JOURNAL OF CHEMICAL EDUCATION

Article

pubs.acs.org/jchemeduc

Validating Chemistry Faculty Members' Self-Reported Familiarity with Assessment Terminology

Mary Emenike,[†] Jeffrey R. Raker,[‡] and Thomas Holme*[,‡]

[†]Rutgers Learning Center, Rutgers University, New Brunswick, New Jersey 08901, United States

[‡]Department of Chemistry, Iowa State University, Ames, Iowa 50010, United States

Supporting Information

ABSTRACT: With the increasing emphasis placed upon chemistry instructors and departments to assess and evaluate their courses and curricula, understanding the structure of chemistry faculty members' knowledge and use of assessment terms and concepts can shed light on potential areas for targeted professional development. Survey research that might accomplish this objective often relies on self-reported responses from the target audience, and such information is sometimes difficult to assess in terms of validity. As an example of an internal mechanism to help establish validity, it is possible to include an "internal standard" item early in the survey. For the sake of understanding faculty members' familiarity with assessment terminology, an item that asked participants to identify analogous pairs of terms comparing assessment measures (*assessment validity* and *assessment reliability*) to laboratory measures (*accuracy* and *precision*) served this purpose. Using ordered logistic regression, participants who answered the analogy question completely correctly were more likely to report higher levels of familiarity with the assessment terms. Because the self-reported data appears to be valid, these data can be further used in subsequent analyses in order to determine the general familiarity trends among chemistry faculty regarding assessment terminology.

ANALOGOUS TERMS		Laboratory Measures: Accuracy	Laboratory Measures: Precision
Assessment Measures	Validity	☑	☒
	Reliability	☒	☑

KEYWORDS: *First-Year Undergraduate/General, Second-Year Undergraduate, Upper-Division Undergraduate, Chemical Education Research, Interdisciplinary/Multidisciplinary, Testing/Assessment*

FEATURE: Chemical Education Research

INTRODUCTION

It is increasingly apparent that all educators—including those for chemistry courses—face increased expectations related to assessment of student learning. Entire collections of anecdotes about how campuses have embraced outcomes assessment have been produced.[1] The idea that assessment should be learner centered,[2] for example, calls on college faculty to view assessment as more encompassing than merely obtaining student scores on content tests. Along with this idea, calls for assessment as a scholarly pursuit within academia are also increasing.[3] Some studies suggest faculty participation is lower among scientists than other disciplines.[4] Within chemistry education, Towns[5] reported a case study of several chemistry departments' assessment plans. The American Chemical Society's Examinations Institute (ACS-EI) reported on a needs assessment survey of chemistry faculty members from across the United States of America.[6,7] The overall goal of this survey was to measure the current state of chemistry faculty members' involvement with assessment efforts and knowledge of assessment terminology. Findings from this survey uncovered similar results related to departmental assessment efforts as reported by Towns's case study.[5]

In addition to research reports, recent editorials in this journal have emphasized the need for chemistry educators—not just chemistry education researchers—to undertake some of the responsibility for understanding the role of assessment in chemistry education. Pienta,[8] for example, described his observation that chemical educators appear to form a bimodal distribution on this issue, with some who value assessment significantly and others who appear to deny the possible role of enhanced assessment efforts to assist student learning. Bretz[9] has also addressed the challenges that might be faced by chemistry educators who are tasked with assessing student learning at the course and departmental level, noting specifically that for many college faculty members, new requirements for enhanced assessment efforts appear to require extensive learning of new techniques.

As Bretz[9] further suggested, chemistry faculty might not know certain terminology related to assessment. Jargon is the term for specialized use of words or expressions by a particular group that are or become difficult for others to understand. If faculty members consider conversations about assessment to include too much jargon, there is a chance that communication will break down and the endeavor to develop assessment plans might stall. On the flip side, Slevin[10] noted over a decade ago that there is more than the inclusion of jargon in the discourse of assessment that might discourage faculty. The exclusion of words such as *intellect, intelligence, imagination, wonder, contemplation, truth, inquiry,* and *collegiality* in common descriptions of assessment efforts tend to imbue a sense of

Published: August 21, 2013

distrust among faculty for demands that they somehow improve their assessment. Thus, it is possible that when it comes to assessment, the language that is used plays an outsized role in how worthwhile ideas are ultimately implemented. The most likely way to assess this premise is to conduct survey research with chemistry instructors and seek to measure their familiarity with assessment terms that may be considered jargon by some fraction of the target audience.

PURPOSE

As part of a larger needs assessment survey of chemistry faculty members whose implementation is described elsewhere,[6,7] the ACS-EI collected data on faculty members' familiarity with assessment terminology in order to determine the structure of this knowledge and whether specific terms were considered jargon. As is true for any survey instrument, steps must be taken to assess the validity of the findings. As an internal validation item for this self-report of term familiarity, a question was included early in the survey to measure participants' knowledge of two terms related to assessment. The findings reported herein serve to answer the research question: *What is the validity of chemistry faculty members' self-reported familiarity with assessment terminology*? With validity established, models to understand how faculty members parse their understanding of assessment terms can be constructed, and one such effort is reported elsewhere.[11]

ASSESSMENT TERMINOLOGY

Faculty members were asked to report their familiarity with 13 assessment terms (Figure 1). Brief working definitions are provided for these terms (Table 1), along with descriptive rationales for their inclusion in this survey. These terms were chosen based on results from focus groups that met during the development of the survey. While they are not an exhaustive listing of possible assessment terms, they were (i) the terms most often mentioned by education research specialists in their focus group sessions and (ii) hypothesized to provide a range of familiarity within the chemistry instructor pool sampled in this work. The 13 terms are grouped into four theoretical assessment categories: program assessment, instrument assessment, item assessment, and general statistics.

The program assessment terms (*formative, summative*, and *interim assessments*) were of interest because they describe different types of assessment that can be used to evaluate students' knowledge and skills for different purposes and because they are terms often associated with project evaluation. Two terms related to instrument assessment (*validity* and *reliability*) provide information about the accuracy and precision of an assessment instrument. Three item assessment terms (*item response theory—IRT, item difficulty*, and *item discrimination*) can be useful to chemistry faculty members who are interested in measuring their students' knowledge, abilities, and attitudes, among other criteria. The remaining terms, categorized as general statistics (*variance, ANOVA—analysis of variance, linear correlation coefficient, factor analysis*, and *Cronbach's alpha*) are used in assessment as well as in standard experimental analysis. These terms were included in this survey because of their importance in instrument design, measurement of student knowledge and characteristics, and other assessment measures (e.g., Cronbach's alpha measures assessment reliability).

Use the following scale to indicate how familiar you are with the terms presented in this table:

1. I have never heard this term before.
2. I have heard this term before but do not know what it means.
3. I have heard this term before but am not confident I know what it means.
4. I have heard this term before and have a sense of what it means.
5. I am completely familiar with this term and know what it means.

	1	2	3	4	5
Formative assessment	○	○	○	○	○
Summative assessment	○	○	○	○	○
Interim assessment	○	○	○	○	○
Assessment validity	○	○	○	○	○
Assessment reliability	○	○	○	○	○
Item response theory	○	○	○	○	○
Item difficulty	○	○	○	○	○
Item discrimination	○	○	○	○	○
Linear correlation coefficient	○	○	○	○	○
Cronbach alpha	○	○	○	○	○
ANOVA	○	○	○	○	○
Factor analysis	○	○	○	○	○
Variance	○	○	○	○	○

Figure 1. Survey question for reporting familiarity with assessment terminology.

METHODS

Survey Design and Participation

Details of the development process for the needs assessment survey and descriptions of the participant recruitment are provided elsewhere.[6,7] In a general sense, the survey elicits self-report appraisal for questions related to chemistry faculty members' views on assessment, experiences with assessment practices, the use and perceived usefulness of assessment tools (e.g., exams, student writing, clickers, laboratory reports), experiences with personal exams or ACS Exams, departmental assessment efforts, and familiarity with assessment terminology. Participants were divided into institutional categories: (i) chemistry doctoral granting institutions (doctoral); (ii) chemistry bachelor and master's degree granting institutions (four-year); and (iii) chemistry associate-degree granting institutions (two-year). Demographic data (e.g., sex, years teaching chemistry, chemistry subdiscipline) were collected at the end of the survey.

A key constraint for survey-based research lies in the nature of self-report data.[21,22] With regard to this survey, self-reported term familiarity of participants might be affected, for example, if they thought they should know about these terms. To check the validity of the self-reported familiarity in this survey, participants' knowledge of two terms related to assessment measurement (*validity* and *reliability*) was assessed using an analogy to laboratory measures (*accuracy* and *precision*; Figure 2).

Validity is a term in assessment measurements that is analogous to accuracy in physical measurement, that is, how well does the instrument measure what it is supposed to be measuring. There are several types of *validity* (e.g., face, construct, content), for which detailed information can be found elsewhere.[18−20]

Reliability, on the other hand, is analogous to precision in physical measurements. *Reliability* refers to how well the instrument will provide the same measurement if nothing has

dx.doi.org/10.1021/ed400094j | J. Chem. Educ. 2013, 90, 1130−1136

Table 1. Working Definitions for the Assessment Terms Investigated through the Familiarity Question in the Needs Assessment Survey (Figure 1)

Term	Working Definition
	Program Assessment[a]
Formative assessment	Formative assessment occurs during instruction and provides feedback both to teachers regarding their teaching practices and to students on their learning outcomes. Formative assessment is typically short in length with frequent occurrences to inform instructors of the current state of their students' understanding.
Summative assessment	Summative assessment occurs less frequently than formative assessment (at the end of a unit, semester, or year) and is used to evaluate student performance. Summative assessment is longer in length and broader in scope than formative assessment.
Interim assessment	Interim assessment falls between formative and summative assessment in frequency, scale, and scope. Interim assessment is used to evaluate students' knowledge and skills over a shorter time period than summative assessments.
	Instrument Assessment[b]
Assessment validity	Validity refers to how well the instrument measures what it is supposed to be measuring. This term is analogous to accuracy in laboratory measures.
Assessment reliability	Reliability refers to how well the instrument provides the same repeated measurements if nothing changed within the subject being measured. This term is analogous to precision in laboratory measures.
	Item Assessment[c]
Item response theory	Item response theory is a psychometric model that explains or predicts test performance based on latent traits or underlying abilities that are not directly observable.
Item difficulty	Item difficulty provides a measure of the item's difficulty. It is reported as the fraction of students who answer the item correctly (a higher difficulty value means an easier question).
Item discrimination	Item discrimination provides a measure of how well the item distinguishes between high and low performing students, as measured by the overall exam. Discrimination is calculated by taking the difference between the fraction of correct answers among top- and bottom-performing students, as measured by total scores on the exam.
	General Statistics[d]
Linear correlation coefficient	Linear correlation coefficient provides the degree of linear relationship or association between two variables.
Cronbach's alpha	Cronbach's alpha provides a measure of the internal consistency among a group of items on an instrument. Cronbach's alpha is often used to determine the reliability of the instrument.
ANOVA	Analysis of variance (ANOVA) compares the variance of means to determine whether statistical differences exist among two or more groups of data.
Factor analysis	Factor analysis is used as a data reduction tool to determine correlations among observed variables based on a small number of unobserved factors (i.e., latent variables).
Variance	Variance measures the probability distribution of the data about the mean.

[a]See ref 12. [b]See ref 13. [c]See refs 14−16. [d]See refs 17−20.

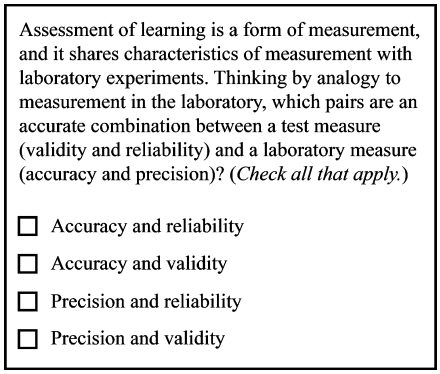

Figure 2. Survey question to validate self-report data using an analogy to laboratory measures.

changed within the subject being measured. It is important to establish an instrument's *reliability* so that any changes observed in the measurements can be attributed to changes in the subject and not an imprecise instrument.

Participants answered this analogy question early in the survey (question 4.2, Figure 2), while they answered the question on familiarity of the terms at the end of the survey (question 14, Figure 1). Placing these two questions far apart in the survey was intended to decrease the influence of answering the analogy question on the participants' reported familiarity to these two terms in a manner akin to a testing effect.[23]

Data Analysis

The majority of data collected through the needs assessment survey was binary (yes/no) or categorical. Because both binary and categorical responses are discrete variables, logistic regression was used to compare responses based on different groups of participants. Unlike standard linear regression models, logistic regression models relate the dependent and independent variables with a nonlinear logit function. For binary data with only two groups ("yes", $y = 1$, and "no", $y = 0$), the relationship is easier to interpret than for data in multiple ordered categories. Although the categorical data collected in the needs assessment survey included both ordered and unordered response categories, the analyses described herein will focus on ordered logistic regression because (i) the analogy question and familiarity response categories were ordered and (ii) binary logistic regression (BLR) analysis of this data set has been used and reported elsewhere.[7] Additional information about ordered logistic regression (OLR) is provided in the Supporting Information.

For the analyses described herein, the proportional odds OLR model was used, which assumes that the dependent variable represents an underlying continuous measure.[24] The proportional odds OLR model compares the probability of being at or below a certain point to the probability of being beyond that point.

$$\log\left[\frac{\Pr(y \le m|x)}{\Pr(y > m|x)}\right] = \tau_m - x\beta \qquad 1 \le m < M \tag{1}$$

or

dx.doi.org/10.1021/ed400094j | J. Chem. Educ. 2013, 90, 1130−1136

$$\Pr(y_i = m|x_i) = F(\tau_m - x_i\beta) - F(\tau_{m-1} - x_i\beta) \quad (2)$$

where F is the cumulative density function for the random error; β is the regression coefficient for the logit; τ indicates the cut-points or thresholds for each level; and the ordered categories are 1, 2, ..., m, ..., $M - 1$, M. In certain cases, statistical differences cannot be determined if the proportional odds assumption is violated; one typical cause of this violation results from having too few participants (<10%) in one of the categories.

Because of the nonlinear relationship in logistic regression, an odds ratio (OR) is typically calculated, which results in a linear relationship between the dependent and independent variables; the linear OR is consequently much easier to interpret than the β-coefficient. The OR is related to the β-coefficient in the above equations by the expression

$$\mathrm{OR} = e^{\beta} \quad (3)$$

For example, an odds ratio of 1.7 in proportional odds OLR would be interpreted as follows: the odds of a participant being in one category is 1.7 times greater than the odds of that participant being in any lower category. Data were analyzed using the statistical software package, Stata.[25,26] β-coefficients, ORs, and p values are reported for the statistics reported herein and in the Supporting Information.

RESULTS AND DISCUSSION

Basic Participant Demographics

The survey was completed by 1546 participants (35% female; 15 ± 9 average years of teaching experiences; 21% two-year, 51% four-year, and 28% doctoral). Detailed demographic information and statistical analyses of the sample of participants based on institution type, sex, number of years teaching, area of specialization, and use of ACS Exams have been reported elsewhere.[7] To validate the self-reported familiarity data reported herein, the entire database of participants was combined. Statistical analysis of the analogy question responses by demographic information are provided in the Supporting Information.

Analogy Question Responses

Of the four answer choices available for the analogy question (Figure 2), only two correspond to correct analogy pairs: *validity and accuracy* and *reliability and precision*. Because participants were asked to choose "all that apply", responses ranged from leaving the question blank (N = 99; 6%) to choosing one (N = 560; 33%), two (N = 828; 54%), or three options (N = 10; 0.6%) to choosing all four options (N = 49; 3%).

Participants were categorized into four ordered groups based on their answers to the analogy question (Table 2). The first (lowest) group (N = 99, 6%) consisted of participants who left the question blank. Participants in the "incorrect" category, the second lowest group (N = 455, 29%), included (i) those who chose three or four response options, (ii) those who chose two response options and one or both of those options were incorrect, and (iii) those who chose one response option and it was incorrect. The second highest group (N = 381, 25%) consisted of participants who chose one response only and it was one of the two correct options. Finally, participants who answered the question completely correctly—choosing only the two correct analogy pairs—comprised the final (highest) group (N = 611, 40%). While less than half of the survey participants answered the analogy question with both correct responses, an additional quarter of participants answered with a single correct response.

Table 2. Categories of Responses to the Comparison of Assessment Measures (Validity and Reliability) to Laboratory Measures (Accuracy and Precision)

Order of Answer Categories with Number of Responses (%) Total Responses, N = 1546			
Blank	Incorrect	One Correct	Two Correct
Lowest	→		*Highest*
99 (6)	455 (29)	381 (25)	611 (40)

Validating Self-Reported Familiarity Data with Analogy Question Responses

While the responses to the analogy question are interesting, the primary purpose of this analogy question was to validate the self-reported familiarity of assessment terms in the final question of the survey (Figure 1). Both *validity* and *reliability* were included in the list of terms; therefore, participants' reported familiarity with these two terms could be directly compared with their response categories in the analogy question.

For each analogy question response category, participants spanned the range of familiarity for these two terms. For the 604 faculty members who answered the analogy question completely correctly, 6% reported they have never heard either term before, while 22% reported they were completely familiar with these two terms. For the 442 faculty members who answered the analogy question incorrectly, 15% reported that they had never heard these terms before, while 10% reported complete familiarity with these two terms. These data suggest that a larger percentage of faculty members reported complete familiarity with these terms when they answered the analogy question correctly than when they answered it incorrectly. Conversely, a higher percentage of faculty members reported no familiarity with these terms when they answered the analogy question incorrectly than when they answered it correctly.

Another way of comparing the responses is to consider the correctness of the analogy question to the reported familiarity of the two terms (Table 3). For the 181 participants who reported they have never heard of the term *assessment validity*, only 23% answered the analogy item with two correct responses. By contrast, 57% of those participants who reported that they were completely familiar with the term *assessment validity* were in the two-correct category. Further parsing of these data is better considered visually, as provided in Figure 3, which shows the comparison of these data by percentages of participants' answers to the analogy question based on their reported familiarity with the term *assessment validity*. It is apparent that more participants who reported higher levels of familiarity with the term *assessment validity* answered the analogy question correctly than those participants who reported lower levels of familiarity with this term. The trend in responses is similar for *assessment reliability* (Table 3, not depicted in a figure).

OLR was used to compare the reported familiarity responses to the discrete categories of analogy question responses; details of all comparisons are provided in Table 4. For each assessment term, *validity* and *reliability*, participants who answered the

dx.doi.org/10.1021/ed400094j | J. Chem. Educ. 2013, 90, 1130−1136

Table 3. Categories of Responses to the Analogy Question Based on Reported Familiarity of Assessment Terms Validity and Reliability

Term	Familiarity Category	Total Responses	Answer Categories with Number of Responses (%)			
			Blank	Incorrect	One Correct	Two Correct
Assessment validity	Never heard	181	19(10)	68(38)	53(29)	41(23)
	Heard, but unknown	142	8(6)	50(35)	38(27)	46(32)
	Heard, but not confident	359	30(8)	110(31)	93(26)	126(35)
	Heard, sense of meaning	596	24(4)	168(28)	147(25)	257(43)
	Completely familiar	237	9(4)	48(20)	46(19)	134(57)
Assessment reliability	Never heard	177	16(9)	67(38)	51(29)	43(24)
	Heard, but unknown	144	8(6)	50(35)	37(26)	49(34)
	Heard, but not confident	353	26(7)	112(32)	89(25)	126(36)
	Heard, sense of meaning	598	30(5)	163(27)	151(25)	254(42)
	Completely familiar	243	10(4)	51(21)	49(20)	133(55)

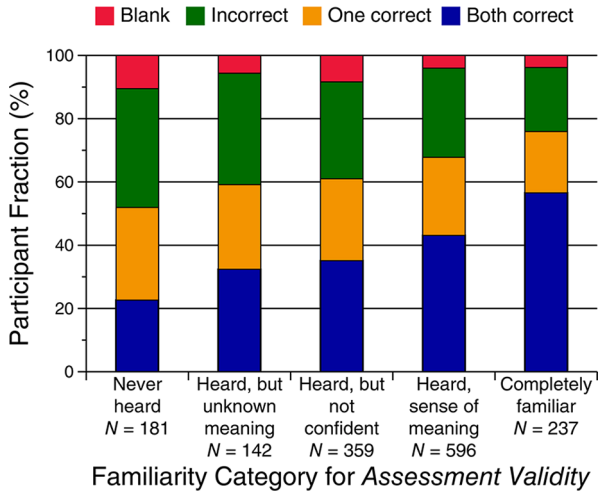

Figure 3. Distribution of participants' answers to the analogy question based on their familiarity with the term *assessment validity*.

analogy question completely correctly (checking the correct analogy and not checking the incorrect analogy) are generally roughly twice as likely to report familiarity with these terms in the self-report section of the survey. Nearly all remaining comparisons between groups show no statistically significant difference in the reported familiarity with either term with one exception. Those who correctly answer only the *validity analogy* item report higher familiarity than those who left the analogy items blank. (Table 4, β = 0.528; OR = 1.7, p = 0.015).

Because the OLR analysis shows only one situation where the participants who answered only the *assessment validity* analogy differ from participants who answered only the *assessment reliability* analogy question correctly, the designation of the "one-correct" category appears to model the data adequately. In other words, in nearly all cases, the odds of reporting higher familiarity with the terms *assessment validity* or *assessment reliability* was not statistically significantly different among faculty members who answered only one analogy relationship correctly, who answered the analogy question incorrectly, or who left the question blank. Thus, as an internal reference for the self-report data, the analogy item serves the function of validating that faculty members who identify these terms accurately also report higher levels of familiarity with the terminology. Anything less than complete correct responses on the analogy items results in lower reports of familiarity, precisely as would be expected.

While Figure 3 provides a summary view of the self-reports of familiarity among the different groups, the visualization does not particularly emphasize this comparative concept. Figure 4 yields more easily visualized comparisons by plotting reported familiarity (from Table 3) of the participants as the cumulative percentage for each group defined by the analogy question. A more detailed description of the cumulative distribution function is provided in the Supporting Information. This visual

Table 4. Ordered Logistic Regression Analysis Comparing Levels of Reported Familiarity among Groups of Participants Based on Their Answers to the Analogy Question

Participants Who Reported		Results by Assessment Term, β (OR)	
Higher Familiarity	Lower Familiarity	Assessment Validity	Assessment Reliability
Answered both analogy questions correctly	Answered only one analogy question correctly	0.628[a] (1.9)	0.528[a] (1.7)
	Answered the analogy question incorrectly	0.741[a] (2.1)	0.697[a] (2.0)
	Did not answer the analogy question (left question blank)	1.080[a] (3.0)	0.806[a] (2.2)
Answered only *validity* analogy question correctly	Answered only *reliability* analogy question correctly	*b*	*b*
	Answered the analogy question incorrectly	*b*	*b*
	Did not answer the analogy question (left question blank)	0.528[c] (1.7)	*b*
Answered only *reliability* analogy question correctly	Answered only *validity* analogy question correctly	*b*	*b*
	Answered the analogy question incorrectly	*b*	*b*
	Did not answer the analogy question (left question blank)	*b*	*b*
Answered only one analogy question correctly	Answered the analogy question incorrectly	*b*	*b*
	Did not answer the analogy question (left question blank)	*b*	*b*
Answered the analogy question incorrectly	Did not answer the analogy question (left question blank)	*b*	*b*

[a] $p < 0.001$. [b] Not statistically significant. [c] $p = 0.015$.

dx.doi.org/10.1021/ed400094j | J. Chem. Educ. 2013, 90, 1130−1136

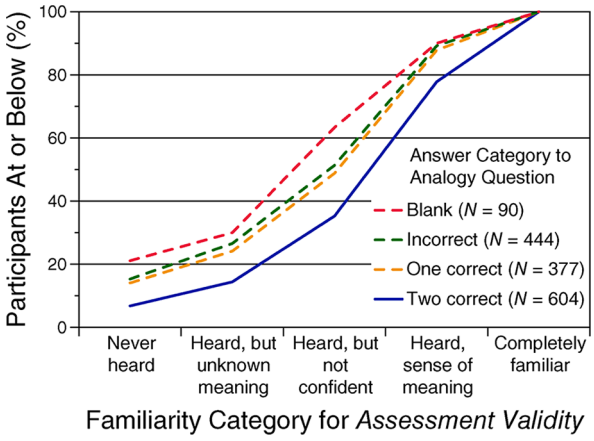

Figure 4. Cumulative percentages of familiarity with term *assessment validity* based on answers to the analogy question (the group of participants indicated by a solid blue line are statistically significantly different from groups of participants indicated by dashed lines).

representation highlights the similarity among participants who left the analogy question blank, answered it incorrectly, or answered only one pair correctly. Moreover, these groups of participants reported distinctly lower levels of familiarity than those participants who answered the analogy question completely correctly (keep in mind that reporting lower levels of familiarity will result in higher plots on the graph).

The OLR analyses suggested that there were two groups of participants based on their reported familiarity: those who answered the question completely correctly (blue line in Figure 4) and those who did not (dashed lines in Figure 4). The self-reported familiarity of faculty members who answered both analogy pairs correctly was statistically different from all other participants. The self-reported familiarity of faculty members who answered only one analogy pair correctly was not statistically different from those who answered it incorrectly or left the question blank. Although the self-reported familiarity was not perfect (i.e., not all faculty members who reported complete familiarity with the terms answered the analogy question completely correctly), the self-reported data appeared to follow a trend that participants who reported higher levels of familiarity answered the analogy question more correctly.

Trends for the Remaining Assessment Terms

The trend of greater familiarity was consistent for nearly all 13 terms ($p \leq 0.05$) for participants who answered the analogy question completely correctly relative to participants who did not. However, the terms *item difficulty*, *factor analysis*, and *variance* showed no statistically significant differences between the faculty members who answered the analogy question completely correctly and those who left it blank. In addition, statistical differences between these two groups of participants could not be determined for the terms *item discrimination*, *Cronbach's alpha*, and *ANOVA* because the proportional odds assumption was violated (too few participants in at least one "familiarity" category, see the Supporting Information). Statistical differences could also not be determined—for the same reason—between participants who answered the analogy question correctly and those who answered it incorrectly for the terms *item response theory* and *Cronbach's alpha*.

While the analogy question only included two terms for directly validating the self-reported data, these two terms were useful for indirectly validating some other terms in the list of assessment jargon. The four terms that could not be indirectly validated had too few participant responses in one or more familiarity category because of two reasons. Three terms (*item discrimination*, *item response theory*, and *Cronbach's alpha*) had too few participant responses in high familiarity categories, which indicated that chemistry faculty members were not highly familiar with these terms. The term *ANOVA*, on the other hand, had too few participant responses in low levels of familiarity, which indicated that chemistry faculty members were highly familiar with this term. Indeed, we would expect chemistry faculty members to be quite familiar with *ANOVA* as it is a common statistical technique used in almost all subdisciplines of chemistry.

CONCLUSIONS

The addition of a survey question that measured the content knowledge of two terms related to assessment allowed for (i) determination of chemistry faculty members' understanding of these two terms and (ii) comparison of self-reported familiarity to this measured understanding. The range of responses to the two aspects of the *select-all-that-apply* validation question suggested there might be degrees of understanding related to these two terms. Responses ranged from leaving the question blank, to answering incorrectly, to answering one analogy pair correctly, to answering both analogy pairs correctly.

Yet when participants' self-reported familiarity with the two assessment terms were analyzed based on their answer category to the analogy question, participants who answered completely correctly (i.e., correctly matching each assessment term with its one analogous laboratory term) were found to report statistically higher familiarity than participants who answered in any other category. Furthermore, there was no statistical difference among the self-reported familiarity levels of participants in any of the not completely correct analogy question categories. Therefore, it appears that the participants' self-reported familiarity with assessment terminology is consistent with their knowledge of the key terms *validity* and *reliability*. Findings were similar for reported familiarity with the other 11 terms related to assessment. While participants can always choose to overrepresent or underrepresent their responses with self-reported data, the data summarized here suggest that the reported familiarity with the assessment terms accurately represents the current state of assessment terminology familiarity for chemistry faculty members. This structure of chemistry faculty members' knowledge related to assessment will be analyzed through structural equation modeling using this validated self-reported data.[11]

ASSOCIATED CONTENT

Supporting Information

Discussion of the use of cumulative density functions, summary of demographic information of the survey sample, breakdown of responses to the internal validation, analogy items on the survey as a function of demographic variables, complete breakdown of the relationship between self-reported familiarity for all assessment terms and correctness level of the internal validation item, and odds ratios from ordered logistic regression for all 13 assessment terms as related to correctness on the internal validation item. This material is available via the Internet at http://pubs.acs.org.

dx.doi.org/10.1021/ed400094j | J. Chem. Educ. 2013, 90, 1130−1136

AUTHOR INFORMATION

Corresponding Author

*E-mail: taholme@iastate.edu.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

The authors thank Kristen Murphy and Jacob Schroeder for their work in the development of the survey that resulted in the database of responses that are analyzed here and Kimberly Linenberger for contributing to the comparison statistics between sample participants and the population from our national database of chemistry faculty members and for providing comments on this report. This work was supported by NSF DUE No. 0920266; any opinions, findings, conclusions and/or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of the National Science Foundation (NSF).

REFERENCES

(1) Maki, P. L. *Coming to Terms with Student Outcomes Assessment*; Stylus Publishing: Sterling, VA, 2010.
(2) Webber, K. L. The Use of Learner-Centered Assessment in U.S. Colleges and Universities. *Res. Higher Educ.* **2012**, *52*, 201−228.
(3) Wang, X.; Hurley, S. Assessment as a Scholarly Activity? Faculty Perceptions of and Willingness to Engage in Student Learning Assessment. *J. Gen. Educ.* **2012**, *61*, 1−15.
(4) Yanowitz, K. L.; Hahs-Vaughn, D. L. Changes in Student-Centered Assessment by Postsecondary Science and Non-Science Faculty. *Teach. Higher Educ.* **2007**, *12* (2), 171−184.
(5) Towns, M. H. Developing Learning Objectives and Assessment Plans at a Variety of Institutions: Examples and Case Studies. *J. Chem. Educ.* **2010**, *87* (1), 91−96.
(6) Emenike, M. E.; Schroeder, J.; Murphy, K.; Holme, T. A. A Snapshot of Chemistry Faculty Members' Awareness of Departmental Assessment Efforts. *Assess. Update* **2011**, *23* (4), 1−2−14−16.
(7) Emenike, M. E.; Schroeder, J.; Murphy, K.; Holme, T. A. Results from a National Needs Assessment Survey: A View of Assessment Efforts within Chemistry Departments. *J. Chem. Educ.* **2013**, *90* (5), 561−567.
(8) Pienta, N. Striking a Balance with Assessment. *J. Chem. Educ.* **2011**, *88* (9), 1199−1200.
(9) Bretz, S. L. Navigating the Landscape of Assessment. *J. Chem. Educ.* **2012**, *89* (6), 689−691.
(10) Slevin, J. F. Engaging Intellectual Work: The Faculty's Role in Assessment. *Coll. Engl.* **2001**, *63*, 288−305.
(11) Raker, J. R.; Emenike, M. E.; Murphy, K. L.; Holme, T. A. Using Structural Equation Modeling To Understand Chemistry Faculty Familiarity of Assessment Terminology: Results from a National Survey. *J. Chem. Educ.* **2013**, DOI: 10.1021/ed300636m.
(12) Perie, M.; Marion, S.; Gong, B.; Wurtzel, J. The Role of Interim Assessments in a Comprehensive Assessment System: A Policy Brief, The Aspen Institute, 2007. http://www.aspeninstitute.org/sites/default/files/content/docs/education/ed_PolicyBriefFINAL.pdf (accessed Aug 2013).
(13) Joppe, 2000, in Golafshani, N. Understanding Reliability and Validity in Qualitative Research. *Qual. Rep.* **2003**, *8* (4), 597−607.
(14) Hambleton, R. K. Item Response Theory: Introduction and Bibliography. *Psicothema* **1990**, *2* (1), 97−107.
(15) Sevenair, J. P.; Burkett, A. R. Difficulty and Discrimination of Multiple-Choice Questions: A Counterintuitive Result. *J. Chem. Educ.* **1988**, *65* (5), 441−442.
(16) Holme, T.; Murphy, K. Assessing Conceptual and Algorithmic Knowledge in General Chemistry with ACS Exams. *J. Chem. Educ.* **2011**, *88* (9), 1217−1222.
(17) Everitt, B. S. *The Cambridge Dictionary of Statistics*, 2nd ed.; Cambridge University Press: New York, 2002.
(18) Abel, S. K.; Ledermen, N. G. *Handbook of Research on Science Education*; Lawrence Erlbaum Associates, Inc.: Mahwah, NJ, 2011.
(19) Creswell, J. W. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches*, 3rd ed.; Sage Publications, Inc.: Los Angeles, CA, 2009.
(20) Kelly, A. E.; Lesh, R. A. *Handbook of Research Design in Mathematics and Science Education*; Lawrence Erlbaum Associates: Mahwah, NJ, 2000.
(21) Mayer, D. P. Measuring Instructional Practice: Can Policymakers Trust Survey Data? *Educ. Eval. Policy Anal.* **1999**, *21*, 29−45.
(22) Ryan, K.; Gannon-Slater, N.; Culbertson, M. J. Improving Survey Methods with Cognitive Interviews in Small- and Medium-Scale Evaluations. *Am. J. Eval.* **2012**, *33*, 414−430.
(23) Roediger, H. L.; Karpicke, J. D. The Power of Testing Memory: Basic Research and Implications for Educational Practice. *Perspect. Psychol. Sci.* **2006**, *1* (3), 181−210.
(24) Fullerton, A. S. A Conceptual Framework for Ordered Logistic Regression Models. *Sociol. Methods Res.* **2009**, *38*, 306−347.
(25) Long, J. S.; Freese, J. *Regression Models for Categorical Dependent Variables Using Stata*, 2nd ed.; Stata Press: College Station, TX, 2006.
(26) StataCorp. *Stata Statistical Software: Release 11*; StataCorp LP: College Station, TX, 2009.