International Journal of Science Education
2010, 1–24, iFirst Article

RESEARCH REPORT

Development and Validation of Instruments to Measure Learning of Expert-Like Thinking

Wendy K. Adams[a,b]* and Carl E. Wieman[c,d]
[a]*Department of Physics, University of Northern Colorado, Colorado, USA;* [b]*Acoustical Society of America, New York, USA;* [c]*Physics and Science Education Initiative, University of British Columbia, Vancouver, Canada;* [d]*Physics and Science Education Initiative, University of Colorado, Boulder, USA*

This paper describes the process for creating and validating an assessment test that measures the effectiveness of instruction by probing how well that instruction causes students in a class to think like experts about specific areas of science. The design principles and process are laid out and it is shown how these align with professional standards that have been established for educational and psychological testing and the elements of assessment called for in a recent National Research Council study on assessment. The importance of student interviews for creating and validating the test is emphasized, and the appropriate interview procedures are presented. The relevance and use of standard psychometric statistical tests are discussed. Additionally, techniques for effective test administration are presented.

Keywords: *Assessment; Formative assessment; University; Science education; Evaluation; Assessment design*

In recent years, there has been a growing effort to develop assessment tools that target students' development of expert-like mastery of specific science topics. These involve questions that accurately probe whether students understand and apply particular concepts in the manner of a scientist in the discipline. Such assessment tools are intended to measure student learning in courses to provide Formative Assessment of Instruction (FASI). We present the methodology involved in developing and validating such assessment tools. This same methodology works equally well

*Corresponding author. Department of Physics, University of Northern Colorado, Campus Box 127, Greeley, Colorado 80639, USA. Email: wendy.adams@colorado.edu

ISSN 0950-0693 (print)/ISSN 1464-5289 (online)/10/000001–24

DOI: 10.1080/09500693.2010.512369

for developing assessment tools to measure other aspects of student thinking, such as their perceptions of a field of science and how it is best learned.

In this paper, we will describe how a faculty member or education researcher can create a valid and reliable assessment tool of this type. We have found that a post-doctoral researcher in the field of the topic content with a few months experience in education research training can carry out such a process. Once this tool is created and validated, it provides a way to compare instruction across institutions and over time in a calibrated manner. The methodology we describe for test construction matches well with the 'Standards for educational and psychological testing' (American Educational Research Association [AERA], American Psychological Association [APA], & the National Council on Measurement in Education [NCME], 1999) and is closely aligned with what is called for in the National Research Council's study of assessment (NRC, 2001). We provide a detailed discussion on how to implement this general methodology in specific domains. We will also discuss appropriate statistical analyses that are part of validity evidence and how interpretation of these statistics depends on the nature of the assessment.

One particularly important part of both the development and validation of a FASI is the use of student interviews. There is a large body of literature on the use of student interviews for the purpose of understanding student thinking (Berardi-Coletta, Buyer, Dominowski, & Rellinger, 1995; Ericsson & Simon, 1998), but student interviews are rarely used when developing educational tests, although the value of this kind of information is stressed in the 2001 NRC report. 'The methods used in cognitive science to design tasks, observe and analyze cognition, and draw inferences about what a person knows are applicable to many of the challenges of designing effective educational assessments' (NRC, 2001, p. 5). Here, we will discuss how to use this technique from cognitive science in developing and validating a FASI.

Beginning with the Force Concept Inventory (FCI; Hestenes, Wells, & Swackhamer, 1992), a number of FASI-type instruments have been developed to measure student learning of science at the university level in a systematic way, and these are having a growing impact on teaching and learning. While there are similarities in various groups' approaches to the development and validation of such instruments, the procedures are not fully consistent partially due to the fact that, to our knowledge, no one has written down the process for this specific type of instrument in its entirety. As an example, while describing the steps for evaluating student misconceptions on a particular topic, Treagust (1988) writes about the value of conducting student interviews to determine misconceptions and distracters. However, after reviewing 16 FASI-type instruments for physics, chemistry, geosciences, and biology that were developed in the past 20 years only Redish, Steinberg, and Saul (1998) and Singh and Rosengrant (2003) used student interviews during both the development and validation processes. The developers of the FCI did student interviews after developing the survey; however, the results of those interviews showed that student reasoning was not consistent 'but Newtonian choices for non-Newtonian reasons were fairly common' (Hestenes et al., 1992, p. 148), raising some questions about validity. For this reason, the authors go on to state that scores below 80% can only be used as an upper bound on student understanding. Another reason that development procedures are not consistent could stem from the fact that developers of FASI-type instruments are necessarily content experts so may not be familiar with work in cognitive science or the field of assessment design. For example, the FCI authors discuss grouping of questions that address different aspects of the content but were later criticized for not using statistical measures such as a factor analysis to support these groupings (Heller & Huffman, 1995; Huffman & Heller, 1995) and since have retracted the suggestion of using groups of statements and recommend looking at only the total score (Hestenes & Halloun, 1995). In most of the other examples listed, the appropriate psychometric statistical tests have either not been carried out or the results of such tests were misinterpreted, often due to confusion as to the distinctions between single-construct and multiple-construct assessments. Finally, the validity and reliability of any results obtained with FASI-type instruments depend on how it is administered, but test administration options and tradeoffs are seldom if ever discussed in the literature presenting the instruments.

This paper is intended to describe the complete process of developing a FASI based on previous work and our own experience in such a way that a content expert can create a valid and reliable instrument for their discipline. Individually, and with collaborators, we have now developed nine assessments that test expert-like mastery of concepts; four have been published (Chasteen & Pollock, 2010; Goldhaber, Pollock, Dubson, Beale, & Perkins, 2010; McKagan, Perkins, & Wieman, 2010; Smith, Wood, & Knight, 2008) and several that measure expert-like perceptions about the learning and application of various science subjects with two that are published (Adams et al., 2006; Barbera, Perkins, Adams, & Wieman, 2008). This process has now become relatively refined and straightforward.

Development Framework

Our process, as discussed in detail below, follows the four phases outlined in the Standards for Psychological and Educational testing (AERA, APA, & NCME, 1999, p. 37):

> Phase 1. Delineation of the purpose of the test and the scope of the construct or the extent of the domain to be measured;
>
> Phase 2. Development and evaluation of the test specifications;
>
> Phase 3. Development, field testing, evaluation, and selection of the items and scoring guides and procedures; and
>
> Phase 4. Assembly and evaluation of the test for operational use.

Phase 1

The basic theoretical idea behind all these instruments is that there are certain ways of thinking that experts within a particular subject share. 'Studies of expert-novice differences in subject domains illuminate critical features of proficiency that should be the targets for assessment' (NRC, 2001, p. 4). Ericsson and others have identified how experts have particular mental structures by which they organize and apply information (Ericsson, Charness, Feltovich, & Hoffman, 2006). Expert chess players see patterns in the arrangement of the pieces that tell them how the game is progressing and identify optimum strategies. Physicists organize knowledge around fundamental concepts. When faced with a problem, they recognize patterns in the problem that cue concepts that will be productive for working out a solution. Identifying the unique characteristics that make up specific areas of expertise in different disciplines is an active field of research that is extensively reviewed in Ericsson et al. (2006). Expert thinking, however, goes beyond how information is organized and applied to include discipline-specific heuristics for monitoring problem-solving and other aspects of thinking such as perceptions of how the subject is best learned and where it applies. For example, physicists perceive physics as describing the real world and that it is best learned in terms of broadly applicable concepts. When learning something new, they believe that it should be carefully examined as to how it makes sense in terms of prior knowledge of physics and the behaviour of the world.

A suitable educational goal is to have students thinking more like experts and approaching the mastery of the subject like an expert, and so it is desirable to have test instruments that measure student thinking on a scale that distinguishes between novice and expert thinking (Shavelson & Ruiz-Primo, 2000). This requires a process for first identifying consistent expert thinking and then creating a valid test for measuring the extent to which students learn to think like experts during their instruction in any particular course.

The NRC calls for having three elements in the foundation to all assessments: *cognition, observation, and interpretation* (NRC, 2001). The expert–novice differences of value to teachers define the cognition that is being probed by the FASI; the questions themselves are the tasks that elicit this thinking (observations); and the validation process—demonstrating that the FASI measures what is intended—determines the interpretation by delineating what inferences about student thinking can be drawn from the results of the FASI.

There are several practical requirements for such an instrument if it is to be used on a widespread basis so it can impact instruction. (1) It must measure value added by the instruction, and hence it must be possible to administer it on both pre- and post-instruction basis. (2) It must be easy to administer and grade in the context of a normal course schedule without any training. It is typical to have to make some tradeoffs between the breadth and depth of what is measured and ease of administration, but this is possible to do without seriously compromising the validity. (3) The instrument must test 'expert thinking' of obvious value to teachers. (4) It needs to be demonstrated that it measures what it claims (evidence of validity).

It is unrealistic to have a test that meets goals 1–3 above and tests everything of importance for students to learn in any given course, particularly in a format that can be easily given and graded in a small amount of time. So a more practical goal is to test mastery of a limited set of important, hard-to-learn concepts. Usually these will serve as an effective proxy for measuring how effectively concepts are being learned in the course as a whole. This is supported by the evidence that certain pedagogical approaches are superior in supporting conceptual learning, independent of the particular concept being taught (Bransford, Brown, & Cocking, 2000 and references therein). In selecting which concepts to include, consideration should be given to maximizing the range of institutions where the test can be used. The FCI is a good example of a test that is designed according to this principle. It covers a relatively small subset of the concepts of force and motion that are covered in a typical first-term college physics course. However, this particular set of concepts is taught and valued by nearly everyone, students have particular difficulty with them, and do worse than many teachers expect. The results from this test have been sufficiently compelling to cause many teachers to adopt different instructional approaches for all of their introductory physics course material.

Phase 2

Development and evaluation of the test specifications include: item format, desired psychometric properties, time restrictions, characteristics of the population, and test procedures.

According to classical and modern test theory: 'In test construction, a general goal is to arrive at a test of minimum length that will yield scores with the necessary degree of reliability and validity for the intended uses' (Crocker & Algina, 1986, p. 311). Although the intended use of a FASI is to measure how well students are thinking like experts, the primary goal is not to obtain a summative assessment of student learning; rather it is to provide formative assessment of teaching. Thus the results from the students as a group are more important than ranking individual students, which is fundamentally different from many other assessments. This makes this a low-stakes assessment and the testing approach must be tailored to this use. It also relaxes a number of constraints on test design (NRC, 2001). In addition to not needing to cover all of the course material, the instrument will also often be probing many different facets of thinking and learning, rather than a particular block of content (a single construct), such as how well the student can describe the molecular anatomy of genes and genomes. This makes the test design and statistical tests of the instrument rather different.

Psychometricians typically use either *item analysis* or Item Response Theory (IRT) to determine which items in the pool of potential test items will be the best to construct the most efficient, reliable, and valid test. However, the standard acceptable range of values for these statistical tests was determined for single construct, and summative tests intended to provide maximum discrimination among individual students. Because a FASI serves a much different purpose, statistical-item analysis takes a secondary role to student interviews, which provide much more information about the validity of the questions, as well as often indicating why a question may have statistics that fall outside of the 'standard' psychometric range. That makes it particularly essential that the students interviewed for both the development and validation of the test represent the breadth of the population for which the test is to be used.

When creating a FASI, it's desirable to use a forced answer (multiple-choice or likert-scale) test. This format is easy to administer and to grade. Also, unlike open-ended questions that are graded with a rubric, it easily provides reliably consistent grading across instructors and institutions. To be useful, FASIs usually need to be administered in a pre- and post-format to normalize for initial level of knowledge. We have found that surveys of perception, which can be completed in less than 10 minutes, can be given online, while conceptual tests (20–50 minutes) need to be given during class time, with a careful introduction, so that students take the test seriously in order to obtain a useful result. See Test Administration Tips at the end of this paper for more details.

Phases 3 and 4

These two aspects of test construction comprise the bulk of the work needed to develop a new FASI. Here, we will list and then describe in detail the six steps that are required to develop and validate the test. These six steps are undertaken in a general order; however, it is usually an iterative process. Often part of the validation procedure reveals items that need modification, and then it is necessary to go back and create and validate these new or modified items. The entire process takes very roughly a half person year of a good PhD level person's time to carry out. That effort will likely need to be spread out over one to two years of calendar time, due to constraints of the academic schedule and availability of student test subjects. Key elements of the process are the student interviews that are carried out at Steps 2 and 5. Although both sets of interviews largely rely on a think-aloud format, they are distinctly different. The validation interviews in Step 5 are far more limited in scope than the interviews done in Step 2. Step 5 interviews follow a much stricter protocol as discussed below.

(1) Establish topics that are important to teachers (in our case, college or university faculty members).
(2) Through selected interviews and observations, identify student thinking about these topics and the various ways it can deviate from expert thinking.
(3) Create open-ended survey questions to probe student thinking more broadly in test form.
(4) Create a forced answer test that measures student thinking.
(5) Carry out validation interviews with both novices and subject experts on the test questions.
(6) Administer to classes and run statistical tests on the results.

Modify items as necessary.

Establish aspects of thinking about the topic that are important to faculty members. 'Educational assessment does not exist in isolation, but must be aligned with curriculum and instruction if it is to support learning' (NRC, 2001, p. 3). Establishing important aspects of thinking on the topic is usually the easiest step for a test designer who knows a subject well. These can be done through a combination of the following four methods: (1) reflect on your own goals for the students; (2) have casual conversations or more formal interviews with experienced faculty members on the subject; (3) listen to experienced faculty members discussing students in the course; and (4) interview subject matter experts. One can pick up many aspects of the topic that are important to faculty members who are experienced teaching the subject simply from casual conversations. Merely ask them to tell you what things students have done and said that indicated they were not thinking as the faculty member desired. Normally what is desired is to have the student thinking like an expert in the discipline. An even better source of information is to listen to a group of thoughtful teachers discussing teaching the subject and lamenting where students are not learning. This automatically picks out topics and thinking where there is a consensus among teachers as to what is appropriate for students to learn; otherwise they would not be discussing it. Such discussions also identify where there are conspicuous student difficulties and shortcomings in achieving the desired mastery. This information can be collected more systematically by interviewing faculty members who are experienced teachers to elicit similar lists of expert thinking that students often fail to achieve. As an example, a number of faculty members across multiple domains have mentioned domain-specific versions of ‘When the student has found an answer to a problem, be able to evaluate whether or not it is reasonable’.

Interviewing content experts, even if they have not taught extensively, such as research scientists, in a systematic fashion about the topics under consideration can also be valuable. The ideas that come up consistently in such interviews represent the core ideas that are considered important in the discipline. In our studies in physics, we see significant evolution of thinking between graduate students and faculty members about many fundamental topics in physics, and so for physics and likely other subjects, graduate students should not be considered to be subject experts (Singh & Mason, 2009).

Usually interviewing 6–10 faculty members/content experts is quite adequate to determine the important relevant expert thinking. FASIs are useful only if they focus on ideas that are shared by an overwhelming proportion of experts. So if something does not emerge consistently from interviewing such a small number of experts, it is best not to include it. In this step and the following, one is carrying out the steps called for in the NRC report: ‘Understanding the contents of long term memory is especially critical for determining what people know; how they know it; and how they are able to use that knowledge to answer questions, solve problems, and engage in additional learning’ (NRC, 2001, p. 3).

Selected interviews and observations to understand student thinking about these topics and where and how it deviates from expert thinking. Interviewing and observing students to understand thinking about the topics determined in Step 1 above is in accordance to the call for developing additional cognitive analyses of domain-specific knowledge and expertise (NRC, 2001). As mentioned above, much of the non-expert-like student thinking that is of concern will already be revealed by thoughtful experienced teachers, which in many cases, will include the test developers. In some fields, there is also a significant literature on student misconceptions that is highly relevant. However, it is important to also do direct student observations such as observing and participating in course help sessions and conducting systematic student interviews. These are likely to reveal quite unexpected student thinking and perspectives if one listens carefully and avoids 'filling in blanks' with one's expert knowledge and assumptions.

Course help sessions (a session, often optional, where students come in to work together on homework, often with some form of assistance by TAs) provide a wealth of information on student knowledge, how it is connected, and in what context students apply it. Help sessions are a very useful setting to start understanding student thinking because they have a lot of students working on the course material in one public location, allowing for efficient collection of information. Using the help session to specifically look for topics where student thinking deviates from the experts' often reveals information that may have gone unnoticed previously. Begin by simply observing, listening to student discussions, and taking notes—particularly noting what content students think relates to particular homework questions, and how they use this content. After collecting observational data on how students respond to questions without researcher interference, it is often useful to work with a few students and carefully probe more deeply into what they are thinking both during informal help session interviews, or more formal arranged interviews.

To carry out interviews, one starts by soliciting student volunteers that span the full range and beyond (age, gender, ethnicity, background preparation, grade point average, career aspirations, etc.) of the student population to be tested. It is important to interview as broad a range of students as possible. Differences in thinking, both between different students and between students and experts become much more obvious when one looks at extreme cases. As there can be aspects of student thinking that an expert would never guess, because the perspective is so different, the enhanced 'signal' provided by considering extreme cases is almost always very helpful. For example, when discussing the idea of an electric or magnetic field with first-year physics or other science majors, they typically have an idea of some sort of invisible force; however, when discussing this idea with non-science majors we found that many visualized a sort of field similar to a farm field or field of grass. Another example came from a general science course for elementary education majors. The first author was extremely surprised to discover that many of these students believe that the continents float on the Earth's oceans.

Student interviews can take a variety of forms, but ideally one will pose the student a question or problem on a subject where there is a clear expert consensus, and have them simply explain it or solve it, using a 'think-aloud' protocol (Ericsson & Simon, 1998, p. 182). A think-aloud protocol is where the student is told to think aloud while working a particular problem. The interviewer is restricted to prompting the students to think aloud when they become quiet but must not ask the students to *explain* their reasoning or ask them *how* or *why* they did something. Once questions of this nature are asked, the students' thinking is changed when they attempt to answer—usually it is improved (Berardi-Coletta et al., 1995; Ericsson & Simon, 1998). However, for a think-aloud interview to be successful, the material that the student explores or solves must be very carefully chosen. Since Step 2 is one of the initial steps of FASI development and the goal here is to *find* the material that is appropriate for your test, it's typically necessary to deviate at times from the think-aloud protocol to ask questions that probe certain areas more directly. These interviews may require the interviewer to ask students to explain their answer or to say more about what a particular term or concept means to them. Specific sorts of questions that can be useful are to ask students to explain how they think about or answer exam or homework questions, particularly where students in a course do worse than the teacher expected. Topics that teachers have seen student difficulties or misconceptions on are also useful to ask about. If creating a test about perceptions, it is useful to explore things that students say about how they study or learn the subject, or what they see as constituting mastery in the subject that are in conflict with what experts believe. For example, 'I am well prepared for this physics test, I made up flashcards with all the formulas and definitions used in this course and have spent the last week memorizing them'.

During the interview, it is important to get the students to say what they think, rather than what they think the interviewer wants to hear, so one needs to be careful to avoid cuing students to think or respond in a certain way. An interview should never be a conversation. Deviating from the strict think-aloud protocol is more challenging because the interviewer still must spend most of his/her time simply listening quietly and resisting any urge to interject; however, occasional prompts are necessary. For example, to ask, once a student feels that s/he has finished an answer, 'why did you choose to use this process' or 'what does this term mean to you?'. These are very minimal probes but enough to flesh out the details of why students chose the concepts/strategies that they used and what those mean to them. Because these sorts of probing questions do alter student thinking and could likely help students think of connections they may not have in an actual testing situation, strict think-aloud interviews must be performed for validation once the test is constructed. See Section 'Carry out validation interviews on test questions' for details.

It is often very useful to have an independent source listen to the recording or watch the video of an interview, particularly with an inexperienced interviewer, to see if the interviewer participation was kept to a minimum and the interpretation of the student responses was accurate. When possible, it is even better to have an experienced interviewer sit in with the new interviewer for the first one or two interviews. Students respond quite positively to being interviewed with two interviewers when one is in training. We have seen that in addition to having good interview skills, an interviewer must also have a high level of subject expertise in order to properly interpret what students are saying about the content, particularly to detect when it may deviate from expert-like thinking in subtle ways.

All interviews should be recorded either via audio or video with audio. Immediately following each interview and before the next interview, the interviewer should spend roughly half an hour summarizing the interview. Some interviewers find it useful to listen to or watch the recordings to check their summaries, but we find this becomes redundant with experienced interviewers who listen closely. As one is simply trying to get consistent general ideas of student thinking from such interviews, it is seldom worth the time and expense of transcribing and individually coding such interviews.

Create open-ended survey questions to probe student thinking more broadly in test form. Once patterns in student thinking begin to appear, then the data from help sessions and interviews can be coded more systematically to identify the type and frequency of student thinking about any particular topic. One should pay particular attention to student thinking that is most conspicuously different from that of experts in ways that faculty members may not expect. FASI questions that demonstrate such unexpected thinking are likely to have the most impact on instruction because they surprise and inform the teacher. These questions are also often the most sensitive to improved methods of instruction, because the faculty members did not realize the problem and hence were doing little to address it. As an example, chemistry faculty members at the University of Colorado created a short FASI-type instrument for physical chemistry. One of the questions addressed the notoriously difficult concept area of gas laws and isotherms. After seeing poor results one semester on this question (only a 5% increase in score), the faculty member developed a series of clicker questions to help the students work through the concepts and had a 44% gain the following semester.

It is difficult to give a simple number for how many students should be interviewed, as this depends so much on the topic and the student responses. If there is a great range of student thinking on a particular topic it will require more interviews. Twenty is probably a reasonable upper limit however. If certain common patterns of thinking have not emerged with 20 interviews, it probably means that the thinking on the topic is too diverse to be measured, and hence the topic is not suitable for such an instrument. More typically, within a dozen or so interviews, one will have a sufficiently clear impression to know which topics should be written into FASI questions.

Guided by the expert and student interviews, one then creates open-ended questions that probe student thinking on topics where there are apparent discrepancies between expert thinking and that of students. These open-ended questions are then given to an entire class. This tests, with a larger sample, the issues that arose in the student interviews. When possible, phrasing the question in terms of actual student wording that emerged during interviews is most likely to be interpreted as desired by the students. Examples of productive questions are 'Describe what an electric field is and how you picture it' or 'How long did this feature take to form?' (see Figure 1).

The responses from the class(es) to these questions are then systematically coded according to the patterns of student thinking, similar to what is done for the interviews. These open-ended responses from students provide the best source of possible answer choices to be used in the FASI multiple-choice test.

Create forced answer test that measures student thinking. As discussed above, there are major practical benefits to a FASI composed of multiple-choice questions. When appropriate distracters have been chosen, the test not only easily provides the teacher with information about how many students get the answers correct but also provides information on the incorrect thinking. This matches the NRC guidelines: 'assessments … should focus on making students' thinking visible to both their teachers and themselves so that instructional strategies can be selected to support an appropriate course of future learning' (2001, p. 4).

Classroom diagnostic tests that are developed to assess individual students on a particular topic sometimes use other question formats including two-tier. These assessments are focusing on representative coverage of concepts within a topic area so that the researcher/instructor can characterize in detail student learning on a particular topic. Two-tier questions first ask a multiple-choice question—typically a factual type question with only two possible answers. This is followed by asking a second multiple-choice question about the 'reason'. Such two-tiered questions, while valuable for guiding instruction, are not ideal for the goals of a FASI-type instrument, because FASI instruments aim to have a minimum number of questions, all of which focus on student reasoning, and so would be compromised predominantly of the second portion of the two-tier questions. Brevity and ease of scoring and interpretation are more important for a FASI than detailed characterization of student learning. For these reasons, two-tier questions are outside the scope of this paper.

The primary challenge in creating good multiple-choice questions is to have incorrect options (distracters) that match student thinking. Typically three to five distracters are offered, although there are exceptions. Actual student responses during interviews or to open-ended questions are always the most suitable choices for the multiple-choice question responses, both incorrect and correct (Figure 1). This is the language students are most likely to understand as intended. If one cannot find suitable distracters from among the student responses, then probably the question is unsuitable for use in a multiple-choice form. Care should be taken that wording of the distracters does not inadvertently steer the students toward or away from other answers if they are using common test-taking strategies (Figure 2). For example students avoid options if the answer does not seem 'scientific' or involves a statement of absolutes such as 'never' or 'always'. They also look to see which choices have different length or grammatical form.

Some teachers are bothered by providing distracters that are so inviting to students that they can score lower than if they simply selected answers at random. However, that emphasizes the fundamentally different purpose between these tests and standard summative tests used in classes. Here the purpose is to determine

How long did this feature take to form? Choose the *BEST* answer.

(a) minutes or less

(b) 100s of years

(c) 10s of 1,000s of years

(d) 100s of 1,000s of years

(e) 1,000,000s of years or more

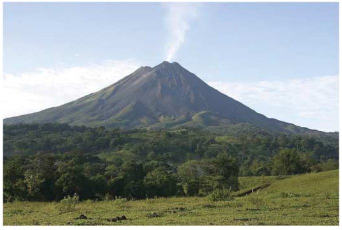

Figure 1. Multiple-choice question from the LIFT (Landscape Identification and Formation Test) created after using open-ended questions to collect student responses (Jolley, 2010). Image Copyright© Shmuel Spiegelman, Image Source: Wikimedia Commons http://commons.wikimedia.org/wiki/File:Arenal-Volcano.jpg

student thinking, and so all options should be very inviting to at least some students. The incorrect options should match established student thinking, which should result in student answers that are as non-random as possible.

When creating possible answers for multiple-choice questions, a danger to avoid is the creation of answers that have multiple parts. A particularly common failing is to ask questions of the form, ‘A is true because of B’. There are two problems with this form of question. First, it takes much more time for a student to read and interpret because it requires combining multiple ideas that may be tightly linked for an expert in the subject but are far less closely linked for the student. Second, the student may well believe A is true, but not for the reason given in B, or s/he may believe the reason B but not that A is true. So the student will then need to struggle over how to interpret the question and what answer s/he should give, and it makes interpretation of his/her responses problematic. We have consistently seen difficulties with these types of multiple-part answers in our validation interviews.

In creating a test that measures perceptions, rather than concepts, it is typical to have the item format be statements rather than questions. Students respond on a likert-scale ranging from ‘strongly agree’ to ‘strongly disagree’ (Crocker & Algina, 1986; Kachigan, 1986). Because these take students much less time to answer than conceptual questions, the upper limit for a perceptions survey is about 50 simple clear statements. The limitations on how FASIs can be administered argue strongly against any concept test requiring more than 30 minutes for all but the slowest students to complete. This typically means 20–30 clear multiple-choice questions.

Carry out validation interviews on test questions. Once multiple-choice questions have been created, they need to be tested for correct interpretation by both experts (teachers of the subject) and students. In the case of experts, one needs to verify that all agree on the correct response and that the alternative answers are incorrect. All experts also need to agree that the question is testing an idea that they expect their students to learn. Typically only 6–10 experts might be interviewed on the test itself, as there is normally a high level of consensus; however, several times that number need to take the test to ensure consistent expert answers. Teachers will often have suggestions on making the wording more accurate. It is not unusual to have situations where teachers want to word questions to be more technical and precise than is actually necessary or suitable for optimum probing of student thinking. It is necessary in those cases to try to find suitable compromises that teachers will accept, but are still clear to students.

It is considerably more work to produce appropriate wording for students than for teachers, often involving multiple iterations and testing with student interviews. Student interviews are necessary to verify that students interpret the question consistently and as intended. It is also necessary to verify that students choose the correct answer for the right reasons and that each wrong answer is chosen for consistent reasons. Figure 2 is an example of a question, which all experts felt had clear appropriate wording and answer choices; however, when students were interviewed they were able to choose the correct answer without any reasoning about genetics. Twenty to forty student interviews are typically required to establish validity with good statistical confidence. We stress how essential these interviews are. Unfortunately, it is not unusual to encounter test questions, even those used in very high-stakes contexts, that multiple experts have examined carefully and on that basis they are considered valid. However, in interviews on test questions, we have consistently seen that multiple experts can conclude that a question is perfect, but then students can have a completely reasonable but quite different interpretation of the question from what was intended. Ding, Reay, Lee, and Bao (2008) have also observed differing student and expert interpretation of questions.

A single DNA nucleotide change of an A to a T occurs and is copied during replication; is this change in DNA sequence necessarily a mutation?

(a) Yes, it is a change in the DNA sequence.

(b) Yes, but only if the nucleotide change occurs in a sex cell (sperm or egg).

(c) Yes, but only if the nucleotide change occurs in the coding part of a gene.

(d) Yes, but only if the nucleotide change occurs in the coding part of a gene and alters the amino acid sequence of a protein.

(e) No, because A and T are similar enough, they can substitute for each other.

Student who earned a D in genetics: ‘I don’t like to see the word only in answers. Answers with only are never true. There are four yes answers and one no, so I will go with answer (a)’.

Figure 2. Multiple-choice question from the GCA (Genetics Concept Assessment) which had to be reworded due to the correct answer being chosen for the wrong reasons during student interviews (Smith et al., 2008)

Student interviews are much more sensitive than interviews of teachers. *With teachers, you want their opinions; with students you're trying to probe their thinking*. This is where work in cognitive science can provide methods that allow the interviewer to learn about student thinking without altering it. These interviews differ from the interviews used to determine student thinking about concepts as described in Section 'Selected interviews and observations to understand student thinking about these topics and where and how it deviates from expert thinking'. Interviews on the test items must follow a strict think-aloud protocol (Ericsson & Simon, 1998). To get accurate information from the interviews it is important that the interviewer does not alter the student thought processes by asking questions. The student should be put in an environment very similar to that in which the test will be administered. Thus, the students sit down and are given the test to fill out just as if they were doing it in a classroom setting. While they are filling out the test, they are asked to think out loud. The interviewer should only say things like 'please tell me your thoughts as you go'; but, never ask them to explain *how* they interpreted each question and *why* they chose the answer they did as this is likely to alter the thought processes (Berardi-Coletta et al., 1995):

> When participants are thinking aloud, their sequences of thoughts have not been found to be systematically altered by verbalization. However, when participants are asked to describe and explain their thinking, their performance is often changed-mostly it is improved. (Ericsson & Simon, 1998, p. 182)

It requires some preparation to put people into a comfortable think-aloud mode. We always start interviews with 5 or 10 minutes of 'icebreaker'-type questions to get the student comfortable with the interviewer. For example, asking them about their major, year in school, classes they like or dislike, future career plans, etc. Often the interviewer will follow the icebreakers with some practice think-aloud exercises. This provides practice for the game of thinking aloud so that interview information from the very first question on the FASI will be valuable. It is sometimes difficult to interpret student thinking from think-aloud interviews since 'inner speech appears disconnected and incomplete' (Vygotsky, 1962, p. 139), but asking for clarification must be avoided. For example, some students use the strategy of picking an answer as soon as they see one they like without reading all the choices. If the interviewer asks for an explanation of each possible choice, they are no longer seeing the student in authentic test-taking mode.

After the interviewer and student have used the think-aloud protocol to go through the entire test, then the interviewer can go back and ask for some further explanation on each item. But these explanations must be considered with care as they do not represent the student thinking while taking the test. The responses will include some reflection and new thoughts that are generated by the interviewers' questions and the need to turn the students' thoughts into intelligible explanations. However, these follow-up explanations can provide some useful information such as why students skipped over certain options (e.g. it did not look scientific enough) or identify options that for the student have multiple ideas contained in a single answer.

By definition, these interviews only provide validation evidence for the population that is interviewed. Therefore the broader the range of students used in the validation interviews, the more broadly the FASI can be safely used. Consider interviewing students of both genders, various ethnic backgrounds, academic specializations, and academic performance. It is typical to have to go through two or three and sometimes more iterations of the questions and possible answers to find wording that is consistently interpreted as desired, so that when the correct expert answer is chosen, it is because the students were thinking in an expert-like manner, and when the students select an incorrect answer it is because they have the non-expert-like thinking the choice was intended to probe.

Administer to classes and run statistical tests on the results. The final step of the development is to administer the test to several different classes and then perform statistical analyses of the responses to establish reliability and to collect further evidence for validity. Class sizes of a few hundred or more are desirable but fewer will suffice in many cases if the statistics are handled carefully. There are many psychometric tests that will provide useful information; however, many of the commonly used statistical tests are specifically designed for assessments that measure a single construct or factor. One characteristic of FASIs is that they usually measure thinking about multiple concepts, so the results of statistical measures must be interpreted accordingly.

Reliability. Traditionally, three broad categories of reliability coefficients have been recognized (AERA et al., 1999):

(1) Administer parallel forms of the test in independent testing sessions. Then calculate an alternate forms coefficient;
(2) Administer the test to two equivalent populations and obtain a test–retest stability coefficient; and
(3) Calculate internal consistency coefficients that measure the relationship of individual items or subsets of items from a single administration of the test.

These three types of sampling variability are considered simultaneously in the more modern *generalizability theory* to create a standard error of measurement (Cronbach, Gleser, Nanada, & Rajaratnam, 1972).

Internal consistency coefficients, (3) above, can also be described as measures of task variability. Because the goals of a FASI include probing multiple concepts with a minimum number of questions and it is not designed to accurately measure the mastery of an individual student, task variability is not a good reflection of reliability of the instrument. The time required to create a parallel validated form of a FASI, as in (1), vastly exceeds the benefits. This makes (2), administering the test to two equivalent populations and obtaining a test–retest stability coefficient, also described as sampling variability due to occasions, the primary method for measuring reliability of a FASI. Note that all three forms of reliability listed above apply to the reliability of the instrument when used on a group and not for individuals. Individuals may have fluctuations that will average out when a group is evaluated as a whole.

A test–retest stability coefficient measures the consistency of test results if the same test could be given to the same population again under identical circumstances. Of course this is impossible because it would require that giving it the first time does not have any impact on the test takers or that they have not changed in any other way between the first and second administrations. However, when administering tests to large university courses (enrolment over 150), one has the ideal situation. The test can be administered again the following year to the same course. The population of students who enrol in a course is very similar from year to year if the university maintains constant admissions criteria. Each year's students have similar preparation for the course, similar experience in college and are of similar demographic composition from year to year. The FASI should be given at the very beginning of each course and then a Pearson Correlation Coefficient can be calculated between the two sets of results. For the FASIs we have been involved with creating, we consistently see coefficients over 0.90 when they are administered in this way; but, there is no agreed upon accepted value.

It is quite common to see the statistic Cronbach's α or the Kuder–Richardson reliability index (KR-20) quoted as a measure of reliability (Cronbach, 1951; Kuder & Richardson, 1937). These indices would fall under (3) above, internal consistency coefficients. They are primarily useful for a single-construct test. Both indices depend on both the correlation between questions and the number of questions (Crocker & Algina, 1986; Field, 2009). However, in the words of Cronbach: 'Coefficients are a crude device that do not bring to the surface many subtleties implied by variance components. In particular, the interpretations being made in current assessments are best evaluated through use of a standard error of measurement' (Cronbach & Shavelson, 2004, p. 394). In fact, having a high correlation between items, which results in a higher value for α or KR-20, means that these items are repetitive. The way a FASI is typically administered puts a premium on minimizing the time required to complete the assessment and hence the number of questions. So a low Cronbach's α or KR-20 on a FASI would be quite reasonable, and a high Cronbach's α or KR-20 on a FASI does not guarantee that the test will be more reliable for its intended use and may be an indication that there are redundant questions that should be removed.

Item analysis. Item analysis is a term broadly used to describe any statistical property of examinees' responses to an individual test item. For FASIs we have found that *item difficulty, item discrimination*, and *point biserial correlation* provide useful information. Knowing these various statistics for each question helps describe how the questions on the test relate to each other and the test as a whole. It is also useful to provide this information when the test is published to inform test users about what sorts of values are likely to be observed for each question. However, as previously mentioned, these statistics are often used for single factor summative assessments of individual students, so the more common interpretation of their values and acceptable *cut off* ranges do not apply with a FASI.

Item difficulty is simply the percentage of students who got the item correct. It is valuable to have a range of difficulty levels. A wider range means the assessment can provide feedback on student learning for a range of student backgrounds and levels of mastery (majors and non-majors courses for example). If all items are very difficult for someone who has not had the course, it is possible that the FASI will only be valuable when given as a post-test (McKagan et al., 2010)—more detail is given in the Test Administration Tips section at the end of this paper.

The discrimination index is a measure of how well an item differentiates between strong and weak students, as defined by their overall test score. $D = p_u - p_l$, where p_u is the portion of students in the upper group who answered the item correctly and p_l is the portion of students in the lower group who answered the item correctly. There is no agreed upon percentage that should be used to determine these groups. We have seen between 27% and 50% used in the literature, but when working with large numbers of students, the results of the discrimination index will not fluctuate substantially based on your choice (Crocker & Algina, 1986). Again, the standard criteria often stated for the 'desirable' range for the discrimination index are not useful, as these are established based on single-construct tests designed to achieve maximum discrimination between students.

The discrimination index can be useful for identifying some important different types of questions. A question that nearly every student gets wrong before instruction but nearly all get correct after good instruction would have a very low discrimination index. However, such a question would be quite valuable on a FASI, as it would discriminate effective from ineffective teaching (formative assessment) by showing that most students were mastering the item in question when taught well. From a completely different perspective, namely teacher psychology, it is good to have at least a few questions on the test where most students do badly at the beginning of the course and well at the end for all reasonable forms of instruction. They show the teacher that students are learning something and the test is measuring student learning. If a teacher saw that there was no improvement on any of the questions, the natural reaction may be to conclude the students were hopelessly incompetent or the test was meaningless. Showing that students are learning some concepts but not others will be more likely to get the teacher to think about how they can make changes to improve their teaching. Another desirable type of question with very low discrimination index is one that probes understanding of a vital concept and the results show that no student learned it. The results of this type of question tend to have a large impact on teachers, particularly if the concept being measured is one that the teacher had erroneously believed students were learning based on results from typical course exams. Physics education researchers have found a number of such psychologically powerful examples (Mazur, 1997).

The point biserial coefficient (ρ_{pbis}) provides an additional measure of how consistent each question is with the whole assessment, and so the same general caveats to using standard ranges of desirable values apply as have already been discussed for Cronbach's α. This statistic is simply the Pearson Correlation Coefficient between a dichotomous variable (one that can only take on the values 0 or 1) and a continuous variable (test score). If there is a particular concept that students learn much better or worse than the other concepts, you want to know that, particularly if you see that those relationships vary considerably with different student populations and courses. For a question that revealed such variation, the standard point biserial coefficient would be much lower than is considered acceptable in traditional assessments. But for a FASI, that would be a fine question. It identifies certain concepts that are much more difficult for students to learn, and/or it identifies that this learning is quite dependent on the specifics of how the concept is taught in a way that is not true for the course material and other concepts as a whole. This provides very valuable feedback for improving how that topic is taught. So a low biserial coefficient can often be considered more desirable, because that shows that the question is probing a particular concept more specifically, rather than testing general knowledge.

Item Response Theory (IRT) is commonly used to create a response curve (probability of a student with a particular ability to answer the question correctly) for each item and/or to create a scaled score for the whole test based on what is known about each item. Practical uses of IRT are the construction of equivalent test forms or development of tests that discriminate at a particular level of ability (Crocker & Algina, 1986). Neither of which are goals for a FASI. In addition, the underlying assumption of IRT is that student scores on a test are based on a single latent trait—some general ability that may or may not be what the test was designed to measure. There are ways to handle multiple traits but these become much more complicated and are not as well understood. However, the idea of knowing more about which students answer an item correctly is of potential interest to developers of a FASI. It is also interesting to create an item response curve for each distracter to learn which students choose each distracter (Ding & Beichner, 2009).

An alternative to IRT is to use total scores as a substitute for the latent trait abilities to create an item response curve for each item (and its distracter if desired) on the FASI. This can be done by graphing the proportion of students who choose the correct response versus the total test score. This is a much easier method that will render useful results for a FASI (Ding & Beichner, 2009; Morris et al., 2006). Full on IRT requires specialized statistical software and large sample sizes from 200 (1 parameter Rausch—arguably not appropriate for a FASI) to 1,000 students (Crocker & Algina, 1986).

Test analysis. There are a few statistics that characterize the test as a whole, for example reliability measures. Another useful whole test statistic is Ferguson's delta (δ) or the *coefficient of test discrimination* (Ferguson, 1949). This statistic measures the discriminatory power of an entire test by investigating how broadly the total scores are distributed over the possible range. Ferguson's delta ranges from 0 to 1. A test with $\delta > 0.9$ is considered to be a test with high discriminatory power. Note that this statistic also depends on the population of students used to calculate δ so it can be interesting data when comparing different courses.

Correlations with other measures such as course outcomes. Validity evidence can be added to the FASI by calculating the Pearson Correlation Coefficient between the FASI results and other measures that are valued in the course such as exams, homework, course questionnaires, or other standard tests. These correlations can be used to establish that the FASI is measuring something that teachers care about. It can also inform the teachers about the incoming thinking of their students. We see this as adding to the evidence of validity in four different forms: (1) Evidence based on *relationships to other variables*, which includes correlations with course outcomes or other standards; (2) *Predictive and concurrent* evidence—does it predict and show what it claims; (3) *Convergent* evidence—the FASI results relate closely to course grades; and finally (4) *discriminate* evidence—a FASI that measures understanding of physics concepts should relate closer to a test of logical reasoning skills than it does to a test of the students' writing ability for example.

With many FASIs only some of the above correlations were studied before introduction of the FASI. A few examples can be found in Adams et al. (2006), Hestenes and Halloun (1995) and Smith et al. (2008). However, many of these sorts of correlations were done in later studies. All of them are by no means required before a new FASI is considered to have enough evidence of validity and reliability to be published. As this research is conducted for a variety of courses and universities, it strengthens the validity generalization of the FASI and each new set of data and correlations makes for useful research studies of their own.

Factor analysis. A factor analysis uses student responses to determine groups of questions that are answered in a correlated manner by students, thus indicating aspects of student thinking that are closely linked. The usefulness of such analysis depends on the design and purposes of the FASI. Some assessments have presented categories based on the author's expert view of the questions; but were criticized because these expert categories did not hold up under a factor analysis (Heller & Huffman, 1995; Hestenes & Halloun, 1995; Huffman & Heller, 1995). The factor analysis revealed that students did not answer questions within an expert category consistently, demonstrating that these categories do not represent single consistent ideas in students' minds. This is relevant when using the scores of individual categories to inform teaching about the particular concept the category may be identifying. With perception surveys, it is especially useful to perform a factor analysis because perceptions can be broad and novices may organize these ideas quite differently from experts.

Factor analysis is data-intensive, requiring a minimum of 10 times the number of responses as questions to carry it out in a statistically meaningful way. There are also multiple-factor analysis methods, a discussion of which is beyond the scope of this paper. See Ding and Beichner (2009), Heller and Huffman (1995), Hestenes and Halloun (1995), Huffman and Heller (1995), and Ramlo (2008) for discussion of factor analysis performed on concept inventories and Adams et al. (2006) for discussion of factor analysis for perception surveys.

Test Administration Tips

Because there are many subtle variations to how a FASI can be administered, when the test is published, it is important to include details of how the developer administered the test when validating it. This is both in keeping with AERA, APS, and NCME guidelines for educational and psychological testing and ensuring that results will be appropriately comparable. Here, we will describe the methods we have found to be effective and discuss how variations on these could affect the validity of a test result.

It is important to get as high a response rate as possible and have the students take the test seriously. However, there are obvious problems with having a test like this count directly toward a student's grade, because then the students will be motivated to learn about the test questions and memorize the answers without understanding them, potentially making the test worthless.

Here is the process that we have found to be effective at achieving the goals of a conceptual FASI. The exam is given on the first day of class, in class. We make it clear to the students that it will not count in their course grade, what is important is that they complete it, and that it will benefit them to do so, because it will improve instruction in the course by establishing their level of prior knowledge. We tell them that they are not expected to know the correct answers, what we are interested in is their current thinking about these topics, and they should put down their best guess but not agonize over the question if they do not know the answer. Nearly all students take the exam seriously when given under these circumstances. Students should be required to turn in both their answer sheets and the test itself to reduce the chances of test questions being circulated.

We have found we can get a serious effort from nearly all students in a course by giving the 'post' version of the test in class on the next to last day of class, with the guidance to the students that they will not be graded on their answers, but that these are all questions on important topics for the course and hence will provide guidance in studying for the final exam and that the FASI results will be reviewed and questions that students do particularly poorly on will be reviewed as part of the exam review. Again, students should be required to turn in both their answer sheet and the test itself.

In some courses, such as introductory quantum mechanics, students have no previous experience with the material, and so their pre-instruction scores are simply random guessing. Once a teacher has given the FASI and verified this lack of incoming knowledge on the subject, it is adequate in subsequent offerings of the course at that institution to only measure post-instruction results. However, it is never safe to assume knowledge of incoming student thinking about a subject without measuring it. Unfortunately, it is not adequate to assume preparation based on past courses and grades they have received.

On tests of expert-like perceptions, students always have opinions and so it is important to measure both pre and post. Pre-information is a way to learn about the population; however, to learn about the impact of instruction it is particularly important to only consider results that have matched pre- and post-tests for individual students, because selection effects are so important in such tests. It is reasonable to expect a strong correlation between a student's perceptions about the subject and whether or not they will be inclined to fill out a survey. Selection effects exist for tests of conceptual mastery as well, but they are not as tightly coupled to performance on the test.

Tests of perceptions can be given online with less risk of students consulting other sources to try and find a 'right' answer, and so for simplicity of data compilation and to avoid using class time, they are commonly administered online during the first and last weeks of a course. It is good to give the pre-test as early as possible, as our studies have demonstrated that students' perceptions are affected by the course very early in the semester. There are particular challenges to getting students to put serious thought into filling out online assessments, and we have developed some strategies to achieve this (Adams et al., 2006). A small number of bonus points are given to the students to provide incentive to fill out the test, reminder emails are sent out a couple of times during the week. Some instructors also find it effective to add the survey to the first homework assignment where students get the credit for simply completing the survey. Criteria such as dummy questions[1] and time used to fill out the test are used to filter out tests that were not answered with adequate care.

If the test is altered—removing question(s), adding question(s), changing the wording of question(s), etc.—results cannot then be compared with the published results, and in accordance with AERA, APS, and NCME guidelines this fact should be noted when the test is published. The reliability and validity were collected for the FASI as it was constructed only, and the value of comparisons with the results of modified tests is quite limited.

Some authors publish their instrument with the development and validation article while others only distribute the actual test when requests are received. Opinions are divided on this topic; however, the FCI and CLASS appear to still be quite effective tools and are publically available. While there was fear that students would get answers, invalidating the tests, there is no indication that this has happened, even for the FCI, the oldest and most widely used of such tests. Scores on the FCI have remained stable over decades in the absence of pedagogical changes. This is not surprising. Because the results of these tests typically do not count toward the students' grade, they have no incentive to take the time required to search out the answers and memorize them. If there is concern about compromising the questions, Maloney, O'Kuma, Hieggelke, and Van Heuvelen (2001) suggest not using the real name of the FASI when administering to students. They also name the file differently so that web searches using the official test name do not pull up the test.

Another common concern is that it is possible that since students have seen the FASI on the pre-course test it will inflate their post-course results. The empirical evidence is that such an effect is very small, as there have been many examples of pre–post improvements (gain) that are actually zero, most notably the extensive results obtained over many years with the FCI (Hake, 1998) by many instructors. In our experience, a surprisingly large fraction of the time, students will indicate they do not even remember taking the pre-test, let alone remember any questions from it. With many years of giving multiple FASI tests, we have yet to encounter a student who brought up one of the FASI pre-test questions at some point during the course. This lack of retention is not so surprising when one considers that these are questions students answer with no stakes involved, usually the student does not know how to answer most of them, they get no feedback on the answers, and they usually see each question for well under a minute and then not again until several months later.

Summary

In 2001, the National Research Council called for research on new forms of assessment that can be made practical to measure the effectiveness of instruction. They noted that the most serious barrier in traditional assessment design is the requirement of a team of experts (scientists, educators, task designers, and psychometricians) for creating a new test. They also recognize that further work is needed to integrate what is known from cognitive science into assessment design.

Many tests of instruction, both conceptual understanding and perceptions about a discipline and how it is best learned, have been created in the sciences. These tests were developed and validated using many of the same general procedures. However to our knowledge, these procedures have not been written down. Here, we described how a formative assessment of instruction can be created without a team of interdisciplinary experts by using student and expert interviews with methodology from cognitive science. The assessments that have been created using these procedures have been validated and published in peer-reviewed journals and widely adapted on an international scale. This provides evidence that this method is successful at creating a useful instrument for formative assessment of instruction.

Acknowledgements

We would like to thank the members of the Carl Wieman Science Education Initiative (CW-SEI) at the University of British Columbia and the Science Education Initiative (SEI) at the University of Colorado, especially Sarah Gilbert, Laurie Langdon, and Jennifer Duis for their helpful comments. This work was supported by the University of Colorado through the SEI and the University of British Columbia through the CW-SEI.

Note

1. We use this statement to discard the survey of people who are not reading the questions: Please select agree-option four (not strongly agree) for this question to preserve your answers.

References

Adams, W. K., Perkins, K. K., Podolefsky, N., Dubson, M., Finkelstein, N. D., & Wieman, C. E. (2006). A new instrument for measuring student beliefs about physics and learning physics: The Colorado learning attitudes about science survey. *Physical Review Special Topics—Physics Education Research, 2*(010101), 1–14.

AERA (American Educational Research Association), APA (American Psychological Association), & NCME (National Council on Measurement and Education). (1999). *Standards for educational and psychological testing.* Washington, DC: Author.

Barbera, J., Perkins, K. K., Adams, W. K., & Wieman, C. E. (2008). Modifying and validating the Colorado learning attitudes about science survey for use in chemistry. *Journal of Chemical Education, 85*, 1435–1439.

Berardi-Coletta, B., Buyer, L. S., Dominowski, R. L., & Rellinger, E. R. (1995). Metacognition and problem transfer: A process-oriented approach. *Journal of Experimental Psychology: Learning Memory and Cognition, 21*(1), 205–223.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn, brain, mind, experience, and school.* Washington, DC: National Academy Press.

Chasteen, S. V., & Pollock, S. J. (2010). Tapping into juniors' understanding of E&M: The Colorado Upper-Division Electrostatics (CUE) diagnostic. *2009 Physics Education Research Proceedings, 1179*, 7–10.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* Fort Worth, TX: Holt, Rinehart and Winston.

Cronbach, L. J. (1951). Coeffiecient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297–334.

Cronbach, L. J., Gleser, G. C., Nanada, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements.* New York: John Wiley.

Cronbach, L. J., & Shavelson, R. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement, 64*(3), 391–418.

Ding, L., & Beichner, R. (2009). Approaches to data analysis of multiple-choice questions. *Physical Review Special Topics—Physics Education Research, 5*(020103), 1–17.

Ding, L., Reay, N. W., Lee, A., & Bao, L. (2008). Effects of testing conditions on conceptual survey results. *Physical Review Special Topics—Physics Education Research, 4*(010112), 1–6.

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (2006). *The Cambridge handbook of expertise and expert performance.* New York: Cambridge University Press.

Ericsson, K. A., & Simon, H. A. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture, and Activity, 5*, 178–186.

Ferguson, G. A. (1949). On the theory of test development. *Psychometrika, 14*, 61–68.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). Los Angeles: Sage.

Goldhaber, S., Pollock, S., Dubson, M., Beale, P., & Perkins, K. (2010). Transforming upper-division quantum mechanics: Learning goals and their assessment. *AIP Conference Proceedings, 1179*, 145–148.

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics, 66*(1), 64–74.

Heller, P., & Huffman, D. (1995). Interpreting the force concept inventory: A reply to Hestenes and Halloun. *Physics Teacher, 33*(8), 503–511.

Hestenes, D., & Halloun, I. (1995). Interpreting the force concept inventory: A response to March 1995 critique by Huffman and Heller. *Physics Teacher, 33*(8), 502–506.

Hestenes, D., Wells, M., & Swackhamer, G. (1992). Force concept inventory. *Physics Teacher, 30*(3), 141–158.

Huffman, D., & Heller, P. (1995). What does the force concept inventory actually measure? *Physics Teacher, 33*(3), 138–143.

Jolley, A. R. (2010). *Identifying landscapes and their formation timescales: Comparing knowledge and confidence of beginner and advanced geoscience undergraduate students* (Undergraduate honours thesis). Department of Earth and Ocean Sciences, University of British Columbia, Canada.

Kachigan, S. K. (1986). *Statistical analysis.* New York: Radius Press.

Kuder, G., & Richardson, M. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160.

Maloney, D. P., O'Kuma, T. L., Hieggelke, C. J., & Van Heuvelen, A. (2001). Surveying students' conceptual knowledge of electricity and magnetism. *American Journal of Physics, Physics Education Research Supplement, 69*(S1), S12–S23.

Mazur, E. (1997). *Peer instruction: A users manual, series in educational innovation.* Upper Saddle River, NJ: Prentice Hall.

McKagan, S. B., Perkins, K. K., & Wieman, C. E. (2010). The design and validation of the quantum mechanics conceptual survey. Retrieved July 27, 2010, from http://arxiv.org/abs/1007.2015

Morris, G., Branum-Martin, L., Harshman, N., Baker, S., Mazur, E., Dutta, S., … McCauley, V. (2006). Testing the test: Item response curves and test quality. *American Journal of Physics, 74*(5), 449–453.

NRC (National Research Council). (2001). *Knowing what students know. The science and design of educational assessment.* In J. W. Pellegrino, N. Chudowsky, & R. Glaser (Eds.), *Committee on the foundations of Assessment* (Board on Testing and Assessment Center for Education Division of Behavioral and Social Sciences and Education) (pp. 1–14). Washington, DC: National Academy Press.

Ramlo, S. (2008). Validity and reliability of the force and motion conceptual evaluation. *American Journal of Physics, 76*(9), 882–886.

Redish, E. F., Steinberg, R. N., & Saul, J. M. (1998). Student expectations in introductory physics. *American Journal of Physics, 66*(3), 212–224.

Shavelson, R. J., & Ruiz-Primo, M. A. (2000). On the psychometrics of assessing science understanding. In J. Mintzes, J. H. Wandersee, & J. D. Novak (Eds.), *Assessing science understanding* (pp. 304–341). San Diego, CA: Academic Press.

Singh, C., & Mason, A. (2009, July). *Revisiting categorization.* Paper presented at the American Association of Physics Teachers 2009 summer meeting (BC05), Ann Arbor, MI.

Singh, C., & Rosengrant, D. (2003). Multiple-choice test of energy and momentum concepts. *American Journal of Physics, 71*(6), 607–617.

Smith, M. K., Wood, W. B., & Knight, J. K. (2008). The Genetics concept assessment: A new concept inventory for gauging student understanding of genetics. *CBE Life Sciences Education, 7*(4), 422–430.

Treagust, D. F. (1988). Development and use of diagnostic tests to evaluate students' misconceptions in science. *International Journal of Science Education, 10*(2), 159–169.

Vygotsky, L. S. (1962). *Thought and language.* Cambridge, MA: MIT Press.