CBE—Life Sciences Education
Vol. 14, 1–11, Summer 2015

Article

The Faculty Self-Reported Assessment Survey (FRAS): Differentiating Faculty Knowledge and Experience in Assessment

David I. Hanauer* and Cynthia Bauerle†

*English Department, Indiana University of Pennsylvania, Indiana, PA 15705-1094; †Undergraduate and Graduate Science Education Programs, Howard Hughes Medical Institute, Chevy Chase, MD 20815-6789

Submitted October 14, 2014; Revised January 7, 2015; Accepted February 23, 2015
Monitoring Editor: Marshall Sundberg

Science, technology, engineering, and mathematics education reform efforts have called for widespread adoption of evidence-based teaching in which faculty members attend to student outcomes through assessment practice. Awareness about the importance of assessment has illuminated the need to understand what faculty members know and how they engage with assessment knowledge and practice. The Faculty Self-Reported Assessment Survey (FRAS) is a new instrument for evaluating science faculty assessment knowledge and experience. Instrument validation was composed of two distinct studies: an empirical evaluation of the psychometric properties of the FRAS and a comparative known-groups validation to explore the ability of the FRAS to differentiate levels of faculty assessment experience. The FRAS was found to be highly reliable ($\alpha = 0.96$). The dimensionality of the instrument enabled distinction of assessment knowledge into categories of program design, instrumentation, and validation. In the known-groups validation, the FRAS distinguished between faculty groups with differing levels of assessment experience. Faculty members with formal assessment experience self-reported higher levels of familiarity with assessment terms, higher frequencies of assessment activity, increased confidence in conducting assessment, and more positive attitudes toward assessment than faculty members who were novices in assessment. These results suggest that the FRAS can reliably and validly differentiate levels of expertise in faculty knowledge of assessment.

INTRODUCTION

> I think it's a real gorilla in the room, in that it's very difficult to try to assess what you're doing, you need very specific goals and objectives and then how do you know you're meeting them?
>
> —Faculty assessment workshop participant

CBE Life Sci Educ June 1, 2015 14:ar17
DOI:10.1187/cbe.14-10-0169
Address correspondence to: Cynthia Bauerle (bauerlec@hhmi.org).

Backdrop of Undergraduate STEM Educational Reform

The national call for undergraduate science, technology, engineering, and mathematics (STEM) education reform focuses on evidence-based teaching as the standard for faculty practice (President's Council of Advisors on Science and Technology, 2012). Evidence-based teaching implies not only that STEM educators draw on the educational literature to select and apply instructional practices that will lead to student success but also that they attend to student outcomes by gathering evidence in their own classrooms about what their students are learning. Outcomes assessment has thus emerged as a central theme of the national discourse about how to improve student learning in STEM fields. Increasing awareness and support for assessment development, especially among STEM faculty members with little prior formal training, may be understood as an important driver of undergraduate STEM education reform (Hanauer *et al.*, 2009; Hanauer and Bauerle, 2012).

The standard of "scientific teaching," approaching the practice of teaching in the same way scientists approach research, has set the tone for STEM education reform (Handelsman *et al*., 2007). The recent American Association for the Advancement of Science (AAAS) *Vision and Change in Undergraduate Life Science Education* report outlines a curricular framework of core concepts and competencies that biology students should learn (AAAS, 2011). Similarly, the 2009 Association of American Medical Colleges (AAMC)–Howard Hughes Medical Institute (HHMI) *Scientific Foundations for Future Faculty* report outlines student competencies for premedical education (AAMC, 2009). These frameworks for undergraduate life sciences are in close alignment with each other and also parallel curricular frameworks developed for secondary biology education (College Board, 2012; National Research Council, 2012). Thus, science education is moving away from traditional content-focused approaches and being reframed in terms of demonstrable student skills and habits of mind.

As faculty members transform STEM curricula to be more focused on supporting student development of core competencies, they must also adapt their approaches toward learning assessments appropriate to this context. While certain practices, such as summative assessment tests typically used at the end of a unit or course, are well established as an integral part of STEM teaching, faculty members may be much less familiar with integrated and ongoing formative assessment strategies that provide critical real-time feedback to students in support of their learning (Shavelson *et al*., 2008; Yin *et al*., 2008; Hanauer *et al*., 2009; Hanauer and Bauerle, 2012). Formative assessment strategies are powerful tools for improving teaching effectiveness and student learning outcomes, yet many STEM faculty members have little experience with these practices and may be reluctant to incorporate them into their teaching.

As STEM departments consider how best to improve the success of their students, it is important to consider faculty knowledge and attitudes toward assessment (Bauerle and Hanauer, 2013). Deciding how to engage faculty members in departmental conversations about assessment or how to design programs of support that respond to faculty need is complicated by the varied experiences and knowledge that STEM faculty members bring to the table. Faculty members may be unfamiliar with the literature on learning outcomes or may not have a clear understanding of assessment terms and strategies. Faculty members who do not feel confident in their ability to apply assessment practices may be reluctant to try new approaches in the classroom. Because assessment is a topic that often comes up in a charged context (e.g., faculty evaluation or accreditation), the term itself may carry a negative connotation for faculty members, obscuring the potentially positive impact assessment practices can have on teaching practice (Schilling and Schilling, 1998).

Faculty Perceptions of Assessment

Previous work has examined faculty responses to assessment using attitudinal or knowledge-based approaches. Schilling and Schilling (1998) used a qualitative approach to capture the reasons for faculty resistance to assessment. Their report documents faculty perceptions of assessment as redundant, irrelevant, unreliable, intrusive, time-consuming, bureaucratic, and an imposition on academic freedom. Wang and Hurley (2012) used a quantitative approach to look at attitudes, perceptions, and needs of liberal arts college faculty regarding assessment. They found that faculty resistance is linked to a perception that assessment is not valued as a scholarly activity.

The American Chemical Society (ACS) initiated a large-scale, self-reporting survey study of department assessment efforts, including types of assessment used, motivation to use assessment, and faculty roles in departmental assessment initiatives (Emenike *et al*., 2013a). They found that, beyond the standard practice of administering summative exams, time demands on grading are a primary deterrent to faculty engagement with additional assessment practices. Furthermore, faculty members report their engagement in assessment practice is largely extrinsically motivated, that is, in response to requirements imposed for accreditation or certification purposes. A follow-up to the study by Emenike and colleagues focused more specifically on individual faculty members self-reported degree of familiarity with assessment terms (Raker and Holme, 2014). Results reflected that different levels of knowledge among participants were related to demographic variables such as institution, subdiscipline and teaching experience, and individuals' prior knowledge of statistics (Raker and Holme, 2014). In this study, validity of self-reporting familiarity of assessment vocabulary was confirmed using an internal validity approach.

Context for the Current Study

In our efforts with faculty members engaged in competency-based curricular development in undergraduate life sciences education, assessment development has emerged as a central focus for faculty-development work. In one selective, multi-institutional curricular design initiative we supported, faculty members developed novel interdisciplinary modules organized around core scientific competencies in the *Scientific Foundations for Future Faculty* report and considered how to develop strategies for measuring competency gains. Program participants found that standard content-focused assessments were insufficient to measure student learning gains specified in their competency-based approaches and that end-of-unit summative tests did not accommodate real-time assessment students need in support of their learning. Instead, faculty members needed to develop assessments specific to the educational goals of their competency-based curricula. To support this work, we engaged faculty members in a multiyear assessment-development program that served to develop a common understanding and vocabulary among participants and provided a collective context for them to develop effective assessment approaches specific to their curricular projects. Given the broad array of faculty experience around assessment, we were motivated to find a way to gauge faculty knowledge, attitudes, and prior assessment experience.

We describe a novel survey instrument, the Faculty Self-Reported Assessment Survey (FRAS), which expands on previously described tools in a comprehensive design that collects information about faculty knowledge, attitudes, confidence, and practices. Survey design is informed by principles of active assessment, a strategy developed for understanding student-demonstrated learning in STEM contexts, and thus is purposely aligned with faculty goals

within a competency-focused curricular context (Hanauer *et al.*, 2009). The FRAS was constructed for use in the context of a development model in which faculty members build their capacity for assessment practice through integration with competency-based curriculum development. Thus, it is specifically designed as a tool for faculty members engaged in developing the skill sets they need to respond to priorities and practices recommended in multiple national reports on undergraduate STEM education reform. The FRAS is based on a "self-reporting of familiarity" approach and includes four distinct sections that query assessment knowledge, practices, confidence, and attitude. A list of assessment-related terms is expanded from previous studies and is ordered into categories to identify specific assessment areas of interest or gaps in knowledge among a survey population. Specifically, survey design reflects categories of "assessment program design," "assessment instruments and scoring systems," and "processes of validation." In addition, the FRAS queries engagement with assessment based on self-reported frequency of conducting different assessment practices and self-reported attitudinal responses. It also includes a confidence-level scale to measure the degree to which faculty members feel able to conduct different assessment activities. The FRAS is intended as an assistive tool for providing useful data about assessment knowledge and experience that can identify shared areas of faculty interest and need, inform development programs or departmental conversations about student learning outcomes, and guide the development of shared language and understanding about learning-assessment practices.

In this paper, we report on our study of the psychometric properties of the FRAS instrument, in which we formally addressed two research questions:

- Does the FRAS function in a valid and reliable way?
- Does the FRAS distinguish between novice and advanced survey populations in a meaningful way?

METHODS

Development and Design of the Faculty Assessment Survey

Extending previous research that had looked at faculty knowledge of assessment, the FRAS was designed to have four sections: knowledge of assessment terms, frequency of assessment practice, confidence in conducting assessment activities, and positive/negative attitudes toward assessment. Participants self-report their responses according to standard five-point Likert scale adapted for each section of the instrument. The first of these sections is a self-report of vocabulary items. This particular list of vocabulary items was based on prior work by the ACS and reported by Emenike *et al*. (2013b). But the original list was extended to include a wider range of terms specifically relating to issues of formative assessment and assessment as a development process. The FRAS organizes assessment knowledge and practice into four functional categories: program design, instrumentation, scoring systems, and processes of validation. The underpinning assumption of asking about familiarity with assessment vocabulary is that this reflects levels of exposure and ability in terms of assessment. The validity of this assumption was partially evaluated by Emenike *et al*. (2013b) by regressing vocabulary item response with an assessment expertise analogy task. Results from their study suggested a relationship between higher familiarity and demonstrated expertise on the analogy task.

The other sections of the survey were motivated by findings from previous research of the importance of faculty members' attitudes about and motivation toward assessment. Thus, the FRAS was designed to query attitude and sense of confidence and also included a frequency of practice scale to access the degree to which assessment work was being conducted. Together, the four sections of the survey offer a comprehensive view of faculty assessment understanding, attitudes, confidence, and frequency of practice, and should allow distinction between faculty members with different assessment experience. This design facilitates an understanding not only of explicit assessment knowledge but also psychological responses to assessment, such as confidence and attitude, and behavioral data about frequency and types of faculty practice. By covering different assessment categories and providing information on conceptual, psychological, and practical aspects, the FRAS is designed to serve as a comprehensive platform for evaluating faculty experience and goes beyond previous instruments, which looked at faculty understanding and ability in terms of assessment. The Supplemental Material presents a full version of the FRAS.

Study Design

The overall validation design was composed of two distinct evaluations. The first involved evaluating the psychometric properties of the FRAS. Science education faculty members completed a full version of the survey, and this was then subjected to statistical analysis of its psychometric properties. The second evaluation consisted of a direct comparison between two groups of faculty members defined according to their level of assessment knowledge and experience. The design of this study was reviewed and approved by the Institutional Review Board at Indiana University of Pennsylvania (IRB log no. 13-257). The study was conducted in full accordance with the guidelines of the approved IRB protocol.

Participants

The participants for the first study were 95 STEM faculty members at institutions participating in a large educational program funded by a private science organization. The participants all teach at the undergraduate level and were from 90 different institutions. To maintain confidentiality and avoid any negative implications, we collected no identifying information that might enable researchers or funders to directly or indirectly identify faculty participants used in this study.

The participants for the second study were drawn from two different STEM faculty programs supported by a private science organization. All participants were experienced faculty members with years of STEM teaching experience and thus have likely had informal exposure to assessment knowledge and practice in their regular educational settings. Faculty members in the first group (n = 11) participated in a formal 3-yr assessment-development program that included educational workshops designed to enhance assessment knowledge and support implementation of assessment

practice in the science classroom. For the purposes of this study, this group was defined as having "advanced" assessment knowledge and experience. Faculty members in the second group ($n = 17$) were enrolled in but had not yet begun a formal introductory assessment-development workshop. This group was defined as having "novice" levels of assessment knowledge and experience. As with the first study, no personal identifying data were collected.

Data Analysis for Survey Development

Following accepted validation procedures, analysis involved empirically evaluating the reliability, dimensionality, and validity of this new tool before its distribution for more applied usages (Netemeyer *et al*., 2003; Bachman, 2004). For the evaluation of the psychometric properties of the FRAS, the following analytical and statistical procedures are reported in the next section. Dimensionality of the vocabulary familiarity section of the FRAS was evaluated using a principal components analysis (PCA) approach (Cattell, 1978). Dimensionality was further explored using a focused correlational approach looking at the relationships among specific sections of the vocabulary items and frequency of usage and confidence scales.

For reliability, Cronbach's alpha was calculated to establish the internal consistency of the whole survey and each of its subsections. For validity, a known-groups approach utilized two groups of STEM faculty members, one composed of STEM faculty members with significant formal assessment knowledge (advanced) and a second made up of STEM faculty members with limited formal assessment knowledge (novice). If the FRAS has the ability to differentiate levels of faculty assessment knowledge, then systematic differences between these two groups should emerge.

RESULTS

Dimensionality

The dimensionality of the survey was primarily evaluated using a PCA approach. The aim of this analysis was to provide some understanding of the internal structure of the survey by empirically specifying those items that group together based on an underpinning intercorrelational relationship. As specified in the previous section, this tool comprises four different sections: self-reported knowledge of assessment vocabulary (knowledge section), self-reported frequency of engaging in different assessment practices (practice section), self-reported confidence in conducting different assessment activities (confidence section), and attitude toward different assessment activities (attitude section). The limited number of participants ($n = 105$) and large number of items meant that a full PCA could not be conducted on the whole of the survey (Hair *et al*., 1979; MacCallum *et al*., 2001; de Winter *et al*., 2009). Accordingly, a PCA approach was applied to explore the internal structure of the most controversial aspect of the survey, the self-reported familiarity with assessment vocabulary items. A PCA on this section of the survey offers some understanding of the groupings of items and, by inference, the underlying structure of knowledge concerning assessment.

As a first stage in the PCA, the assumption of correlational interrelationships of at least 0.3 and above were checked in the correlation matrix. On this initial stage of the analysis, several items were found to have low and negative correlations with other items and were accordingly removed from the analysis, and the PCA was repeated without these items. The vocabulary items that were removed were "parallel form reliability," "checklist," "Cronbach's alpha," "multiple-choice question," "cloze," and "true/false." All remaining items met the criteria of at least a 0.3 correlation, suggesting that a PCA can be used as a suitable analytical procedure. In a reiteration of the PCA, the assumption of sufficient sampling size was tested. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy (MSA) was calculated, and the result (KMO = 0.84) was well above the 0.5 benchmark, suggesting a sufficient sample size for this analysis. Additionally, an MSA for each vocabulary item on the scale was computed, and results for all included items were from 0.77 to 0.94, indicating adequate sizes for each of item.

PCA with direct oblimin rotation with Kaiser normalization was conducted to examine the internal structure of faculty self-reported knowledge of assessment. For determining the number of components to include, a trianalytical approach was utilized. This included the Kaiser criterion of keeping only components with eigenvalues above 1, graphing and visual analysis of the scree plot of eigenvalues, and parallel analysis comparison of component eigenvalues with Monte Carlo procedures randomly generated eigenvalues. The Monte Carlo procedure functions as a null hypothesis and provides threshold levels for eigenvalues that can be used above chance at the 0.05 significance level. As can be seen in Table 1, comparison of eigenvalues from current study data and the Monte Carlo simulation revealed that only two components were above the chance level, suggesting a two-component solution. The Kaiser criterion by itself allowed the inclusion of four possible components. The scree plot suggested a two-component solution. Consideration of the percentage of variance revealed that the first component accounted for 43.3% of the variance, with the second component raising this above the 50% threshold of variance suggested as an acceptable outcome for a PCA. Based on these three analyses, it was decided to specify a two-component solution for the PCA.

A PCA with direct oblimin with Kaiser normalization rotation and a forced two-component solution was computed. Two components were extracted that explained 53.6% of the total variance of the observed variables. Table 2 presents the component pattern matrix and regression coefficients for each variable on each of the components. The first component, which accounted for 44.6% of the total variance, was constructed only from items related to assessment-program design and instruments (assessment program and

Table 1. Eigenvalues, percentage of variance explained, and Monte Carlo simulation eigenvalues

Component	Eigenvalue	% Variance explained	Monte Carlo simulation eigenvalue
1	9.53	43.3	2.06
2	1.91	8.7	1.87
3	1.65	7.5	1.72
4	1.41	6.4	1.61

Table 2. Component pattern matrix and regression coefficients for a two-component solution for the FRAS

Question	Component 1 (program and instrument)	Component 2 (validation)	Assessment category by survey design
Scenario questions	0.78		Instruments
Assessment program	0.73		Assessment program
Student learning outcomes	0.72		Assessment program
Formative assessment	0.69		Assessment program
Summative assessment	0.67		Assessment program
Student competencies	0.67		Assessment program
Problem-solving questions	0.63		Instruments
Performance assessment	0.59		Instrument
Assessment validity	0.57		Validation
Assessment reliability	0.54		Validation
Assessment task	0.52		Instrument
Portfolio	0.49		Instruments
Authentic assessment	0.47		Instrument
Rubrics	0.44		Scoring systems
Holistic scales	0.43		Scoring systems
Item difficulty	0.38		Validation
Interrater reliability		–0.96	Validation
Intrarater reliability		–0.96	Validation
Test–retest reliability		–0.88	Validation
Internal consistency		–0.71	Validation
Content validity		–0.67	Validation
Item discrimination		–0.46	Validation
Alternative assessment		–0.41	Instrument
Analytic scales		–0.32	Scoring systems

instrument). The second component, which accounted for 9.0% of the total variance, consisted primarily of items from the realm of assessment validation (validation). This two-structure solution suggests a basic differentiation in knowledge between the design and development aspects of assessment practice and that of processes of validation.

To further explore the proposed structure of the faculty assessment knowledge based on the PCA, we examined the means of each of the design categories of the survey. Table 3 presents descriptive data for each of the vocabulary items by survey design category. The grand means of the survey categories reveal that, for this administration of the FRAS, faculty members reported having the least knowledge of validation items and the greatest knowledge of assessment-program items. While these are only descriptive data on a limited sample of 95 faculty members, it is supportive of the general outcome of the PCA, which suggests a distinction in faculty knowledge concerning assessment validation. Overall, the outcome of the PCA and consideration of the descriptive data suggest that faculty responses to the FRAS reflect an underlying distinction in the grouping of assessment knowledge between assessment design/instruments and assessment validation. This result will need further investigation with a larger and more diverse sample of faculty members.

To further augment this exploration of dimensionality, we also used a simple correlational approach. While this does not mollify the need for a full factor analysis of the survey at a later stage, it does offer some initial investigation of aspects of item relationship. In particular, as part of the validation process, the relationship among specific vocabulary items and frequency of engaging in assessment activities needed to be investigated. The logic behind asking participants about their familiarity with particular assessment vocabulary items is that this should indicate degree of exposure and expertise in assessment. By correlating ratings of familiarity with reported frequency of engaging in different assessment activities, it is possible to begin to check this assumption of a relationship. Accordingly, in the present validation study, in addition to the PCA for the assessment vocabulary items, two correlational studies were conducted: 1) Pearson's *r* correlation coefficients were calculated for all vocabulary items relating to knowledge of an assessment program and the rating of frequency of constructing an assessment program; and 2) Pearson's *r* correlation coefficients were calculated for all vocabulary items relating to assessment instruments and the rating of the frequency of writing formative assessments. A final correlational study that emerged from the findings of the PCA consisted of investigating the relationship between the vocabulary items relating to the process of assessment validation and the participant's overall confidence rating ("Overall, I am confident in my abilities to appropriately assess my course and my students"). The assumption underpinning this investigation was that if, as indicated in the PCA that validation is a separate grouping related to higher levels of assessment ability, this should also be correlated with overall confidence in conducting assessment.

The results of these correlation studies are presented in Table 4. As can be seen for all three analyses of relationship, positive, significant, but low correlations were found on all items. For the analysis of the relationship between the self-reported ratings of familiarity with assessment-program vocabulary items and the rating of the frequency of constructing an assessment program, the correlation

Table 3. Means, grand means, and SDs for vocabulary item scales by assessment category ($n = 95$)

Question	Mean[a]	SD	Assessment category
Assessment program	3.39	1.30	Assessment program
Student learning outcomes	4.55	0.59	Assessment program
Formative assessment	3.77	1.37	Assessment program
Summative assessment	3.81	1.33	Assessment program
		Grand mean	3.93
Portfolio	4.20	1.06	Instruments
Assessment task	3.28	1.30	Instrument
Performance assessment	3.20	1.32	Instrument
Authentic assessment	2.57	1.44	Instrument
Alternative assessment	2.39	1.38	Instrument
Problem-solving questions	4.58	0.89	Instrument
Scenario questions	4.09	1.12	Instrument
		Grand mean	3.47
Rubrics	4.71	0.61	Scoring systems
Analytic scales	3.07	1.25	Scoring systems
Holistic scales	2.36	1.20	Scoring systems
		Grand mean	3.59
Assessment validity	3.35	1.27	Validation
Item discrimination	2.69	1.38	Validation
Assessment reliability	3.37	1.28	Validation
Content validity	2.68	1.39	Validation
Item difficulty	3.18	1.42	Validation
Interrater reliability	2.84	1.62	Validation
Intrarater reliability	2.84	1.62	Validation
Test–retest reliability	2.64	1.45	Validation
Internal consistency	3.28	1.39	Validation
		Grand mean	2.98

[a]A five-point Likert scale from 1, "I have never heard this term before," to 5, "I am completely familiar with this term."

coefficients ranged from 0.35 to 0.41. This weak but significant relationship between these items suggests there is a connection between the familiarity with assessment-program vocabulary and the frequency of designing an assessment program. For the analysis of the relationship between self-reported familiarity with assessment instrument vocabulary and the frequency of writing formative assessments, four items were above the 0.3 correlation benchmark. This suggests that familiarity with authentic assessment, performance assessment, assessment tasks, and scenario questions is related to the frequency of writing formative assessments. Finally, for the analysis of the relationship between familiarity of assessment-validation vocabulary items and the rating of overall confidence in assessment, four items were found to be above the 0.3 correlation benchmark. The items assessment reliability, assessment validity, item difficulty, and item discrimination were positively correlated with overall confidence in conducting assessment, suggesting these are related to overall assessment confidence. The relationships revealed in these three studies suggest that self-reported familiarity with vocabulary is connected to relevant self-reported frequency of assessment activity and overall confidence in assessment ability. This supports the underpinning assumption that reported familiarity with assessment vocabulary translates into professional functioning in terms of assessment activity and ability. However, future research is required to consider the dimensionality of the full survey.

Consistency

Following the outcome of the PCA process, the final version of the FRAS was analyzed for its internal consistency. The internal consistency of the whole instrument (including the vocabulary familiarity, frequency of assessment practice, confidence, and attitude scales) was calculated using Cronbach's alpha with the result $\alpha = 0.96$, which indicates very high levels of consistency for the tool. To further interrogate internal consistency, we calculated item-total correlations for each item. This analysis consists of correlating each item with the sum of the items (total score) and allows the identification of items that might reduce reliability (Guilford, 1953). In the current analysis, all items were situated between 0.965 and 0.964, indicating a fluctuation of only 0.001 in the resultant Cronbach's alpha. Because the whole instrument includes 50 items (and this can artificially inflate Cronbach's alpha), internal consistency was also calculated individually for each of the four sections of the FRAS. The internal consistency for the vocabulary familiarity sections was $\alpha = 0.94$; for the confidence scales, $\alpha = 0.93$; and for the attitude scales, $\alpha = 0.91$. For each of these sections, consideration of the item-total correlations did not necessitate any changes in the included items. For the frequency of assessment practice scales, one of the items—"I have defined my course in terms of student competencies"—reduced the overall internal consistency, and it was removed from the survey. The resultant measure of internal consistency for this section of the instrument was $\alpha = 0.91$. Accordingly,

Table 4. Pearson *r* correlations for assessment program, formative assessment, and validation ($n = 105$)

Assessment vocabulary item	"I have constructed an assessment plan for my course"	"I have written formative assessments"	"Overall I am confident in my abilities to appropriately assess my course and my students
Assessment program			
Assessment program	0.38***		
Student learning outcomes	0.41***		
Formative assessment	0.35***		
Summative assessment	0.37***		
Instruments			
Portfolio		0.24***	
Assessment task		0.38***	
Performance assessment		0.39***	
Authentic assessment		0.36***	
Alternative assessment		0.28***	
Problem solving		0.28***	
Scenario questions		0.39***	
Validation			
Assessment validity			0.36***
Item discrimination			0.31**
Assessment reliability			0.34***
Content validity			0.27**
Item difficulty			0.33***
Interrater reliability			0.29**
Intrarater reliability			0.29**
Test–retest reliability			0.28***
Internal consistency			0.24*

*$p < 0.01$.
**$p < 0.005$.
***$p < 0.001$.

based on the results of the reliability analysis, the whole scale and its subsections are to be considered highly reliable. The distributed version of the FRAS can be found in the Supplemental Material.

Differentiating between Novice and Advanced Faculty Assessment Knowledge and Experience

The final stage of the FRAS validation process consisted of empirically evaluating the ability of this tool to differentiate between groups of faculty members with different levels of assessment knowledge and experience. The process used here consisted of a known-groups validation process and, as described earlier, is basically a statistical comparison of the two groups. The core assumption here is that if the FRAS can demonstrate statistical differences between two known groups, then the tool has validity in terms of its ability to differentiate levels of knowledge and experience.

As a first stage, the data set was explored using descriptive approaches. Normality of each of the items on the FRAS was assessed using the Kolmogorov-Smirnov test with a Lilliefors significance correction and the Shapiro-Wilk test of normality. For the majority of the items and for all items in relation to one or the other group, the significance levels of the two tests of normality were below the 0.001 threshold, indicating the data are not normally distributed, and a parametric analytical approach is therefore not appropriate for this data set. Accordingly, a nonparametric comparative Mann-Whitney *U*-test was calculated. All rating scales across the four categories of information (knowledge, practice, confidence, and attitude) were analyzed for systematic differences between defined novice and advanced groups.

Table 5 summarizes the means, SDs, Mann-Whitney *U*-test values, and significance levels for all items on the self-reported faculty knowledge of assessment vocabulary items. As can be seen in Table 5, there are significant differences between novice and advanced faculty groups on self-reported assessment vocabulary familiarity ratings. Of the 23 items on the FRAS, 15, or 65%, were found to be significantly different ($p < 0.01$), with the directions of the means revealing that advanced faculty members report higher levels of familiarity with these terms (except in the case of "analytical scales"). The FRAS does have the ability to differentiate levels of formal assessment knowledge and experience.

Consistent with the PCA study of the psychometric properties of the FRAS, validation seems to function as a specific indicator in differentiating faculty assessment knowledge and experience. Differences between advanced and novice groups were most pronounced in the category of validation of assessment instruments. In this category, eight of the nine terms presented were reported as being significantly better known to the advanced group than to the novice group. This set includes the terms "assessment validity," "assessment reliability," "content validity," "item difficulty," "intrarater reliability," "test–retest reliability," and "internal consistency." This result substantiates PCA and correlational study results and reflects that validation of instruments may be a knowledge category

Table 5. Mann-Whitney *U*-test comparisons for advanced and novice groups on self-reported degrees of familiarity of assessment knowledge terms[a]

Question	Mean and SD advanced	Mean and SD novice	Mann-Whitney *U*-test	Significance	Assessment category
Assessment program	4.18 (0.75)	3.00 (1.27)	54**	0.003	Assessment program
Student learning outcomes	4.82 (0.4)	2.59 (1.7)	64**	0.006	Assessment program
Formative assessment	4.64 (0.5)	4.53 (0.62)	128	0.49	Assessment program
Summative assessment	4.64 (0.5)	3.59 (1.54)	73**	0.01	Assessment program
Portfolio	4.27 (1.42)	3.71 (1.53)	95.5	0.08	Instruments
Assessment task	3.82 (1.47)	3.12 (1.36)	80*	0.03	Instrument
Performance assessment	3.73 (1.42)	1.88 (1.31)	49**	0.001	Instrument
Authentic assessment	3.55 (1.69)	3.53 (1.58)	133	0.74	Instrument
Alternative assessment	3.27 (1.67)	1.47 (1.17)	59**	0.005	Instrument
Problem-solving questions	4.73 (0.19)	1.18 (0.12)	8.5**	0.001	Instrument
Scenario questions	4.27 (0.38)	4.41 (0.32)	127	0.61	Instrument
Rubrics	4.73 (0.46)	4.18 (1.07)	99	0.09	Scoring systems
Analytic scales	2.36 (1.56)	4.41 (0.87)	63.5**	0.005	Scoring systems
Holistic scales	2.18 (1.76)	2.41 (1.37)	126.5	0.5	Scoring systems
Assessment validity	4.36 (0.67)	1.94 (1.19)	11.5**	0.001	Validation
Item discrimination	4.18 (0.75)	3.24 (1.52)	101.5	0.11	Validation
Assessment reliability	4.18 (1.16)	3.0 (1.22)	63**	0.005	Validation
Content validity	4.18 (1.25)	2.18 (1.51)	50.5**	0.002	Validation
Item difficulty	3.91 (1.3)	2.88 (1.31)	67**	0.009	Validation
Interrater reliability	3.73 (0.54)	2.94 (1.47)	98.5	0.18	Validation
Intrarater reliability	3.64 (1.74)	2.29 (0.38)	76.5*	0.02	Validation
Test–retest reliability	3.36 (1.69)	2.00 (1.5)	72**	0.01	Validation
Internal consistency	3.27 (1.73)	1.88 (1.11)	60.5**	0.009	Validation

*$p \leq 0.05$.
**$p \leq 0.01$.
[a]Five-point Likert scale from 1, "I have never heard this term before," to 5, "I am completely familiar with this term."

that specifically correlates with advanced assessment experience. This result suggests that conceptual or working knowledge of standard practices for validating assessment instruments may serve as a significant determinant in defining faculty assessment expertise.

In relation to familiarity with terminology associated with assessment instruments, four of the seven terms tested were found to be significantly different ($p < 0.01$). In particular, the concepts of "assessment task," "problem-solving questions," "alternative assessment," and "performance assessment" were rated as more familiar to the advanced group than to the novice group. The difference in degree of familiarity was particularly pronounced in relation to the concept of "performance task," with the novice group only assigning a mean of 1.88 ("I have heard this term before but do not know what it means"). For the items relating to assessment-program design, three of the four items were found to be significantly different ($p < 0.01$), with the advanced group reporting higher familiarity on the terms "assessment program," "student learning outcomes," and "summative assessment." The novice group gave only a mean familiarity rating of 2.59 ("I have heard this term before but I am not confident what it means") for the term "student learning outcomes," suggesting potential difficulties in defining an assessment program. There were no significant differences between groups for the scoring system items, except for the term "analytical rating scales," which was scored higher by the novice than the advanced group. Overall, the results of the comparative analysis of this section of the instrument suggest that the FRAS has the ability to systematically differentiate between faculty members with different levels of assessment knowledge.

For further evaluation of the ability of the FRAS to differentiate between levels of faculty interaction with assessment, Mann-Whitney *U*-tests were performed and associated significance levels were determined for the assessment practice, confidence, and attitudinal sections of the instrument. Figure 1 depicts the comparison between advanced and novice group results in assessment practice. Five of seven items were significantly different at the $p < 0.01$ level. The direction of the difference is important, with the advanced faculty group self-reporting higher incidence of practice than the novice group. Figure 2 shows results for comparison between advanced and novice groups in assessment confidence. Eleven of 13 items were determined to be statistically different at the $p < 0.01$ level, with advanced group mean significantly higher than the novice group mean. The comparison between advanced and novice group results in attitudes toward assessment is displayed in Figure 3. Again, the direction of difference is as expected, with advanced group reporting more positive attitudes toward assessment. Additional statistical data related to Figures 1–3 are available in the Supplemental Material. Overall, these results suggest that the FRAS differentiates levels of faculty knowledge and experience in relation to assessment.

The advanced group reported significantly higher frequencies than the novice group (rating mean greater than 3.0) for six assessment practices, including "constructing an assessment plan," "defining learning outcomes," "writing summative assessments," and "providing feedback

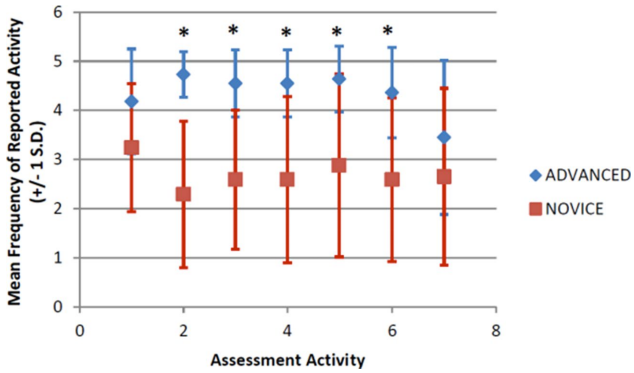

Figure 1. The FRAS instrument differentiates levels of faculty assessment experience in practice. Mean faculty responses in response to statements about frequency of engagement in assessment practices from advanced (n = 11) and novice (n = 17) groups were analyzed by the Mann-Whitney U-test of nonparametric distribution. The means and SDs for advanced (blue) and novice (red) groups are plotted on the chart. Responses were numerically coded from 1, "Never," to 5, "All the time." Statements are: 1) "I have constructed an assessment plan for my course," 2) "I have defined student learning outcomes for my course," 3) "I have written formative assessments," 4) "I have written summative assessments," 5) "I have provided feedback to students based on formative assessments," 6) "I have provided feedback to students based on summative assessments," and 7) "I have written a report based on assessment data." Asterisks indicate a significance level of $p < 0.01$ or less. Advanced mean value = 4.35, SD = 0.86; novice mean value = 2.69, SD = 1.60.

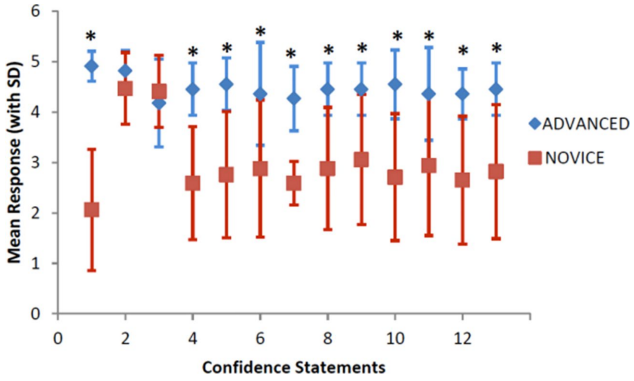

Figure 2. The FRAS instrument differentiates levels of faculty confidence in assessment competency. Mean faculty responses in response to statements about confidence of assessment practices from advanced (n = 11) and novice (n = 17) groups were analyzed and plotted. Responses were numerically coded from 1, "Strongly disagree," to 5, "Strongly agree." Statements extend "I am confident in my ability to": 1) "define the important components of my course," 2) "define my course in terms of student learning outcomes," 3) "define my course in terms of student competencies," 4) "design formative assessments," 5) "design summative assessments," 6) "evaluate the quality of the assessments I have designed," 7) "analyze the formative assessments I have designed," 8) "analyze the summative assessments I have designed," 9) "provide students with the relevant feedback based on the formative assessments I have designed," 10) "explain to specific students the outcomes of their summative assessment performance," 11) "report assessment outcomes to administrators," 12) "I am confident that my assessments accurately reflect the teaching objectives of my course," and 13) "Overall I am confident in my abilities to appropriately assess my course and my students." Advanced mean value = 4.47, SD = 0.61; novice mean value = 2.99, SD = 1.12.

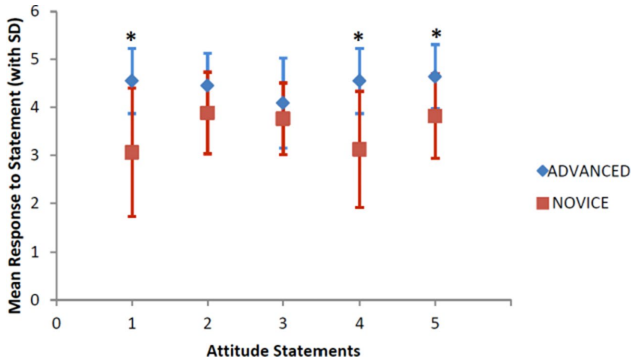

Figure 3. The FRAS instrument differentiates faculty attitudes toward assessment. Mean faculty responses in response to statements about assessment attitudes from advanced (n = 11) and novice (n = 17) groups were analyzed as in Figure 1. The means and SDs for advanced (blue) and novice (red) groups are plotted on the chart. Responses were numerically coded from 1, "Extremely negative," to 5, "Extremely positive." Statements extend "How do you feel about": 1) "using assessment in your course," 2) "reporting to faculty on assessments of your class," 3) "reporting to administration on assessments of your class," 4) "providing feedback to your students based on formative assessment," and 5) "providing feedback to students based on summative assessment." Advanced mean value = 4.46, SD = 0.73; novice mean value = 3.53, SD = 1.00.

on formative and summative assessments." Advanced respondents reported very high frequency of engagement in the practice of "providing formative feedback to students" over novice respondents, suggesting that formal assessment development correlates with adoption of novel assessment practices.

For scales relating to degrees of confidence in ability to perform assessment activities, the advanced group reported significantly higher levels ($p < 0.01$) of confidence for 10 of the 13 activities. Importantly, there is a significant 1.63 mean difference between advanced and novice groups on the overall evaluation of confidence in their ability to "appropriately assess" their courses and their students. These results suggest that knowledge and experience correlates with increased faculty confidence.

The final issue that the survey addresses is the attitudinal response of faculty members to various assessment activities. The advanced group had higher mean ratings than the novice group, signifying a more positive attitude about assessment activities on all the items, with significant difference reflected for three of the five items. There was a significant 1.43 mean difference in faculty members' positive attitude toward "providing feedback to students from formative assessments," with the advanced group reporting extremely positive attitudes toward this activity. As with the confidence ratings, the data on faculty attitude toward different assessment activities suggest that formal exposure to assessment knowledge and experience correlates with stronger positive attitudes toward assessment and higher frequency of adoption of assessment practices.

The advanced group in this study consisted of faculty members who participated together in a multiyear assessment-development program, while the novice group had no prior collective experience in common. Participants in both groups elected to participate in the program and thus

represent a self-selected group in terms of general interest in and motivation toward assessment. While both groups may likely have been more motivated toward assessment development and thus willing to engage actively in the study, FRAS results reflect a wide range of experience with assessment among the novice group, which is likely to be generally reflective of the diversity of STEM faculty experience. Specifically, in addition to increased group means in all sections of the FRAS, decreased SDs for all statistically significant items were observed in the advanced group results compared with the novice group. Thus, the advanced group displayed a convergence of self-reported scoring, suggesting that participating in a formal assessment-development program correlates with emergence of a common vocabulary and shared understanding of assessment practices.

Overall, the survey data from the comparative known-groups study suggest that the FRAS is a useful and valid tool for differentiating levels of STEM faculty assessment knowledge and self-reported experience. Significant differences were found on all sections of the survey and in the direction expected. The advanced group had significantly higher levels of content knowledge (as measured by familiarity with assessment terminology), were involved in assessment activities more frequently, had higher levels of confidence concerning their ability to conduct assessment activities, and had more positive attitudes about assessment.

CONCLUSIONS

The primary aim of this study was to develop and evaluate a novel survey instrument for providing useful data about STEM faculty assessment knowledge and experience. The FRAS offers a comprehensive tool for self-reported information, enabling correlation across aspects of faculty knowledge, attitudes, confidence, and practice. The FRAS was validated using standard approaches and displays strong consistency and reliability, supporting its utility as a novel instrument for differentiating levels of faculty assessment knowledge and experience. The dimensionality of familiarity with the assessment terms section of the survey was analyzed, and the outcome suggests a difference in knowledge in assessment-program design, instrumentation, and assessment validation. The organization of elements into categories enables identification of specific gaps in knowledge that might inform departmental conversations or areas of focus in faculty-development programming. For instance, lower levels of self-reported familiarity for instrument validation items suggests this topic may represent a more advanced assessment knowledge level. In a known-groups validation study, 34 of the 51 individual items on the FRAS were found to be significantly different, with means in the expected direction of higher outcomes for the advanced group.

This study presents a preliminary examination of the psychometric properties of the FRAS as it has been adapted for use in specified contexts. While we report initial tests of dimensionality, full factorial analysis of the instrument with a larger subject group would enable exploration of internal relationships and underpinning constructs within the instrument itself. A second limitation of the study relates to the small numbers and specification of the participants in the known-groups validation study. The sizes of the groups and characteristics of the data set forced a nonparametric analysis of results and limited generalizability. Once again, larger comparison groups drawn with relevantly defined high and low assessment knowledge would allow further validation of this tool. Finally, the current study only analyzes reliability in terms of internal consistency. Future studies of the stability of the tool over administrations should be conducted.

The FRAS is a valid tool with high levels of internal consistency and may be easily administered as a way to gauge faculty interest and need with respect to assessment development. Faculty members with advanced assessment knowledge self-reported higher levels of familiarity with assessment terms, higher frequencies of assessment activity, increased confidence in conducting assessments, and more positive attitudes toward assessments than novice faculty members. It should be noted that, while the FRAS interrogates faculty self-reported awareness and attitudes about assessment and assessment practice, it is not valid as an assessment of the quality of faculty assessment practice. Rather, as a tool developed in the context of an assessment-development program specifically to monitor faculty progress in their self-reported awareness, attitudes, and practices, it is likely to be most useful in providing feedback on assessment-development programs geared toward faculty members with limited assessment experience.

The FRAS was developed for application with STEM faculty members for whom similar professional training and experience may specifically inform their experience with and attitudes toward learning assessment; application of the tool in non-STEM faculty populations has not been tested and is an area for future study. The FRAS may be included as an additional element of formative departmental assessment among other established measures of STEM faculty practice (e.g., COPUS, BioCore Guide, or PULSE departmental rubrics; Smith *et al*., 2013; Brownell *et al*., 2014; PULSE Community, 2014). It is intended as a valid and reliable instrument for informing departmental conversations, identifying areas of interest for assessment learning, or monitoring the outcome of faculty assessment-development programs. The FRAS can be used to establish baseline information and to monitor advances in knowledge, attitudes, and practice as faculty members advance their assessment experience.

As departments respond to national calls to improve undergraduate STEM education, there is increasing awareness of the need to ground science teaching practice in the scholarship of how students learn. Scientific teaching implies that instructors establish an active educational context that supports student engagement and provides mechanisms for regular feedback on students' progress. Thus, effective teaching requires that faculty members adopt assessment methods that allow instructors to capture the complexities of student thinking and action that are inherent to science practice. Developing faculty capacity for learning assessment is a key lever for improving undergraduate STEM education. The FRAS offers one avenue for measuring development in self-reported assessment familiarity and confidence and, as such, may have a formative role in directing future program, curricular, and institutional innovations in the direction of evidence-based science teaching.

ACKNOWLEDGMENTS

The authors thank Ms. Patricia Soochan for careful editing and two anonymous reviewers for critical feedback that served to strengthen the manuscript. D.I.H. was supported by HHMI award #52007402.

REFERENCES

American Association for the Advancement of Science (2011). Vision and Change in Undergraduate Biology Education: A Call to Action, Washington, DC.

Association of American Medical Colleges (2009). Scientific Foundations for Future Physicians, AAMC-HHMI Committee Report, Washington, DC.

Bachman L (2004). Statistical Analyses for Language Assessment, Cambridge, UK: Cambridge University Press.

Bauerle C, Hanauer DI (2013). Essential outcomes, essential inputs. Peer Review *14*, 31.

Brownell SE, Freeman S, Wenderoth MP, Crowe AJ (2014). BioCore Guide: a tool for interpreting the core concepts of Vision and Change for biology majors. CBE Life Sci Educ *13*, 200–211.

Cattell R (1978). The Scientific Use of Factor Analysis in Behavioral and Life Sciences, New York: Plenum.

College Board (2012). AP Biology Course and Exam Description, rev. ed. http://media.collegeboard.com/digitalServices/pdf/ap/ap-biology-course-and-exam-description.pdf (accessed 8 October 2014).

de Winter J, Dodou D, Wieringa PA (2009). Exploratory factor analysis with small sample sizes. Multivar Behav Res *44*, 147–181.

Emenike M, Raker JR, Holme T (2013a). Validating chemistry faculty members' self-reported familiarity with assessment terminology. J Chem Educ *90*, 1130–1136.

Emenike ME, Schroeder J, Murphy K, Holme T (2013b). Results from a national needs assessment survey: a view of assessment efforts in chemistry departments. J Chem Educ *90*, 561–567.

Guilford J (1953). The correlation of an item with a composite of the remaining items in a test. Educ Psychol Meas *13*, 87–93.

Hair J, Anderson RE, Tatham RL, Grablowsky BJ (1979). Multivariate Data Analysis, Tulsa, OK: Pipe Books.

Hanauer DI, Bauerle C (2012). Facilitating innovation in science education through assessment reform. Liberal Educ *98*, 34–41.

Hanauer DI, Hatfull GF, Jacobs-Sera D (2009). Active Assessment: Assessing Scientific Inquiry, Dordrecht, Netherlands: Springer.

Handelsman J, Miller S, Pfund C (2007). Scientific Teaching, New York: Freeman.

MacCallum RC, Widaman KF, Preacher KJ, Hong S (2001). Sample size in factor analysis: the role of model error. Multivar Behav Res *36*, 611–637.

National Research Council (2012). A Framework for K–12 Science Education: Practices, Crosscutting Concepts, and Core Ideas, Washington, DC: National Academies Press.

Netemeyer R, Bearden WO, Sharma S (2003). Scaling Procedures: Issues and Applications, Thousand Oaks, CA: Sage.

President's Council of Advisors on Science and Technology (2012). Engage to Excel: Producing One Million Additional College Graduates with Degrees in Science, Technology, Engineering, and Mathematics, Washington, DC: U.S. Government Office of Science and Technology.

PULSE Community (2014). Home page. www.pulsecommunity.org (accessed 8 October 2014).

Raker J, Holme TA (2014). Investigating faculty familiarity with assessment terminology by applying cluster analysis to interpret survey data. J Chem Educ *91*, 1145–1151.

Schilling K, Schilling KL (1998). Proclaiming and Sustaining Excellence: Assessment as a Faculty Role, Washington, DC: Association for the Study of Higher Education.

Shavelson RJ, Young DB, Ayala CC, Brandon PR, Furtak EM, Ruiz-Primo MA, Tomita MK, Yue Y (2008). On the impact of curriculum-embedded formative assessment on learning: a collaboration between curriculum and assessment developers. Appl Meas Educ *21*, 295–314.

Smith ME, Jones FH, Gilbert SL, Wieman CE (2013). The Classroom Observation Protocol for Undergraduate STEM (COPUS): a new instrument to characterize university stem classroom practices. CBE Life Sci Educ *12*, 618–627.

Wang X, Hurley S (2012). Assessment as a scholarly activity?: faculty perceptions of and willingness to engage in student learning assessment. J Gen Educ *61*, 1–15.

Yin Y, Shavelson RJ, Ayala CC, Ruiz-Primo MA, Brandon PR, Furtak EM, Tomita MK, Young DB (2008). On the impact of formative assessment on student motivation, achievement, and conceptual change. Appl Meas Educ *21*, 335–359.