0.05	1.10	0.18	1.08	1.69	280	> .05		
28 (n = 100)	27 (n = 90)	-0.05	1.10	0.14	1.14	1.17	188	> .05		
28 (n = 100)	13 (n = 110)	-0.05	1.10	0.01	1.05	0.40	208	> .05		
5 (n = 156)	11 (n = 182)	-0.13	1.00	0.18	1.08	2.74	336	.01	0.30	small
5 (n = 156)	27 (n = 90)	-0.13	1.00	0.14	1.14	1.87	244	> .05		
5 (n = 156)	13 (n = 110)	-0.13	1.00	0.01	1.05	1.09	264	> .05		
5 (n = 156)	28 (n = 100)	-0.13	1.00	-0.05	1.10	0.59	254	> .05		

Appendix H Continued

1st Rater	2nd Rater	1st Mean	1st SD	2nd Mean	2nd SD	Wald statistic	df	p-value	Effect size	
18 (n = 144)	11 (n = 182)	-0.23	1.08	0.18	1.08	3.40	324	< .001	0.38	small
18 (n = 144)	27 (n = 90)	-0.23	1.08	0.14	1.14	2.47	232	.02	0.33	small
18 (n = 144)	13 (n = 110)	-0.23	1.08	0.01	1.05	0.40	208	> .05		
18 (n = 144)	28 (n = 100)	-0.23	1.08	-0.05	1.10	1.27	242	> .05		
18 (n = 144)	5 (n = 156)	-0.23	1.08	-0.13	1.00	0.59	254	> .05		
16 (n = 154)	11 (n = 182)	-0.32	1.12	0.18	1.08	4.15	334	< .001	0.45	small
16 (n = 154)	27 (n = 90)	-0.32	1.12	0.14	1.14	3.07	242	.01	0.41	small
16 (n = 154)	13 (n = 110)	-0.32	1.12	0.01	1.05	2.45	262	.02	0.30	small
16 (n = 154)	28 (n = 100)	-0.32	1.12	-0.05	1.10	1.90	252	> .05		
16 (n = 154)	5 (n = 156)	-0.32	1.12	-0.13	1.00	1.58	308	> .05		

Appendix H Continued

1st Rater	2nd Rater	1st Mean	1st SD	2nd Mean	2nd SD	Wald statistic	df	p-value	Effect size	
16 (n = 154)	18 (n = 144)	-0.32	1.12	-0.23	1.08	0.71	296	> .05		
26 (n = 132)	11 (n = 182)	-0.50	1.15	0.18	1.08	5.31	312	< .001	0.61	medium
26 (n = 132)	27 (n = 90)	-0.50	1.15	0.14	1.14	4.10	220	< .001	0.56	medium
26 (n = 132)	13 (n = 110)	-0.50	1.15	0.01	1.05	3.61	240	< .001	0.46	small
26 (n = 132)	28 (n = 100)	-0.50	1.15	-0.05	1.10	3.03	230	.01	0.40	small
26 (n = 132)	5 (n = 156)	-0.50	1.15	-0.13	1.00	2.89	286	.01	0.34	small
26 (n = 132)	18 (n = 144)	-0.50	1.15	-0.23	1.08	2.01	274	.05	0.24	small
26 (n = 132)	16 (n = 154)	-0.50	1.15	-0.32	1.12	1.34	284	> .05		
7 (n = 134)	11 (n = 182)	-0.61	1.16	0.18	1.08	6.17	314	< .001	0.7	medium
7 (n = 134)	27 (n = 90)	-0.61	1.16	0.14	1.14	4.80	222	< .001	0.65	medium

Appendix H Continued

1st Rater	2nd Rater	1st Mean	1st SD	2nd Mean	2nd SD	Wald statistic	df	p-value	Effect size	
7 (n = 134)	13 (n = 110)	-0.61	1.16	0.01	1.05	4.38	242	< .001	0.56	medium
7 (n = 134)	28 (n = 100)	-0.61	1.16	-0.05	1.10	3.77	232	< .001	0.50	medium
7 (n = 134)	5 (n = 156)	-0.61	1.16	-0.13	1.00	3.75	288	< .001	0.44	small
7 (n = 134)	18 (n = 144)	-0.61	1.16	-0.23	1.08	2.82	276	.01	0.34	small
7 (n = 134)	16 (n = 154)	-0.61	1.16	-0.32	1.12	2.16	286	.05	0.25	small
7 (n = 134)	26 (n = 132)	-0.61	1.16	-0.50	1.15	0.78	264	> .05		
21 (n = 88)	11 (n = 182)	-0.97	1.39	0.18	1.08	7.13	278	< .001	0.92	large
21 (n = 88)	27 (n = 90)	-0.97	1.39	0.14	1.14	6.02	186	< .001	0.87	large
21 (n = 88)	13 (n = 110)	-0.97	1.39	0.01	1.05	5.70	206	< .001	0.80	large
21 (n = 88)	28 (n = 100)	-0.97	1.39	-0.05	1.10	5.17	196	< .001	0.73	medium

Appendix H Continued

1st Rater	2nd Rater	1st Mean	1st SD	2nd Mean	2nd SD	Wald statistic	df	p-value	Effect size	
21 (n = 88)	5 (n = 156)	-0.97	1.39	-0.13	1.00	5.21	252	< .001	0.69	medium
21 (n = 88)	18 (n = 144)	-0.97	1.39	-0.23	1.08	4.45	240	< .001	0.59	medium
21 (n = 88)	16 (n = 154)	-0.97	1.39	-0.32	1.12	3.91	250	< .001	0.51	medium
21 (n = 88)	26 (n = 132)	-0.97	1.39	-0.50	1.15	2.73	228	.01	0.37	small
21 (n = 88)	7 (n = 134)	-0.97	1.39	-0.61	1.16	2.09	230	.05	0.28	small