THE VALIDITY ARGUMENT OF A WEB-BASED SPANISH LISTENING EXAM: TEST USEFULNESS EVALUATION

Abstract:

This study describes research used for supporting a validity argument for a new Spanish Listening Exam (SLE), whose scores are intended to place examinees into appropriate levels of university Spanish classes. This study contributes to the field of argument-based approaches to language assessment by implementing Bachman's (2005) assessment use argument framework (AUA). The validity argument is supported by research providing backing for three warrants pertaining to the quality of inference that can be made on the basis of examinees' performance. These warrants, based on three qualities of test usefulness (Bachman & Palmer, 1996) guided test development and research toward a placement test with the purpose of grouping students according to their levels of language ability within the framework of the communicative approach for learning Spanish.

Keywords

argument-based approach, assessment use argument, validity argument, test usefulness

INTRODUCTION

There is little doubt that a language program needs to assess students with a tool that is similar and appropriate to the current pedagogical beliefs and research needs. This constitutes the main rationale for doing research in designing a placement test based on language tasks in a language class domain. The purpose of this study is to apply a recent theoretical validation model, the assessment use argument (AUA) from Bachman (2005), in order to design and implement an online listening exam __ i.e., Spanish Listening Exam (SLE). As Bachman (2005) states, the process of articulating an AUA can be seen as supporting and building a legal case to convince a judge. In this study investigations of SLE test usefulness are included to support the warrants of the validity argument for the listening test in terms of three qualities of consistency,

construct validity and authenticity. The claim to support the warrant for consistency in the validity argument states that the SLE scores are consistent across different test-takers. The claim to support the warrant for the construct validity states that the SLE scores are valid indicators of the SLE construct and not of any other construct. The claim of authenticity looks at whether the listening task characteristics correspond to those of the classroom listening tasks. Investigations of internal consistency, IRT analyses, ANOVA analyses and a post-test study are credible evidence to support the SLE case.

While much of the recent work in validation theory has been conceptual (Bachman, 2005; Kane, 1992, 2004; Kane, Crooks and Cohen, 1999; McNamara, 2006), a growing number of studies are applying theoretical validation models to the actual exam design and implementation, a process that provides useful insights into how exam-construction data can be used to refine the theoretical frameworks. This study contributes to the field of argument-based approaches to language assessment by implementing the innovative meta-structure of Bachman's (2005) validation model. This is the first such study to investigate the applicability of a validation argument to an assessment that is based on language use tasks in the domain of a language class. In addition, since it explores this with speakers of a language other than English, this will contribute to the application of validation arguments in research into tests of languages other than English.

LITERATURE REVIEW

From the 1990s on, validation research has been seen as the process of determining claims and constructing an argument about the inferences and uses made from test scores and evidence collection (Chapelle, Enright, & Jamieson, 2008; Kane 1992, 2004). Kane (1992)

describes the concept of "interpretive argument" (Kane, 1992, p.527) as the inferences and assumptions that are supported by relevant kinds of evidence. A model based on this interpretive argument was built by Kane, Crooks, and Cohen (1999). The interpretive argument model has four components that are linked to each other in terms of three bridges: 1) scoring, 2) generalization and 3) extrapolation. Kane (2001, and 2004) extended his three-bridge formulation of the interpretative argument to four-bridge formulation (test use). The three first bridges are the inferences we make when building a test and with the last bridge Kane addresses the job of test use as illustrated below in Figure 1.

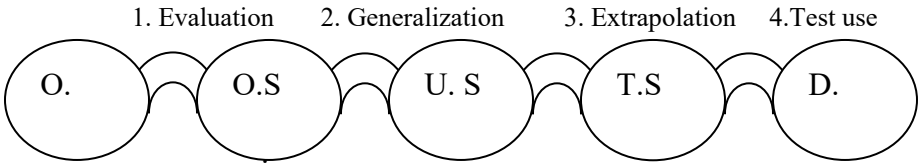


Figure 1: The bridge analogy: Observation (O.) to Observed Score (O.S) to Universe Score (U.S) to Target Score (T.S). Finally, a decision (D.) about the student's score can be made in order to use the score.

The first bridge deals with "evaluation," the score that we infer from a test performance observation. The second bridge or inference is "generalization," from the observed score on a test to a universe score, or the observed score seen as the score that would be expected of a student over a universe of generalization on similar test tasks. The third inference to what Kane et al. (1999) have named "extrapolation" from universe score to the target score represents the context of what test takers can do outside of the test itself. Finally, the fourth bridge deals with "test use" which is a decision about the test taker for which the score is used. With this sequence

of inferences it is possible to interpret the test-score and use the observed score to make a decision.

Researchers have concentrated their attention on test use and the consequences of basing the validity argument (Bachman, 2005, Kane 2004, Mislevy, Steinberg, & Almond, 2003) on Toulmin's (2003) argument approach, which consists of making an assertion of a claim, and if this assertion is challenged, defending it with data, facts, and evidence, to determine if the claim holds up.

The approach to validation taken in this research integrates the various qualities of good tests into a validity argument which supports the score-based inferences of the test. This structure of the validity argument is based on the work of Toulmin (2003). Toulmin's argument approach with different components to support an initial assertion is illustrated in Figure 2:

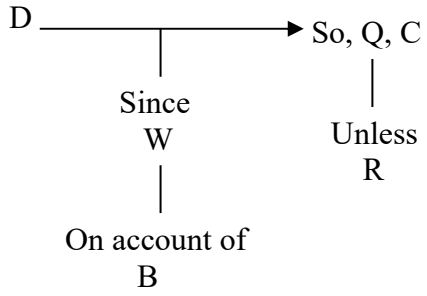


Figure 2 Toulmin's argument approach

Figure 2 illustrates that an inference starts from a datum (D) and from this datum we make a claim (C), conclusion, or what in language testing is called test-score interpretation. A datum is the necessary basis for a claim. A qualifier (Q) indicates strengths (e.g., including 'certainly' or 'necessary' words) or weaknesses (e.g., including 'possibly' or 'presumably') of our claim, while warrants (W) are decisions used in order to interpret test scores or statements to go from a datum

to a claim. These statements support our claims. Backing (B) is based on theory, prior research or evidence collected during our validation process in order to support our warrant. Finally, rebuttals (R) are problems that might potentially arise in which the warrant would not apply along with alternative explanations to solve these problems. (Diagram taken from Toulmin 2003, p.97)

Bachman (2005) draws upon this structure to outline two parts of what he calls an assessment use argument (AUA). The AUA responds to the problem he identifies as arising from the lack of theoretical model for investigating test use and consequences, except for some lists of considerations for the development and use of the test. He states that during the process of articulating an argument that justifies the uses of a language test, we need to keep three important considerations in mind:

> "First of all, there is no guarantee that even valid score-based interpretations will be relevant, useful, and sufficient for the intended uses, or decisions. Second, there is no guarantee that these interpretations will not be subverted for other uses, or be used for decisions other than those for which they were intended. Finally, the validity argument by itself provides no basis for either anticipating or investigating potentially unintended consequences of the way the score-based interpretations are used" (Pp.14-16).

Accordingly, Bachman (2005) suggests an AUA that also employs the same structure as described in Toulmin's approach. Bachman (2006) also believes that the AUA that he proposes is what Kane has called as two parts of an interpretive argument: The descriptive interpretation "... that estimate some variable for the examinees being tested, without specifying any particular uses for the test scores…", and the "decision-based interpretation" that involves making a decision about the examinee based on the descriptive statement (Kane, 2002, p.32). Bachman (2005) extends Kane's model by evaluating the relationship between these interpretations. We could define Bachman's model of use argument as a meta-structure for building an argument

using concepts such as claims, warrant, backing and rebuttal to examine test validation and test use.

The innovative function of this argument is not the types of evidence that may or may not be successful in a validation argument. Rather, its contribution is the nature of the argument meta-structure connecting the assessment performance to score interpretations (e.g., Spanish test results can be used as indicators of students Spanish proficiency) and to intended test use (e.g., linking scores to the proper placement level). This meta-structure allows test developers to consolidate test design, development, scoring interpretations and intended uses within a single model; a feat that, to my knowledge, has not yet been accomplished by any language tester. This meta-structure is comprised of two different arguments.

First, a validity argument links assessment performance to score interpretation. This structure includes the same elements as Toulmin defined them: claim, warrant, backing and rebuttal. Warrants could be based on the qualities of a test: reliability, construct validity, authenticity, and or other factors that are important in supporting the link between test performance and score interpretation. Second, a utilization argument begins where the validity arguments leave off—to link the score interpretation to the appropriate score use. The utilization argument deals with claims (i.e., the decision we intend to make), warrants (i.e., when the score-based interpretation are relevant, useful, sufficient and beneficial to the students based on the decision we will make), and backing (i.e., evidence collected) to support the intended test use, as well as with rebuttals (i.e., reasons for making a different decision) to investigate unintended consequences of test use.

The utilization argument relies on data collected during test use after the test has been operational. For a new test such as the SLE, therefore, the focus is on the validity argument,

whose conclusion will ultimately be the stating point for the utilization argument. The purpose of this study, then, is to describe and examine the validity argument of the SLE that constitutes support for the SLE score-based interpretation. The structure of the validity argument includes claims, warrants, backing and rebuttals which deals with reliability, construct validity and authenticity. The mentioned structure of the validity argument addresses the first three links in Kane's bridge (see figure 1): evaluation, generalization and extrapolation.

THE SPECIFICS OF THE SLE

The SLE is a web-based Spanish placement exam based on listening, intended for use in a Spanish program where the instruction of listening skills is considered important for language learning. The SLE is a linear exam in which the Item Response Theory (IRT) was used in order to improve the test reliability and validity of the test. The SLE is an instrument to be used in conjunction with additional tools that assesses other components (e.g., speaking and writing) of a Spanish language communicative approach. The Spanish classes offered at the University of California, Davis (UCD) in the first and second year were classified by the Spanish department at the American Council for the Teaching of Foreign Languages (ACTFL) level (ACTFL, 1986). Therefore, the standards of the SLE tasks are in terms of the ACTFL proficiency guidelines. The test structure of the SLE includes one task for instructions and practice in English[1], in order to know how to use the software, followed by ten Spanish tasks (i.e., different oral passages followed by questions. The first tasks are for intermediate low [e.g. announcement] jumping to novice high [e.g., descriptions] and following with more difficult tasks for intermediate low and the SLE ends with intermediate mid proficiency level tasks [e.g., dialogues]. Each task has a different number of questions from 5 questions to 11 questions.) Because the test's purpose is to place students in the Spanish lower-division program, the SLE includes ten tasks based on semi-

[1] See Appendix A for similar tasks for the SLE

scripted oral text types (e.g., spontaneous announcements, dialogues, mini-talks and classroom tasks), and local (i.e., looking for explicit details) or inference items (e.g., grammatical such as phonological, lexical and syntactic). The SLE has a total of 82 items for ten oral passages. Reading context and items are presented in advance. Test-takers answered items of some sort (e.g., multiple-choice, limited response, true or false) after listening to the oral input twice, – following the same instructions that the instructors use during the listening comprehension portion of a typical midterm exam. All tasks are equally important and all items are scored from 0 to 1 on grammatical knowledge. This testing method was conceived with a web-based test in mind in which automated scoring is one of the main features, thus, making it possible to process and review student performance rapidly with little cost or effort for students and teachers alike.

THE ASSESSEMENT USE ARGUMENT: THE VALIDITY ARGUMENT FOR THE SLE

The validity argument of the SLE follows Bachman's (2005) which is a blend of Toulmin's argument approach with Bachman and Palmer's (1996) model of test usefulness. Six test qualities are included in the test usefulness: reliability, construct validity, authenticity, interactiveness, impact, and practicality. Although the relative importance of these qualities will vary given the purpose of the test and according to context, test usefulness will help us find a balance among these qualities without ignoring any.

We will state the test qualities used in this paper in a moment because it enables us better to understand the method of the SLE development and the validity research conducted. The validity argument in support of score-based interpretation for the SLE followed Bachman's (2005) validation framework and is illustrated in Figure 3.

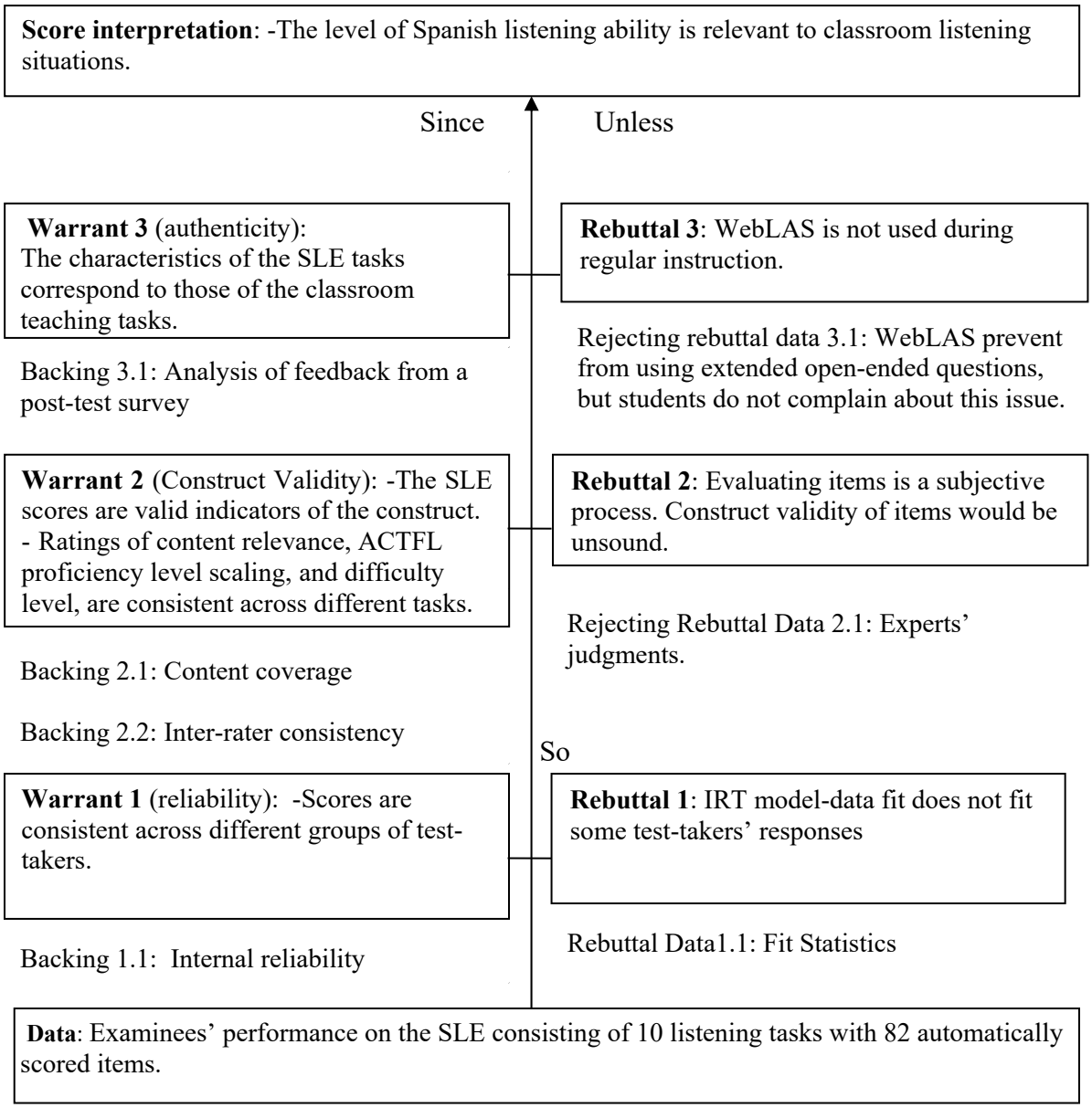


Figure 3. A validity argument for performance on the SLE

The validity argument posed three research questions regarding the usefulness of the SLE as a test, as listed below. It is not possible to cover impact in this paper due to limited space, so the utilization argument (i.e., the intended consequences and decisions of using the SLE) is not addressed here. So, the discussion will be limited to the interpretations about the test takers' language ability, the scores, and the test takers' performance on the SLE which is related to the validity argument during the SLE development. The validity argument supports the link between the SLE performance and the intended listening proficiency interpretations. The framework of AUA helps to identify the elements of both test usefulness and intended use for the development of the SLE. Claims, warrants and rebuttals will be stated for the listening assessment and evidence gathered from the present study will serve as backing to support the claims. The types of claims in the AUA will be supported by different warrants:

Claim 1: The scores that are obtained from test takers' performance are consistent. The backing and rebuttal of the reliability/consistency warrant serve as evidence to examine the first research question.

Claim 2: The SLE scores can be appropriately interpreted as indicators of Spanish proficiency. The rating of different raters across different assessment tasks is consistent. From the validity argument the construct validity warrant is supported by content coverage backing and inter-rater consistency. The backing of the construct validity helps to answer the second research question.

Claim 3: The SLE listening tasks are generalizable to the classroom domain. Generalizations are linked to the relevant placement decisions to be made when using classroom tasks. This claim is supported by the authenticity warrant which states that the characteristics of the SLE tasks correspond to those of the classroom tasks. The authenticity backing addresses the third research question in the study. These three research questions are summarized below:

1) To what extent are the scores that are obtained from test takers' performance consistent? Specifically: 1a. What is the reliability of the SLE? 1b. To what extent do the test-takers' levels of listening ability affect their item performance on the SLE?
2) To what extent can SLE scores be appropriately interpreted as indicators of Spanish proficiency? Specifically: 2a. To what extent is the SLE meaningful with respect to the construct to be assessed? 2b. What item features account for difficulty using the Rasch model? 2c. To what degree are the ratings of oral passages in this study consistent?
3) To what extent are the interpretations about the listening ability generalizable to the instructional domain? Specifically: 3a) To what extend are the SLE tasks similar to the instructed domain according to our test-takers' perceptions?

Qualities of usefulness

Given that the main function of developing this test is its usefulness we have carried out the process of the validity argument. That is, the components of the SLE validity argument are based on Bachman and Palmer's (1996) model that describes test usefulness along with three test qualities evaluated in this study:

Reliability: The consistency of scoring is obtained by using a database with possible responses for the limited production responses (i.e., typing a Spanish word). These responses are collected during the tryout of tasks. The scoring of the rest of items is consistent with the responses of a group of instructors. Consistency of scoring is also obtained through self-scoring.

Construct validity: A series of factors can be used as evidence of construct validity. The construct of the SLE test is based on an interaction between the listening ability, the tasks, and the Spanish course syllabi used at UCD. Since syllabi are altered over the years to reflect

changes in textbooks, thirteen Spanish textbooks are examined in order to identify the main topics (grammar and vocabulary) that normally are taught in the first and second year of the Spanish language curriculum. Buck's (2001) theory-based definition of language knowledge was used to measure the listening ability because of its two components of language knowledge and strategic knowledge. Content coverage evidence is demonstrated through item and task congruency with the instructional content of our elementary and intermediate language program. Finally, oral passage ratings of different raters are checked for consistency. For a detailed explanation of the construct validity see Pardo-Ballester (2008).

Authenticity: The authenticity of the SLE test is consistent with the instructional tasks presented during regular Spanish classes. Authenticity is not understood in terms of the tasks one is expected to do in a Spanish speaking country-- such as ordering a meal or serving food to customers—but, instead, on the basis of similar tasks that students might do during class.

Now, that the qualities of the test have been stated, we will present the four steps that we followed in the validity argument. First, the score interpretations are explained, followed by a presentation of the performance of the SLE tasks obtained from our test takers' data in order to make a score interpretation. Then, the warrants and backing to support the link between SLE tasks performance and score interpretation are presented and finally rebuttals and rebuttal data to support the validity argument.

1 Score-based interpretation

a) Covering the main topics taught in the first two years of a Spanish curriculum was one of the objectives during the development of the SLE. The exam's construct was based on building listening tasks by using different spoken passages. The difficulty or facility of the tasks was based on the grammar points, vocabulary, and content to assess. In addition, task difficulty was

also considered taking into account the accent, rate of delivery and intonation used in the oral passages. This specific construct was built with the purpose of assessing different levels of proficiency (see Pardo-Ballester, 2008).

b) Assessment results: The SLE tasks were piloted with Spanish instructors and students of different proficiency levels in order to consider the complexity of items or possible flaws. Modifications to test tasks and items were made when necessary. For now, data of 147 test takers shows the scoring evidence of different tasks for first-and second-year learners.

2 SLE tasks performance

Instructed domain for the Spanish classes at UCD: The listening ability of three Spanish proficiency levels was evaluated in relation to the ACTFL proficiency listening guidelines. The Spanish department at UCD classified the following Spanish classes at the following ACTFL level:

Spanish 2: Novice-high; Spanish 3, 21: Intermediate-low; Spanish 22, 23: Intermediate-mid.

Different SLE tasks were described at the beginning of this paper: A variety of grammatical points, vocabulary and discourse forms (i.e., descriptive, narrative, expository, argumentative, and instructive) were present in the tasks. In the SLE the tasks in the target domain were varied. This is because students of different listening abilities took this test to make sure that the beginning and intermediate texts selected by instructors represented the intended differences in levels. Web test tasks resemble classroom tasks and midterm exams during a regular quarter, except that in class a paper and pencil format is used.

3 Warrants & backing to support the link between the SLE performance and score interpretation

Warrant 1 (reliability or consistency): Ratings of the oral passages are consistent across different tasks. Students are placed into different proficiency classes according to the score they receive in the Spanish Computer Adaptive Placement Exam (S-CAPE). Results from the S-CAPE imply that there is a relationship between proficiency levels and students' test scores.

- Backing 1.1 (internal consistency): The alpha coefficient is used to estimate if scores are consistent across different groups of test-takers.

Warrant 2 (construct validity): The SLE scores are valid indicators of the construct, and not of other constructs. Ratings of content relevance, ACTFL proficiency level scaling, and difficulty level, are consistent across different tasks.

- Backing 2.1(content coverage): The Rasch IRT model is used to analyze and estimate item difficulty. The Rasch estimation procedures place item and candidate ability on the same scale. With the information about difficulty measures a two-way analysis of variance is used to investigate the content coverage of the SLE.
- Backing 2.2 (Inter-rater consistency): Four raters evaluate the oral passages of the SLE based on Spanish textbooks used to build the listening context, the ACTFL guidelines to guide the levels of the tasks, and other discourse features such as hesitations. Internal consistency is estimated to explore the judges' ratings.

Warrant 3 (authenticity): The characteristics of the SLE tasks correspond to those of the classroom teaching/ learning tasks. Students are presented with a test situation that deals with different spoken tasks similar to those used during the Spanish classes in order to teach or assess Spanish at different levels. Test-takers answer by selecting a choice or by writing a limited answer. These tasks included in the test are authentic because they are relevant to classroom instruction, but reading instructions to complete the test tasks are far from authentic during class

instruction. However, the prompts are authentic to the language instruction, that is, the students will not be surprised in finding the prompts in Spanish, because this is what they normally find when taking a test or during class work. In addition to that, the audio input is provided by Spanish instructors in the same way as when they are teaching.

- Backing 3.1 (authenticity): Results of students' feedback analysis from a post-test survey will help to indicate if there is similarity among SLE tasks and the ones used during the Spanish classes.

4 Rebuttals and rebuttal data to support the SLE validity argument

Rebuttal 1: Results of the IRT model-data fit might indicate that some persons or items on the SLE do not fit the model well (McNamara, 1991).

- Rejecting Rebuttal data 1.1: The identification and deletion of misfitting persons and items might improve the accuracy of the Rasch model.

Rebuttal 2: Listening comprehension deals with different factors related to L2 oral input and deciding the classification of items is a very subjective process. Items' classification could be unsound.

- Rejecting rebuttal data 2.1: Experts' judgment to measure item difficulty is taken into account. The evaluation process is done in two steps. First, each evaluator classifies the item individually and second, the group decides a final classification for all items together.

Rebuttal 3: WebLAS is not used during the regular course test, but rather paper & pencil listening tests are offered with open-ended multiple-choice questions. In addition, Spanish mid-terms which are administered in class measure listening through extended open-ended items while WebLAS offers limited open-ended items.

- Rebuttal data3.1: Automated scoring in the SLE prevents the use of open-ended questions, but students do not complain about not using open-ended items in the SLE because the participants have limited open-ended items where they have to write one word in Spanish.

The framework of assessment use argument helped identify the elements of both test usefulness and intended use for the development of the SLE. In summary, claims, warrants and rebuttals were stated for the listening assessment and evidence was gathered as backing to support the claims. The kinds of evidence collected in the assessment use argument came from a variety of sources.

METHOD

Test takers

147 students enrolled in different Spanish classes at UCD Davis participated in this project. The breakdown of the numbers of participants at those courses is shown in Table 1 corresponding to the ACTFL proficiency levels.

Table 1: Proficiency level students for the SLE

Proficiency Level	*Number of students*	*Percentage*
Novice-high	34	23%
Intermediate-low	60	41%
Intermediate-mid	53	36%

Materials

Audio recorded materials: Ten listening passages were recorded digitally by eight Spanish native speakers. The test developer controlled for volume, background noise, and recording time

using AUDACITY which is "…free, open source software for recording and editing sounds" (AUDACITY, 2001 http://audacity.sourceforge.net/)

Post-test usability survey: After examinees finished the SLE, they completed a survey that asked them to rate a) their level of satisfaction with their experience with the software, the familiarity with the test method used, the similarity with the Spanish classroom-instruction tasks, their strategies used when taking the exam, b) their impression of the difficulty of listening tasks, and c) their comments on using this online exam for assessing their Spanish comprehension.[2]

WebLAS: WebLAS is the acronym of the Web-based Language Assessment System. WebLAS was constructed by programmers working with test developers at University of California Los Angeles (UCLA Department of Applied Linguistics and TESL & Center for Digital Humanities, 2003). This software was used to create and deliver the listening exam on the World Wide Web.

Procedures

The procedure of this study involves three phases in which different Spanish lecturers and Spanish TAs were contacted via email, phone, or in person in order to explain the research study and solicit their collaboration.

The first phase of this project was the development of audio recorded material with a variety of listening input (i.e., dialogues, mini-talks, announcements and others) similar to those used in the classroom for the purpose of teaching. Eleven Spanish native were recruited for the production of the recorded materials. We opted for developing semi-scripted texts in order to control the topics and content (i.e., vocabulary and grammar) of the oral texts described in the

[2] Most of the results from the post-test survey are linked to the utilization argument and therefore they are not discussed in this paper. A forthcoming paper based on the utilization argument will be explored in terms of (1) the SLE tasks' relevance to the language instruction domain for interpreting test scores and deciding student's placement; (2) score utility and test-takers' perceptions about the difficulty of the tasks for the applied purpose; (3) the intended consequences interpreting scores and using them for placement decisions; and (4) the sufficiency for students' placement based on listening comprehension.

test specifications as much as possible. A list of topics was given to the speakers and they had about a minute to prepare the situation before the actual recording. Fifteen oral passages were selected from forty oral stimuli to the final version of SLE.

The second phase was the creation of listening tasks and the paper and pencil trial. The development of items was complicated, because of the limitation of using web-based items. Also, in order to avoid frustration for beginners, when the oral input was difficult, high-frequency vocabulary and simple grammar was used in a number of items. Test developers created the items by listening to the oral input rather than reading the text because this process ensured creating items which were focused on comprehension rather than difficult items which are normally based on memory of small details. To trial these new tasks we visited various classes in April 2005 to administer the listening tasks on paper and pencil. These tasks were considered to be a listening practice for the students. Once they answered all items they were asked to circle the words that they had trouble understanding and then add comments on the difficulty of the tasks. The rationale for asking students about the test difficulty was to ensure the exam's content relevance.

In the third phase of this study, we incorporated all suggestions and/ or recommendations from students and instructors in order to improve or eliminate items, and to evaluate the level of the grammar and vocabulary tested in those tasks. Item analysis such as difficulty of each item was not performed, but decisions for eliminating items were based on suggestions from the group of experts. Our experts were four raters for the oral passages and item content analyses and standard setting procedures. All raters were Spanish graduate students (i.e., Spanish TAs) with experience teaching Spanish courses at the lower level division. Raters were familiar with the Spanish instruction at UCD for at least three years and had taught at least four different

course levels. Three of them were female and one was male. Two of the participants were Spanish native speakers and the other two were near native speakers of Spanish. Nine items were discarded due to their ambiguity or inappropriateness as judged by students and instructors. Items were rewritten for the SLE based on the group responses. At the end of the spring 2005 the SLE, composed of ten listening tasks with a total of 82 items spread all over the tasks, was created using WebLAS. This exam was first tested with instructors and then with students.

Analyses

The WINSTEPS computer program (Version 3.63; Linacre, 2006) was used to calibrate dichotomous items using the Rasch model. WINSTEPS implements a two-step process to estimate item and ability parameters. After pretesting items, the Rasch model was used to calibrate the items. After initial fitting of the data, the fit of the model to both items and persons was assessed by checking the outfit and infit for means squares.

Descriptive statistics and a two-way analysis approach using the SPSS 12.00 (2003) package was used to investigate construct validity in order to determine the interactions between difficulty measures across linguistic items (i.e., lexical, syntactic or phonological items) and task level. A percentage response from a post-test survey was performed in order to explore students' responses about the authenticity quality of the SLE usefulness.

RESULTS

Results provide backing for the warrants of reliability, construct validity, and authenticity, which were investigated by IRT analyses, ANOVA and participants' perceptions of the use of the exam beyond the test setting were analyzed as proof of the test's usefulness

1 Reliability

The first claim in the AUA pertains to the scores that are obtained from test-takers' performance. The warrant for consistency states that the SLE scores are consistent across three different proficiency groups of test takers. Backing to support this warrant of consistency is discussed below. An exception or rebuttal for the first claim was also articulated indicating that person fit for some test-takers' responses does not fit the IRT model. The finding of the rejection of the rebuttal is the rebuttal data in the validity argument and is also discussed below.

1.1 The results of IRT analyses using WINSTEPS

IRT was used to calculate the internal consistency for the SLE scores across different groups of test takers and across the SLE items. The Rasch reliability of case estimate (equivalent of KR20/ Cronbach's alpha) was .87 for 144 students and for 82 items the reliability resulted in .94. This data is the evidence for the warrant which states that the scores are consistent across three proficiency levels (i.e., 144 students with different proficiency levels).

The Rasch modeling sheds some light on the initial data fit in order to explore the rebuttal of the reliability warrant which states that IRT model-data fit does not fit some test-takers' responses.

Criteria for fit according to the WINSTEPS manual were the following:

>2.0 Distorts or degrades the measurement system.

1.5-2.0 Unproductive for construction of measurement, but not degrading.

0.5-1.5 Productive for measurement.

<0.5 Less productive for measurement, but not degrading. May produce misleadingly good reliabilities and separations.

Following these criteria, all items were acceptable. However, three subjects (87, 119,137) had outfit mean squares greater than 2.0 and were dropped from further analysis. After

re-fitting the model without these subjects, the outfit mean squares for all items and subjects were acceptable. Next step for fitting the model was the assessment of item polarity. This analysis assesses whether all items are measuring the same ability in the same direction. Practically, this is assessed using the point measure correlation, with which a positive value indicates the same direction of measurement (i.e., polarity). Results showed that all items displayed the same polarity since all items indicate positive correlations with a range from .10 to .54 values.

Dimensionality was assessed using variance decomposition via principal components (also known as Eigen Analysis). The following criteria were applied:

Variance explained by measures> 60% is good

Unexplained variance explained by 1st contrast (size) < 3.0 is good

Unexplained variance explained by 1st contrast < 5% is good

Table 2 Dimensionality of the SLE

Table of STANDARDIZED RESIDUAL variance (in Eigenvalue units)

			Empirical		Modeled
Total variance in observations	=	168.1	100.0 %		100.0 %
Variance explained by measures	=	86.1	51.2 %		48.5 %
Unexplained variance (total)	=	82.0	48.8 %	100.0 %	51.5 %
Unexplained variance in 1st contrast	=	3.4	2.0 %	4.2 %	
Unexplained variance in 2nd contrast	=	3.0	1.8 %	3.7 %	
Unexplained variance in 3rd contrast	=	2.9	1.7 %	3.5 %	
Unexplained variance in 4th contrast	=	2.8	1.7 %	3.4 %	
Unexplained variance in 5th contrast	=	2.6	1.6 %	3.2 %	

Table 2 shows that nearly 50% of the variance in the data is explained by the model (48.5%). The percentage of unexplained variance represented by the first contrast is only 4.2%, which compares to 3.2% in the 5th contrast. If the data had fit exactly, the shift in variance explained by

measures would have been only 51.2 - 48.5 = 2.7% in absolute terms. These data do not support the existence of multidimensionality beyond the single Rasch dimension.

2 Construct validity

The claim pertains to the interpretations of the SLE scores as indicators of Spanish proficiency. Analyses for item and task difficulty were discussed previously in a paper about the construct validity of the SLE (Pardo-Ballester, 2008). However previous results did not indicate which linguistic features with respect to task level help to discriminate learners based on their listening performance.

2.1 Two-way analysis of variance for content coverage

In order to explore the relationships between linguistic features of items and task proficiency levels in terms of the IRT difficulty of items, a univariate two-way ANOVA was performed and served as backing for content coverage. The difficulty of items is expressed in logits. Descriptive statistics for this comparison are presented in Table 3.

Table 3 Descriptive Statistics of linguistic points and task levels

Linguistic code	Task level	Mean	SD	N of items
Lexical	1	-.79	1.3	10
	2	-.43	1.0	26
	3	.44	.96	4
	Total	-.43	1.1	40
phonological	1	1.2	.60	3
	2	.33	.21	2
	3	1.8	.39	2
	Total	1.1	.72	7
syntactic	1	-.41	1.3	4
	2	.18	1.2	21
	3	.70	1.1	10
	Total	.26	1.2	35
Total	1	-.33	1.4	17
	2	-.13	1.1	49
	3	.77	1.0	16
	Total	-.00	1.2	82

Table 3 shows how the two independent variables were compared to the difficulty logits. The data are organized by mean, standard deviation and number of cases. The means represent the difficulty of the items. With the means we see an order in difficulty for the linguistic features of the items according to the task level (i.e., 1= the easiest, 2= moderate and 3= the most difficult) except for the phonological items of level 2 which is closer in difficulty for lexical and syntactic items of level 3. The mean of the phonological items in level 2 break the pattern of item difficulty which should have been 1.33 or 1.55 for example to keep the difficulty pattern. The means of the lexical and syntactic items together show an order in difficulty in terms of task level (e.g., the mean of the lexical items for level 1 indicate the easiness of this linguistic feature and for level 2 and 3 respectively the difficulty increases according to the means expressed in logit units). The total means show that lexical items are the easiest items, followed by syntactic items and the most difficult are the phonological items. The table revealed relatively high standard deviations across the data. Standard deviations on the items and tasks ranged from .219-1.409. Most standard deviations were about 1. This is considered high for a range of -3 to + 3 logit scale. The high standard deviations across items and tasks may be accounted for by the diversity of different levels of Spanish proficiency.

In order to determine if a significant difference or differences existed between the means of linguistic features of items and task levels, a two-way ANOVA was performed. Results appear in Table 4

Table 4 Two-way ANOVA results for comparing linguistic items and task levels.

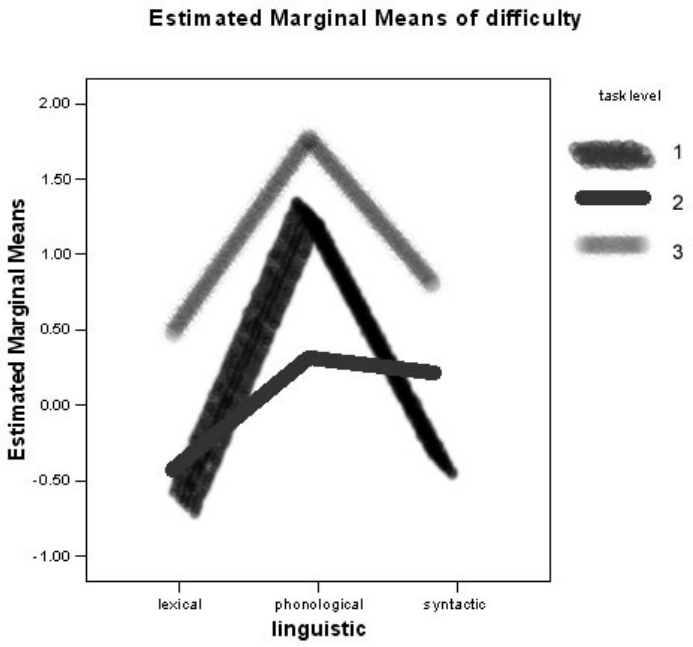
Variable & source	Df	MS	F	Sig.

Linguistic	2	5.242	4.029	.022
Task level	2	3.522	2.708	.073
Linguistic *Task level	4	.616	.474	.755
Error	73	1.301		

Table 4 shows the results of a two-way ANOVA performed to show differences in difficulty of three linguistic items belonging to three task levels. Previous analyses comparing task levels found no difference between levels 1 and 2, but significant differences were found between the other task level pairs. Table 4 shows no significant main effect either for task level ($F=2.708$, $p=.073$) or for the interaction of linguistic and task level ($F=.474$, $p=.755$). However, a significant main effect was present in the linguistic features ($F=4.029$, $p=.022<.05$). A Tukey post-hoc test showed that the sources of this significant difference were between different pairs of linguistic features: 1) between lexical and phonological features (mean difference = -1.6021, $p=.003$, $p<05$), and 2) between lexical and syntactic features (mean difference = -.6973, $p=.027$, $p<05$). No significant difference was found between phonological and syntactic features (mean difference = .9049, $p=.141$).

Figure 4 displays the profiles for the three task levels and three kind of linguistic items.

Figure 4 Linguistic and task level profile



There is a significant interaction between task levels 1 and 2, however no interaction was found between levels 1 and 3 and 2 and 3. The difficulty of the phonological items caused the interaction. This interaction, although unpredicted was possibly due to the difficulty that judges have classifying the code of linguistic features. Results of this analysis suggested that phonological items do not help as a discriminator of L2 listening performance on this test. Moreover, this result suggests that lexicon or syntax will help to discriminate student level.

2.2 Internal consistency

Four instructors rated the ten oral passages used as stimuli in the SLE using three different scales for obtaining content experts' opinions. In order to explore the consistencies among four raters across the ten oral passages coefficient alpha was calculated for the consistency of each scale. The scale for content relevance was about asking raters for level of agreement on linguistic topics included in the oral passages which were relevant to the Spanish textbooks used in

different Spanish courses. Raters also rated the oral passages based on the ACFTL scale for listening ability. Finally, the level of all listening tasks was rated according to a difficulty level scale. The difficulty scale include their raters ratings based on their viewpoint ranking the texts according to: the ACTFL guidelines, rate of delivery, fewer things or people in the text, familiarity with the topic, concrete or abstract content, and dialects versus standard Spanish (see Pardo-Ballester, 2007).

Overall, the inter-rater consistency for the ten oral passages of the SLE appears to be quite respectable (see Table 5).

Table 5 Internal consistency among four raters across oral passages

Scale	*Cronbach alpha*
For content relevance	.889
For ACTFL level	.924
For difficulty level	.939

3 Authenticity

The third claim of the AUA bears upon the generalization of the listening tasks to the classroom domain. Generalization is supported by the warrant about authenticity. This warrant states that a feature of the SLE was the similarity of tasks with the instructed domain. Test-takers' feedback was considered crucial in deciding about the satisfaction with the exam and to provide support for the intended use of placing students in Spanish courses using listening tasks which correspond to the language instruction domain. In addition, the SLE was intended to test how much Spanish students can understand orally outside of the book which is considered to be very important in order to be able to apply the language in the real world. Authenticity was explored by means of our participants' perceptions collected after the test by means of a survey. The

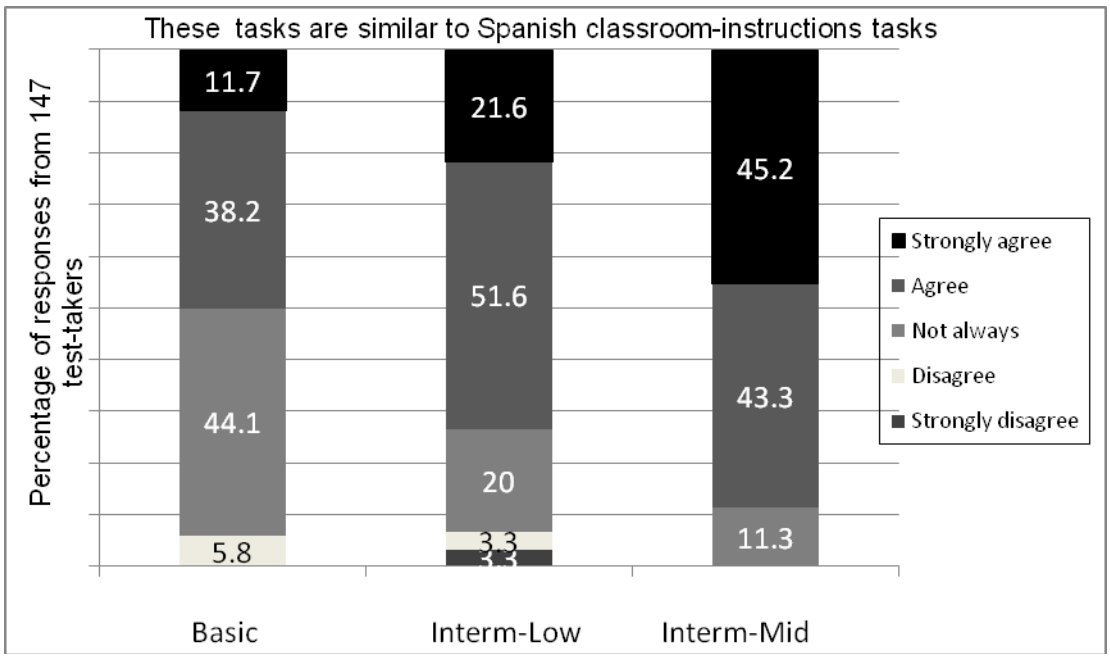
following students like the idea of being assessed through an online exam using similar tasks as those used during Spanish instruction, as well as using a variety of tasks, even though these qualities of the exam represents a challenge for some students:[3]

> "This is the same oral comprehension I've done in all my Spanish classes and it's pretty useful."
>
> "It is similar to the oral component we use when we take our midterms."
>
> "Because it makes you think + process information as you receive it. Simulates a real classroom or conversation"
>
> "B/c it all relates to what we are learning in class"
>
> "Online test shows how we are comprehending what we hear in the classroom & what we can understand"
>
> "It can see how well we understand different types of oral comprehension."
>
> "Yes because it has a variety of different scenarios. They were challenging too"
>
> "This program used a variety of situations instead of just one"
>
> "Because there are a variety of different tasks, of different levels of "easiness".
>
> "Many different types give everyone a chance to show what they know. Some were hard, but others were easier."
>
> "Because you can learn from different comprehension. You get to understand the different listening task, that are sometimes not done in class"
>
> "Some of the activities were easier than others therefore it is easy to see where my level of listening was"

In the post-test survey students have to select a choice in a Likert scale from 1 to 5 about the SLE tasks' similarity with the classroom-instruction tasks, with 1 being strongly disagree and 5 being strongly agree. Figure 5 shows the results in percentage of this statement.

[3] These opinions are as the students wrote them, misspelling are presented as written by students.

Figure 5 Test-takers' opinions on statement 'These listening tasks are similar to Spanish classroom-instruction tasks'



Most of the SLE tasks were for intermediate learners who are more used to the different types of tasks (i.e., dialogues, announcements and others) and can handle both a slow and a somewhat faster rate of speech. Figure 5 shows that higher percentages of the SLE similarity with classroom-instruction tasks are found in the 2 and 3 proficiency levels with "agree" (4) or "strongly agree" (5) answers than in level 1. Although 50% of beginners also agree or strongly agree with the tasks' similarity.

DISCUSSION AND CONCLUSION

This paper has investigated the procedures used to develop a Spanish listening exam and evaluated the claim that it distinguishes among examinees' Spanish listening proficiency pertaining to their classroom listening. The purpose of this exam was to assess the listening proficiency of L2 learners in Spanish. This exam was built to be used in a Spanish program in which students need to be grouped according to their levels of language ability before proceeding

with their instruction. In order to interpret the listening scores as a predictor of what examinees can do, the study included investigations of the test usefulness to support the warrants of the validity argument for a Spanish listening assessment. To ensure that the SLE had a satisfactory level of test usefulness, test development included investigations of internal consistency, IRT analyses, and ANOVA analyses for content coverage in order to support the construct validity, and feedback analyses for the post-test study. These investigations of the test usefulness served as backing evidence to justify the warrants and rebuttals of the validity argument. The validity argument investigated the validity of score-based inferences and uses framed in terms of three types of arguments that represent some of the SLE qualities of usefulness. Results of our investigations are represented in Table 6.

Table 6. Results of the test usefulness for SLE

Qualities	Backing for warrants	Backings for rebuttals
Reliability	-82 items were acceptable and therefore the .94 item reliability was accepted as having good consistency across items. -The Rasch reliability was .87 for 144 students.	-The reliability for persons did not seem consistent because three individuals did not fit the model. However, these three individuals were deleted in order to improve the accuracy of the Rasch model.
Construct Validity	-Two-way ANOVA suggested that lexical and syntactic features were the most helpful for discriminating among L2 listeners (see Figure 4 & Table 4). - Rasch difficulty estimates for the phonological items caused an interaction, which suggests that it would be better to avoid using phonological items altogether in listening tests (See Figure 4 & Table 4). -The rater consistency across different tasks, using different scales and four raters, was more than satisfactory (for content relevance the internal reliability was .88, for the ACTFL scale .92, and a reliability of .93 for difficulty level).The high alpha values for all scales suggest that the raters were in agreement with regard to the difficulty of the tasks	-Experts found difficulties when judging the items of the tests (see Pardo-Ballester, 2007).

	(see Table 5).	
Authenticity	-Participants perceived that the SLE listening tasks corresponded to the instructed domain of the first-and-second Spanish courses (See Figure 5). -Findings support the generalization of the SLE listening tasks to the classroom domain.	-Students did not complain about not using WebLAS during regular instruction.

From the findings we learned that test development is a difficult task that challenges test developers at all times. In this study, we investigated the procedures used to develop a listening exam in response to an evaluation of its usefulness.

The research questions for the validity argument investigated the validity of score-based inferences and uses frames in terms of three types of arguments that represent the SLE qualities of reliability/consistency, construct validity and authenticity.

As mentioned previously, the first claim in the AUA pertains to the scores that are obtained from test-takers' performance. The claim constituted the warrant for reliability/consistency in the validity argument which stated that the SLE scores were consistent across different groups of test takers. With this warrant the first research question was formulated as follow: To what extent are the scores that are obtained from test takers' performance consistent? Specifically: 1a. What is the reliability of the SLE? 1b. To what extent do the test-takers' levels of listening ability affect their item performance on the SLE?

Claim 1 was warranted based on the consistency warrant and despite the rebuttal findings of three persons whose responses did not fit the Rasch model. Consistency was improved by eliminating these outlying data points. Overall, the evidence suggests that test-takers' scores were mostly consistent across different groups of participants.

Rasch reliability for test scores across test-takers was estimated as backing for consistency of test scores. The Rasch reliability (equivalent of KR 20) was .87 for 144 students and for 82 items the reliability resulted in .94. These results are evidence of consistency across three proficiency levels as well as for different difficulty items. The results of IRT analyses served as rebuttal data to respond to the first question. The rebuttal data stated that running IRT analysis for misfitting items or persons could improve accuracy and reliability through identification of misfitting items and persons. Therefore, a way to strengthen reliability could be achieved by deleting items or persons that do not fit the model. Evidence gathered as rebuttal data included fit statistics by means of using WINSTEPS for data analysis.

Results of fit statistics indicated that all items were acceptable and therefore the .94 item reliability was accepted as a good consistency across items. However, the reliability for persons did not seem consistent because three individuals did not fit the model. Consistency was improved by eliminating these outlying data points. Results from the dimensionality check, using variance decomposition via principal components, supported the warrant that the test items comprised a linear unidimensional measurement. Moreover, the SLE was expected to have a wide spread of items and persons with different ability. Evidence from the IRT results guaranteed the variety of the item content, thereby enhancing the content validity of the SLE.

The second claim of the AUA pertains to the interpretation of the test scores. This claim addressed the following tripartite research question: To what extent can SLE scores be appropriately interpreted as indicators of Spanish proficiency? Specifically: 2a. To what extent the SLE is meaningful with respect to the construct to be assessed? 2b. What item features account for difficulty using the Rasch model? 2c. To what degree are the ratings of oral passages in this study consistent? The second claim was supported by two warrants for the construct

validity to investigate that the SLE scores were valid indicators of the SLE construct and not of other constructs. Content coverage and inter-rater consistency were gathered as backing for the warrants. In order to account for content coverage, grammatical knowledge was measured in the SLE. Significant differences were found when comparing the linguistic features among the items. Results about the inter-consistency across different tasks which were based on four different raters using three different scales were very consistent: With respect to the content relevance, internal reliability was .88, for the ACFTL scale .92 was obtained, and a reliability of .93 for difficulty level. The high alpha values for all scales suggest that the raters were in agreement with regard to the difficulty of the tasks.

The third claim stated that the SLE listening tasks were generalizable to the classroom domain. Generalizations were linked to the relevant placement decisions to be made when using classroom tasks. This claim was based on the authenticity warrant and addressed the following research question: To what extent are the interpretations about the listening ability generalizable to the instructional domain? Specifically: 3a) To what extend are the SLE tasks similar to the instructed domain according to our test-takers' perceptions? The authenticity quality looks at whether the listening tasks characteristics correspond to those of the classroom teaching tasks. By means of a post-test survey we learned that the SLE listening tasks were relevant to the classroom domain. Students agreed on the SLE similarity of tasks to the target language use domain as revealed by the high percentages of responses from 147 participants. The generalization claim was warranted based on high percentages of students' responses about the similarity of the test tasks with the classroom domain.

In sum, test-takers' opinions indicated that they perceived the listening tasks as corresponding to the instructed domain of the basic and intermediate Spanish courses. Students' perceptions about

authenticity indicated the variety in content and speech of listening tasks not only in different situations but also on the rate of speech and dialects that students can encounter in different Spanish courses and outside the classroom. Moreover, a high degree of correspondence between the characteristics of test tasks and the classroom tasks was perceived by students, as reflected in their opinions. Test-takers' perceptions about authenticity served as backing to support the interpretation about test-takers' listening ability based on the listening tasks which are relevant to the target language use domain. Despite the fact that online tasks are not used during midterms for these courses, the finding of authenticity supports the generalization of the SLE tasks to the classroom domain.

Last but not least, as mentioned in the introduction, in order to link the scores interpretation to the SLE use we need to include a utilization argument. Further development in investigating validity will investigate the SLE's appropriateness for its use. The utilization argument of the SLE fuses Toulmin's argument model with Messick's (1989) approach in which we state 1) claims about intended consequences of using the SLE and decision to make for using the test; 2) warrants in order to know if our test interpretations are relevant, meaningful, useful, sufficient, and beneficial to our participants; and 3) backing and rebuttals to support the intended use and investigate unintended and intended consequences of the SLE. Since it is the intention of the test developers and program administrator that the SLE is eventually to be used, investigations of the impact, usability and interactiveness qualities of usefulness, standard setting procedures, setting final cut scores and participants' feedback analyses will be included as backing evidence to support the SLE warrants of the utilization argument. To investigate the cut-off scores two different approaches will be presented in order to set standards for placing students in Spanish courses based on their listening proficiency level. The bookmark procedure

will examine the test content and the borderline-group method to evaluate the performance of test takers rather than items. In addition, the validity argument is linked and supported by the utilization argument in such a way that the relevance and utility evidence of the test use are related to the construct validity, while intended consequences and sufficiency are related to the intended purpose of the SLE.

REFERENCES

American Council for the Teaching of Foreign Languages. 1983. *ACTFL Proficiency Guidelines.* Revised 1985. Hastings-on-Hudson, NY: ACTFL Materials Center.

AUDACITY. (2001). Retrieved July 17, 2006 from http://audacity.sourceforge.net

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 1, 1-34.

Bachman, L. F. (2006). *Linking validity and use in educational assessments*. Paper presented in the National Council on Measurement in Education. April, San Francisco.

Bachman, L. F. & Palmer A. S. (1996). Language testing in practice. Oxford: Oxford University Press.

Buck, G. (2001). *Assessing listening*. Cambridge: Cambridge University Press.

Chapelle, C. A., Enright M.K.,& Jamieson J. (2008). Test score interpretation and use (1-25). *In Building a validity argument for the test of English as a Foreign Language*. New York, New York: Routledge.

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin*, 112, 527-535.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement*, 38, 319-342.

Kane, M. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice*, 21(1), 31-41.

Kane, M. (2004). Certification testing as an illustration of argument-based validation. *Measurement: interdisciplinary Research and Perspectives*, 2 (3), 135-170.

Kane, M., Crooks, T. J., & Cohen, A. S. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, 18 (2), 5-17.

Linacre, J.M. (2006). *Winsteps* (Version 3.61.2) [Computer software]. Chicago: Winsteps.com.

McNamara, T. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language testing,* 8, (2), 139-159.

McNamara, T. (2006). Validity in language testing: the challenge of Sam Messick´s legacy. *Language Assessment Quarterly* (3), 1, 31-51.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: Mcmillan.

Mislevy, Steinberg, & Almond. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives*, 1, 3-62.

Pardo-Ballester, C. (2007). *The Development of a Web-Based Spanish Listening Placement Exam*. Doctoral dissertation, University of California, Davis, CA.

Pardo-Ballester, C. (2008). The construct validity of a Web-based Spanish Listening Exam.

In C.A. Chapelle, Y.R. Chung, & J. Xu (Eds.). *Proceedings of the TSLL Fifth Annual Conference: Towards adaptive CALL: Natural language processing for diagnostic language assessment (pp. 209-227). Ames, IA: Iowa State University*.

SPSS (2003). SPSS version 12 for Windows. Chicago, IL: SPSS Inc.

Toulmin, S. E. (2003). *The uses of argument*, 2nd ed. Cambridge, UK: Cambridge University Press.

UCLA, Department of Applied Linguistics and TESL & Center for Digital Humanities. (2003). WebLAS (Web-based Language Assessment System). Retrieved July 17, 2006, from http://www.weblas.ucla.edu/

Weir, C. (2005). *Language testing and validation*. New York: Palgrave McMillan

APPENDIX A

For security of the SLE, the tasks and items presented here are taken from the pilot test, but they have a close resemblance to the actual test items/tasks used in the paper.

En un restaurante mexicano: Imagina que eres un mesero que trabaja en un restaurante. Escucha la conversación 2 veces y presta atención a lo que pide el cliente. Después contesta las preguntas.

1. ¿Cuál es la situación del cliente?
    a. no tiene hambre
    b. tiene prisa
    c. tiene hambre y prisa
    d. no tiene sed

2. El cliente pide la combinación de varios platillos.
    a. cierto
    b. falso

3. Para comer, ¿qué pide el cliente a la carta? (escribe una palabra en español)

4. ¿Qué va a beber el cliente? (Escribe 1 palabra en español)

5. ¿Al cliente le gusta el mole?

a. cierto
b. falso