122nd ASEE Annual Conference & Exposition
June 14 - 17, 2015
Seattle, WA

Making Value for Society

Paper ID #13108

Developing and Validating a Concept Inventory

Miss Natalie Jorion, University of Illinois, Chicago

Natalie Jorion is a research assistant and Ph.D. student of learning sciences specializing in psychometrics at the University of Illinois in Chicago, 1240 W. Harrison St, Chicago, IL 60607; njorio2@uic.edu.

Dr. Brian Douglas Gane, University of Illinois at Chicago

Dr. Brian Gane is a Visiting Research Assistant Professor at the Learning Sciences Research Institute, University of Illinois at Chicago. Dr. Gane's research focuses on psychological theories of learning and skill acquisition, assessment, instructional design, and STEM education.

Prof. Louis V DiBello
Prof. James W Pellegrino, University of Illinois, Chicago

Developing and Validating a Concept Inventory

Introduction

Concept inventories (CIs) have been used to assess undergraduate students' understanding of important and difficult concepts in engineering disciplines. However, research has shown that even meticulously designed CIs often fall short of validly measuring student conceptual understanding.[1,2] CI developers' intentions of measuring particular understandings are sometimes not congruent to the skills and conceptual understandings students use to interpret and respond to items in practice. This incongruity can occur despite developers' expertise, perhaps because of the "expert blind spot." [3,4] Even when developers create items that tap into the intended conceptual understandings, the assessments may not reveal the extent to which students have mastered particular concepts. To create an inventory whose scores are interpretable and meaningful requires that the developers be mindful of validity concerns from the outset. An assessment's validity is the extent to which an assessment measures what it was intended to measure. Evidence related to validity is demonstrated both in analyses of the assessment's content and of examinee response patterns. Evaluating content and response patterns are two parts of an evidentiary argument process.[5]

This paper presents the Evidentiary Validity Framework, an analytic framework that researchers and test developers can use to evaluate validity arguments once the assessment has been developed. It does so by first building on established definitions related to validity, then introducing the evidentiary approach to assessment design.[6] We present two case studies to illustrate using these analyses to evaluate a CI's validity: the Dynamics Concept Inventory[7] (DCI) and the Conceptual Assessment Tool for Statics[8] (CATS). The developers of both tests make claims about overall conceptual understanding of the domain, understanding of specific concepts, and about students demonstrating particular misconceptions or common errors. Our analysis found varying degrees of support for each claim, such as the use of the assessment as a measure of students' overall understanding. Furthermore, only CATS provided evidence for student understanding of specific domain concepts. Neither assessment showed strong evidence for measuring student misconceptions. Overall, this Evidentiary Validity Framework can serve as a guideline for evaluating the validity arguments of CIs. It can also help CI developers plan ahead when creating inventories so that the validity claims are better aligned to student reasoning.

The Importance of Assessment Validity

For an assessment to be useful, it should measure what it was intended to measure and support the intended uses and purposes.[9,10] *Validity* in terms of assessment design is the extent to which the intended interpretations of test scores with respect to some purpose can be supported by multiple sources of evidence. It is a process of building a validity argument in which claims of test score interpretation relative to a given purpose are clarified and evidence is collected to support those claims. An assessment cannot be deemed "valid" in absolute terms; validity is relative in that it is dependent on the proposed interpretation and use of test scores.[11,12]

Part of this argumentation process requires clearly specifying several components. First, assessment developers should establish the claims about student understanding that they intend to make based on test scores and other test outcomes. These claims should be evaluated in terms of their clarity and plausibility. In addition, developers should indicate what kind of evidence would support these claims. Lastly, those multiple forms of evidence should be collected and interpreted to determine the extent to which they do or do not support the intended assessment claims.[5]

The assessment triangle is one schema for analyzing validity properties of assessments used within the classroom involves three integrated components:[9,13]

- *Cognitive* – This involves the extent to which an assessment taps the desired forms of domain and disciplinary knowledge and skill and does so in ways that are not confounded with other aspects of cognition such as language or working memory load. The conceptual underpinnings for particular assessments are evaluated, including what they reveal about student learning and understanding of critical mathematical or scientific concepts, procedures, and practices. Particular attention must be paid to assessing linguistically and ethnically diverse populations.
- *Instructional* – This involves assessment outcomes that support teaching practice and provide timely information to support instructional decisions and actions (e.g., guide the instructor in what content to cover for the whole class, identify areas of content weakness for specific students, or highlight student misconceptions). In general, it concerns the extent to which the assessment provides teachers with information that can be used to effectively support their instruction.
- *Inferential* – This involves how well the assessment results support statistical inferences derived from application of various psychometric methods regarding the measurement properties of the instrument, including classical test theory, item response theory, and multivariate structural analysis.

These three components are associated with general categories of claims about an assessment and need to be specified for any given instrument. Developers or users of a specific CI would be expected to make specific claims depending on the instrument's proposed use. After specifying the claims and warrants, one can then provide the validity argument in which one evaluates the warrants and backing (evidence) that support the various claims. The plausibility of this argument is the crux of the validity argument.

To evaluate the validity of cognitive claims, developers should identify what evidence would demonstrate student knowledge and understanding. This should be based on both empirical research on learning in engineering disciplines,[14] instructor pedagogical content knowledge, student free responses during pilot studies, and student think-aloud studies.[15,16] Student responses can indicate how they are interpreting and reasoning through items. In developing this evidence, it is important to evaluate whether construct-irrelevant variance and construct underrepresentation have been minimized.[9]

To evaluate the validity of instructional claims, one should evaluate evidence that the instrument can provide information that is aligned with instruction and is useful for instructors. These data might arise from interviews with teachers and/or from analysis of student data. For example, an instrument might reveal misconceptions that the majority of the class hold, or it might aid in identifying a subset of students that need remediation on a specific topic. One common claim is that a CI can be used to determine the efficacy of a specific instructional intervention; evidence for this claim would center on showing that the CI is sensitive to different levels of student proficiency on concepts covered in the intervention.

In gathering evidence for inferential claims, different psychometric analyses should be conducted. For instance, as a first step, one might evaluate an instrument's reliability and measurement error. Reliability is the extent to which an instrument provides the same results over a repeated number of administrations. If the instrument has adequate reliability and low measurement error (for its intended use), then this provides further support for using the performance data to model student thinking (e.g., through factor analysis, item response theory or diagnostic classification modeling).

Applying the Evidentiary Validity Framework to Concept Inventories

Rigorous development of a validity argument and pursuit of validity evidence in support of that argument are particularly important for assessments such as concept inventories that are administered across multiple institutions and, in some cases, are used to evaluate educational interventions.[9,17] To investigate the validity properties of an inventory, one must first identify what claim(s) the developers or users are making about their concept inventory. Claims can be about student learning gains, student misunderstandings, and overall mastery of particular concepts. Once these claims are explicated, it is possible to determine how well developers' claims about what is intended to be measured can be supported with empirical and analytic evidence.[5] Generally, CI developers make three claims about their inventories. Following are the claims and examples of methods to validate each particular claim.

1. **Overall mastery of all concepts represented in the CI.** This claim asserts that (1) overall performance on the inventory measures the focal domain knowledge and that (2) individual items provide coherent data that can be aggregated into an overall measure of performance. Researchers can evaluate this claim in three ways. First, the investigators can determine the assessment's overall reliability using a statistic such as Cronbach's alpha.[18] This index is based on the number of items and the extent to which the items are correlated. Researchers can also calculate the standard error of measurement to determine the confidence with which particular scores can be differentiated. Third, they can examine how items perform in regards to the entire assessment, including calculating alpha-if-item-deleted, item discrimination, item difficulty, and item response theory (IRT) model-fit.[19] Item difficulty is the proportion of answers correct. We use the range of 0.2 to 0.8 as reasonable values for item difficulty. Guidelines for acceptable ranges depend on an assessment's purpose and use, and in this case, the 0.2 to 0.8 range is a conservative, "reasonable" range. Item discrimination is the extent to which correctly answering one item corresponds to performance on the rest of the test. We recommend that values should typically be above 0.2. IRT is a psychometric method that focuses on the item instead of the total test score; it models the probability of answering the item correctly given a specific latent trait. The one parameter logistic (1PL) IRT model characterizes item difficulty,

and the two-parameter logistic IRT model characterizes both item difficulty and discrimination.

2. **Mastery of particular concepts**. This claim asserts that the instrument has subgroups of items that represent different domain concepts. Therefore, one can examine performance on groups of items to measure understanding of individual concepts (i.e., calculate subscores). To indicate mastery of particular concepts, the researcher can group together items by construct. Before doing so, the researcher should remove problematic items, such as those with negative inter-correlations, to ensure that problematic items do not detrimentally affect subscores.[13] The particular methods used to investigate subscores include subscale alphas, exploratory factor analysis, and confirmatory factor analysis. These methods can be used to evaluate the extent to which performance on the items within categories and the categories themselves align with the developer's hypothesized constructs.

3. **Propensity for misconceptions or common student errors**. This claim asserts that the instrument is able to reveal common student errors by means of distractor response patterns. There are at least two ways to investigate this aspect of student performance. One way is to split the dataset into high and low performing students, and then compare how the distractors are selected by students in each respective sample. This can indicate which distractors are attractive to students with low conceptual understanding, signifying misconceptions.[20] Alternately, an IRT approach can be used to measure misconceptions. By using a polytomous scoring response model, latent knowledge states can be associated with each respective multiple-choice response. In this way, each answer will have an associated ability level, which can indicate knowledge progressions.[21] Despite the availability of these methods, it can be challenging to apply them to determine if there is evidence of persistent student misconceptions.

Together, these three classes of analyses can provide evidence for the extent to which the developers' claims align with data patterns found in actual student performance. They can be used iteratively to refine an instrument or to assess the appropriateness of interpreting and using CI scores for a particular purpose and within a specific context.

Building an Inventory using Evidence-Centered Design

There are several steps that test developers should take from the outset to improve the likelihood that their assessments will align with their intended measurement purposes and interpretive uses. Designing items representative of the construct to be measured requires developing an evidence-driven conceptual assessment framework.[6] This framework should be based on a *domain analysis* and a *domain model*, which provides information about the target domain. This includes specifying student thinking, desired performance outcomes, problem representations, and the assumed learning model. Ultimately, the conceptual assessment framework enables a blueprint to be developed that links design components called the student model, evidence model, and task model. With this information, the developers can specify a domain's "big ideas" and then develop a student model. A *student model* defines the knowledge, skills, and abilities (KSA) the developers are trying to measure. An *evidence model* specifies how potential observations provide evidence for this KSA. The *task model* provides a format for student work products to provide evidence of these KSAs, and specifies characteristic and variable features of problems. Connecting these three components requires aligning student learning objectives (student model) with opportunities for students to demonstrate knowledge (task model) and ways of measuring differences in understanding (evidence model).

In creating a conceptual assessment framework, learning objectives should have precise cognitive operands such as predict, explain, contrast, and apply instead of more vague descriptors such as know or understand. These operands can be likened to those provided in Bloom's taxonomy of learning objectives. This ensures that the claims about student knowledge and the ways in which students are supposed to demonstrate this knowledge can be clearly specified for purposes of measurement.[2]

Moreover, the grain size of the conceptual measurement category should be commensurate with the learning goals. In our experience we have found that categories that have a cohesive, unified concept with a precise learning objective tend to have more positively correlated items. Textbook chapter topics do not tend to make conceptually cohesive categories, as the grain size tends to be too large and the items too disparate conceptually. For example, "Identify force and moment reactions at the supports and connections of a rigid body" is a more meaningful learning goal compared to just "Equilibrium of rigid bodies." The former has a better specified learning objective compared to the latter.

If part of the student model involves student misunderstandings, developers should construct distractors carefully. This is especially true when an inventory is being used to identify ways to inform instructional decisions and actions. As much as possible, developers should base distractors on student problematic thinking. For example, this thinking could be from discipline-based education research findings, instructors' pedagogical content knowledge, or students' free-response answers during pilot studies. Student think-aloud interviews can serve as a means to check how students are interpreting questions. If one goal of the CI is to help diagnose student misconceptions then further design decisions are required. As much as possible, answer choices should be mapped to relevant and differentiable misconceptions. Mapping all distractors to the same misconception reduces the diagnostic capacity of the instrument. Misconceptions should not be localized within only one item, but items within a conceptual category will likely have associated misconceptions. Only when students demonstrate a recurring misconception across several items can inferences be made that students possess that particular misconception.

As part of the conceptual assessment framework, the developer should specify an appropriate measurement model with which to evaluate the inventory and interpret the results. The measurement model can help guide the design of the assessment. For example, developers may plan to use confirmatory factor analysis to evaluate category cohesiveness. Research shows that having at least three items within a conceptual cluster is necessary to provide sufficient evidence that students have mastered a concept.[2] When investigating a test's validity, it is difficult to determine whether problematic items are a result of idiosyncrasies of the item or item categories. Performance on one or two items is not sufficient evidence to make strong conclusions about the specifics of student thinking. Thus, specifying the measurement model in advance will affect the overall design of the inventory and the ability to pursue data collection and various analyses related to supporting specific validity claims.

Case Examples of Two Concept Inventories

We applied the Evidentiary Validity Framework to two CIs to examine the extent to which the developer's claims aligned with the content and student performance data. The extended, in-depth analysis is in a paper under review.[2] The first CI we investigated was the Dynamics Concept Inventory (DCI)[7] an inventory consisting of 29 multiple-choice items. Five of these items are from the Force Concept Inventory.[22] The developer designated 14 conceptual categories for the inventory, with one to five items per category. The dataset consists of 966 cases of student performance on the entire inventory. The samples were drawn from two large public universities and the inventory was administered toward the end of student enrollment in an undergraduate dynamics course. The developer makes the following claims about the inventory: (1) the overall score is indicative of students' dynamics knowledge, (2) sub-scales can indicate differentiated conceptual knowledge, and (3) incorrect answers can indicate common misconceptions.[7]

Dynamics Concept Inventory

Claim 1: Overall mastery. The total mean score for the DCI was 14.3 out of 29 (SD=4.6). Item difficulties ranged from 0.06 to 0.91. This suggests that several items were too easy for the given population, while others were too difficult. The item discrimination measures ranged from 0.01 to 0.56; several items did not discriminate well between low and high performing students.

The DCI was fairly reliable for a 29-item assessment (α=0.74). Four items had a higher alpha-with-item-deleted values than the overall alpha of 0.74, indicating that dropping these items would improve the overall reliability of the assessment.

The standard error of estimate for this sample was 2.02. This means that for a student with a score of 14, there is a 68% confidence interval that the true score is between 12 and 16.

A two-parameter IRT analysis was used to identify items that did not fit the model. The two-parameter model estimates difficulty and discrimination parameters, but does not estimate lower or upper asymptotes (indicating guessing or slips, respectively). Item response curves can indicate where the items may not conform to the model. Some of the item response curves had an upper asymptote of 0.6, indicating features of these items were causing even high-ability students to "slip" and provide incorrect responses. However, many items fit the model well, suggesting that items could differentiate between students of different ability.

Claim 2: Mastery of particular concepts. Tetrachoric correlations showed that many of the items were weakly related. Several items correlated negatively with other items, indicating that students who answered one item correctly were less likely to answer another correctly.

Based on these analyses, we decided to remove four items for the remaining analyses. The four items had higher alpha-with-item-deleted values and correlated negatively with the other items.

Three of these items also had low item discrimination values and did not fit the two-parameter IRT model.

We calculated subscale reliabilities for the categories with more than one item. Subscale reliabilities were too low for the majority of the categories to warrant subscale reporting.

We ran an exploratory factor analysis on the remaining items. A parallel analysis indicated an eight-factor structure. We used an oblique rotation to allow for correlation among constructs and suppressed factors less than 0.3. The resulting output was matched to the developer's conceptual categories. Five of the eight factors could be identified, while the remaining three did not have clearly designated categories. Eight of the 25 items did not load onto any factors. Although this analysis suggests that there is some alignment of the items to the developer's conceptual categories, as a whole it appears that most of the items do not map onto the developer's categories. Also, given that the developer designated several categories with only one item, we could not run a confirmatory factor analysis on the data because the model would have been unidentified.

Given the results of the exploratory factor analysis, we did not conduct further analyses on the data. Assuming that the DCI data is a representative sample of the target population, these results indicate that the instrument has poor internal structural properties. More high-quality items should be added to the inventory before conducting a confirmatory factor analysis or attempting diagnostic classification modeling.

Claim 3: Propensity for misconceptions. We tried to conduct a distractor analysis on the data. In some cases, too few students picked distractors related to common misconceptions to make the data a reliable indicator of student misunderstanding. In other cases, it appeared that some of the students had more difficulty with the wording of the item than to an ascribed misconception. Overall, a more extensive domain analyses and a larger sample size were needed to make reliable assertions of propensity for misconceptions.

The results of these analyses show that as a whole, the DCI has a modestly reliable total score and standard error of estimation. However, the structural analyses did not support the developer's designated categories. DCI users could not reliably report examinee sub-scale scores. Removing items is not sufficient to increase overall cohesiveness among items; more high-quality items need to be added to allow for reliable reporting of sub-scale functioning.

Concept Assessment Tool for Statics

The second CI investigated was the Concept Assessment Tool for Statics (CATS).[8] The CATS consists of 27 multiple-choice items representing nine total concepts. The data comprises 1,372 cases across several samples.

Claim 1: Overall mastery. The mean score on CATS was 12.8 out of 27 (SD=5.5). Item difficulties ranged from 0.25 to 0.78, except for one item with a difficulty of 0.16, which was more difficult for students. Except for this one item, all fell within the recommended range of difficulty. The item discrimination measure ranged from 0.20 to 0.65, except for one item with a value of 0.18. These measures indicate that each item's score is positively related to the overall proficiency represented by the total score.

The reliability for CATS was strong (α=0.89). Three items had the same alpha-with-item-deleted value to the overall alpha. When we removed these three items from the test pool, the overall alpha remained the same. This suggests that these items may not cohere as well with the rest of the assessment but, given their performance on the other psychometric analyses and recommendations from the developer, they need not be dropped from the inventory.

The standard error of the estimate for CATS was 2.02. This means that given the mean score of 13, students with a score of 11 and 15 cannot be inferred with a 68% confidence to have different true scores.

A two-parameter IRT model indicated that all the items fit the model except for two. One item had a high probability of being answered correctly by low-ability participants, which suggests that the examinees were guessing. The other item had a lower probability of being answered correctly by high-ability participants, indicating that features of the item misled examinees. However, overall the results suggest that the items measured a wide-range of abilities and could differentiate between high and low-performing examinees.

Claim 2: Mastery of particular concepts. We performed structural analyses on the data to determine if the structure of the assessment conformed to the developer's pre-defined constructs. Tetrachoric correlation were calculated for item pairs. There were strong correlations (>0.5) for items within categories, supporting the hypothesis that these items should be related.

For the remainder of the analyses, we removed one item because it correlated weakly with the remaining items. We calculated sub-scale reliabilities for all the nine concepts. Most of the groups yielded a moderate alpha measure for 3 items, between 0.33 and 0.72. For seven of the nine groups, the alphas supported reliable subscale reporting.

Next, we performed an exploratory factor analysis to determine if the data supported multiple domain concepts. A parallel analysis showed that an eight-factor solution was optimal for the number of components. We used an oblique rotation and suppressed factors less than 0.3. With a few exceptions, the resulting eight factors aligned closely with the developer's nine concepts. One possible reason that these items that did not align perfectly with the pre-defined constructs may be that they require overlapping, complex skills that are not encapsulated by the categorical designations.

Given the favorable structural properties of the assessment as per the exploratory factor analysis, a confirmatory factor analysis was run on the data. We first investigated the developer's hypothesized model using an independence model, in which none of the factors were correlated. We then tested a higher order model, which adds a single, higher order factor to the independence model. The latter model fit the data better, with the performance indices within the recommended ranges. This result suggests that the concepts in CATS are differentiable but still related in terms of reflecting a general conceptual understanding of the domain of statics, which supports the developer's claims. These CFA results suggest that other higher order models could be applied to further investigate structural and item properties of CATS. In particular, a bifactor IRT may be a useful supplementary approach to analyze items using an IRT method without the additional restriction of assumptions about unidimensionality.

Claim 3: Propensity for misconceptions. We tried to analyze item distractors to investigate common misconceptions. However, we could not determine with any certainty that students were picking particular distractors because they possessed certain misconceptions. This is because several distractors for the same item mapped onto the same misconception. As a result, it is difficult to determine if students are picking distractors because of item context, features of the item, erroneous thinking, or misconceptions.

Overall, these analyses indicate that the CATS data provide strong evidence for eight or nine constructs, as designated by the developer. The confirmatory factor analysis supported a higher order or general statics knowledge factor. The subscale alphas showed that seven of the nine constructs had an adequate reliability. These findings support the claim that the sub-scores are reasonable measures of conceptual understanding if used for low-stakes purposes. The analyses also provide evidence that the total score on CATS can be used as a single, summative measure of statics knowledge.

Conclusion

Conceptualizing assessment validity as an evidentiary process is an important step in ensuring that assessments measure what their developers claim they are supposed to measure. Assessment scores should also be used and interpreted in ways that are consistent with the developer's intent and existing evidentiary support. This paper presents an analytic framework for evaluating validity arguments, and links the Evidentiary Validity Framework to evidence-centered design. We presented two case studies of different CIs within the domains of engineering to show how to apply this framework. These cases show how evidence can be used to support the claims of developers, in addition to the uses of inventory scores for evaluating student performance and educational interventions. We also provided suggestions on how CIs can be better constructed from the outset.

It should be noted that there are additional claims one might wish to make about the use of CI scores, and these claims would require additional validity evidence. A developer may assert that the instrument has predictive validity or that the instrument is sensitive to changes in instructional practice. To demonstrate predictive validity, for example, a developer could correlate CI scores with performance on other course assignments. These course assignments would also need to be rigorously evaluated to ensure that they measure the intended constructs, and do so without introducing construct-irrelevant variance. Regardless of the specific claims one makes, validation is the process of finding appropriate evidence to construct warrants for the designated assertions. Overall, the Evidentiary Validity Framework can help guide the design and refinement of assessments intended to provide valid empirical evidence about student thinking and understanding in engineering domains.

Acknowledgements
We thank the National Science Foundation for funding this work through grant NSF 0918531, 0918436, 0918552, and 0920242. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

References

1. James W. Pellegrino, Louis V. DiBello, Katie James, Natalie Jorion, and Lianne Schroeder, “Concept inventories as aids for instruction: A validity framework with examples of application,” *Proceedings of the Research in Engineering Education Symposium*, Madrid. (2011), 719-27.
2. Natalie Jorion, Brian Gane, Katie James, Lianne Schroeder, Louis V. DiBello and James W. Pellegrino, “An Analytic Framework for Evaluating the Validity of Concept Inventory Claims,” *Journal of Engineering Education* 106 (forthcoming).
3. Mitchell Nathan and Kenneth R. Koedinger, "An investigation of teachers' beliefs of students' algebra development," *Cognition and Instruction* 18, no. 2 (2000): 209-237.
4. Mitchell Nathan and Anthony Petrosino, "Expert blind spot among preservice teachers," *American Educational Research Journal* 40, no. 4 (2003): 905-928.
5. Michael T. Kane, "Validating the interpretations and uses of test scores," *Journal of Educational Measurement* 50, no. 1 (2013): 1-73
6. Robert Mislevy, Linda S. Steinberg, and Russell G. Almond, "Focus article: On the structure of educational assessments," *Measurement: Interdisciplinary research and perspectives* 1, no. 1 (2003): 3-62.
7. Gary Gray, Francesco Costanzo, Don Evans, Phillip Cornwell, Brian Self, and Jill L. Lane. "The dynamics concept inventory assessment test: A progress report and some results," In *American Society for Engineering Education Annual Conference & Exposition* (2005).
8. Paul Steif and John A. Dantzler, "A statics concept inventory: Development and psychometric analysis," *Journal of Engineering Education* 94, no. 4 (2005): 363-371.
9. James W. Pellegrino, Naomi Chudowsky, and Robert Glaser, *Knowing what students know*: *The Science and Design of Educational Assessment.* Washington, DC (2001).
10. American Educational Research Association, American Psychological Association, National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing (U.S.), *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association (2014).
11. Michael Kane, "Content-related validity evidence in test development," *Handbook of test development* (2006): 131-153.
12. Samuel Messick, "Meaning and values in test validation: The science and ethics of assessment," *Educational researcher* 18, no. 2 (1989): 5-11.
13. James W. Pellegrino, Louis DiBello, Ronald Miller, Ruth Streveler, Natalie Jorion, Katie James, Lianne Schroeder, and William Stout, “An analytical framework for investigating concept inventories,” In J. Pellegrino (Chair), *The Conceptual Underpinnings of Concept Inventories.* Symposium conducted at the meeting of the American Educational Research Association, San Francisco, CA. (2013).
14. Susan Singer, Natalie R. Nielsen, and Heidi A. Schweingruber, eds., *Discipline-based education research: Understanding and improving learning in undergraduate science and engineering*. National Academies Press (2012).
15. Ruth Streveler, Ronald Miller, Aidsa Santiago-Román, Mary A. Nelson, Monica R. Geist, and Barbara M. Olds, "Rigorous Methodology for Concept Inventory Development: Using the 'Assessment Triangle' to Develop and Test the Thermal and Transport Science Concept Inventory (TTCI)," *International Journal of Engineering Education* 27, no. 5 (2011): 968.
16. Dana Denick and Ruth Streveler, "Qualitative analyses of students’ conceptual reasoning," In JW Pellegrino (Chair), *Evaluating and Improving Concept Inventories as Assessment Resources in STEM*

Teaching and Learning. Symposium conducted at the meeting of the American Educational Research Association, Philadelphia, PA. (2014).

17. Erin Bardar, Edward E. Prather, Kenneth Brecher, and Timothy F. Slater, "The need for a Light and Spectroscopy Concept Inventory for assessing innovations in introductory astronomy survey courses," *Astronomy Education Review* 4, no. 2 (2005): 20-27.
18. Lee Cronbach, "Coefficient alpha and the internal structure of tests," *Psychometrika* 16, no. 3 (1951): 297-334.
19. Linda Crocker and James Algina, *Introduction to classical and modern test theory*, Holt, Rinehart and Winston, Orlando, FL (1986).
20. Albert Oosterhof, *Classroom applications of educational measurement*, Prentice-Hall, Inc., Upper Saddle River, New Jersey (2001).
21. Philip Sadler, Harold Coyle, Jaimie L. Miller, Nancy Cook-Smith, Mary Dussault, and Roy R. Gould, "The astronomy and space science concept inventory: development and validation of assessment instruments aligned with the k–12 national science standards," *Astronomy Education Review* 8, no. 1 (2009).
22. David Hestenes, Malcolm Wells, and Gregg Swackhamer, “Force concept inventory,” *The physics teacher,* 30, no. 3 (1992): 141-158.