CBE—Life Sciences Education
Vol. 15, 1–9, Spring 2016

Research Methods

Contemporary Test Validity in Theory and Practice: A Primer for Discipline-Based Education Researchers

Todd D. Reeves* and Gili Marbach-Ad†

*Educational Technology, Research and Assessment, Northern Illinois University, DeKalb, IL 60115; †College of Computer, Mathematical and Natural Sciences, University of Maryland, College Park, MD 20742

Submitted August 31, 2015; Revised October 31, 2015; Accepted November 2, 2015
Monitoring Editor: Ross Nehm

Most discipline-based education researchers (DBERs) were formally trained in the methods of scientific disciplines such as biology, chemistry, and physics, rather than social science disciplines such as psychology and education. As a result, DBERs may have never taken specific courses in the social science research methodology—either quantitative or qualitative—on which their scholarship often relies so heavily. One particular aspect of (quantitative) social science research that differs markedly from disciplines such as biology and chemistry is the instrumentation used to quantify phenomena. In response, this *Research Methods* essay offers a contemporary social science perspective on test validity and the validation process. The instructional piece explores the concepts of test validity, the validation process, validity evidence, and key threats to validity. The essay also includes an in-depth example of a validity argument and validation approach for a test of student argument analysis. In addition to DBERs, this essay should benefit practitioners (e.g., lab directors, faculty members) in the development, evaluation, and/or selection of instruments for their work assessing students or evaluating pedagogical innovations.

INTRODUCTION

The field of discipline-based education research (Singer *et al.*, 2012) has emerged in response to long-standing calls to advance the status of U.S. science education at the postsecondary level (e.g., Boyer Commission on Educating Undergraduates in the Research University, 1998; National Research Council, 2003; American Association for the Advancement of Science, 2011). Discipline-based education research applies scientific principles to study postsecondary science education processes and outcomes systematically to improve the scientific enterprise. In particular, this field has made significant progress with respect to the study of 1) active-learning pedagogies (e.g., Freeman *et al.*, 2014); 2) interventions to support those pedagogies among both faculty (e.g., Brownell and Tanner, 2012) and graduate teaching assistants (e.g., Schussler *et al.*, 2015); and 3) undergraduate research experiences (e.g., Auchincloss *et al.*, 2014).

Most discipline-based education researchers (DBERs) were formally trained in the methods of scientific disciplines such as biology, chemistry, and physics, rather than social science disciplines such as psychology and education. As a result, DBERs may have never taken specific courses in the social science research methodology—either quantitative or qualitative—on which their scholarship often relies so heavily (Singer *et al.*, 2012). While the same principles of science ground all these fields, the specific methods used and some criteria for methodological and scientific rigor differ along disciplinary lines.

One particular aspect of (quantitative) social science research that differs markedly from research in disciplines such as biology and chemistry is the instrumentation used to quantify phenomena. Instrumentation is a critical aspect of research methodology, because it provides the raw materials input to statistical analyses and thus serves as a basis for credible conclusions and research-based educational practice (Opfer *et al.*, 2012; Campbell and Nehm, 2013). A notable feature of social science instrumentation is that it generally targets variables that are latent, that is, variables

CBE Life Sci Educ March 1, 2016 15:rm1
DOI:10.1187/cbe.15-08-0183
Address correspondence to: Todd D. Reeves (treeves@niu.edu).

that are not directly observable but instead must be inferred through observable behavior (Bollen, 2002). For example, to elicit evidence of cognitive beliefs, which are not observable directly, respondents are asked to report their level of agreement (e.g., "strongly disagree," "disagree," "agree," "strongly agree") with textually presented statements (e.g., "I like science," "Science is fun," and "I look forward to science class"). Even a multiple-choice final examination does not directly observe the phenomenon of interest (e.g., student knowledge). As such, compared with work in traditional scientific disciplines, in the social sciences, more of an inferential leap is often required between the derivation of a score and its intended interpretation (Opfer *et al*., 2012).

Instruments designed to elicit evidence of variables of interest to DBERs have proliferated in recent years. Some well-known examples include the Experimental Design Ability Test (EDAT; Sirum and Humburg, 2011); the Genetics Concept Assessment (Smith *et al.*, 2008); the Classroom Undergraduate Research Experience survey (Denofrio *et al.*, 2007); and the Classroom Observation Protocol for Undergraduate STEM (Smith *et al.*, 2013). However, available instruments vary widely in their quality and nuance (Opfer *et al.*, 2012; Singer *et al.*, 2012; Campbell and Nehm, 2013), necessitating understanding on the part of DBERs of how to evaluate instruments for use in their research. Practitioners, too, should know how to evaluate and select high-quality instruments for program evaluation and/or assessment purposes. Where high-quality instruments do not already exist for use in one's context, which is commonplace (Opfer *et al*., 2012), they need to be developed, and corresponding empirical validity evidence needs to be gathered in accord with contemporary standards.

In response, this *Research Methods* essay offers a contemporary social science perspective on test validity and the validation process. It is intended to offer a primer for DBERs who may not have received formal training on the subject. Using examples from discipline-based education research, the instructional piece explores the concepts of test validity, the validation process, validity evidence, and key threats to validity. The essay also includes an in-depth example of a validity argument and validation approach for a test of student argument analysis. In addition to DBERs, this essay should benefit practitioners (e.g., lab directors, faculty members) in the development, evaluation, and/or selection of instruments for their work assessing students or evaluating pedagogical innovations.

TEST VALIDITY AND THE TEST VALIDATION PROCESS

A test is a sample of behavior gathered in order to draw an inference about some domain or construct within a particular population (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education [AERA, APA, and NCME], 2014).[1] In the social sciences, the domain about which an inference is desired is typically a latent (unobservable) variable. For example, the STEM GTA-Teaching Self-Efficacy Scale (DeChenne *et al.*, 2012) was developed to support inferences about the degree to which a graduate teaching assistant believes he or she is capable of 1) cultivating an effective learning environment and 2) implementing particular instructional strategies. As another example, the inference drawn from an introductory biology final exam is typically about the degree to which a student understands content covered over some extensive unit of instruction. While beliefs or conceptual knowledge are not directly accessible, what can be observed is the sample of behavior the test elicits, such as test-taker responses to questions or responses to rating scales. Diverse forms of instrumentation are used in discipline-based education research (Singer *et al.*, 2012). Notable subcategories of instruments include self-report (e.g., attitudinal and belief scales) and more objective measures (e.g., concept inventories, standardized observation protocols, and final exams). By the definition of "test" above, any of these instrument types can be conceived as tests—though the focus here is only on instruments that yield quantitative data, that is, scores.

The paramount consideration in the evaluation of any test's quality is validity: "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests" (Angoff, 1988; AERA, APA, and NCME, 2014, p. 11).[2,3] In evaluating test validity, the focus is not on the test itself, but rather the proposed *inference(s)* drawn on the basis of the test's score(s). Noteworthy in the validity definition above is that validity is a matter of degree ("the inferences supported by this test have a high or low degree of validity"), rather than a dichotomous character (e.g., "the inferences supported by this test are or are not valid").

Assessment validation is theorized as an iterative process in which the test developer constructs an evidence-based argument for the intended test-based score interpretations in a particular population (Kane, 1992; Messick, 1995). An example validity argument claim is that the test's content (e.g., questions, items) is representative of the domain targeted by the test (e.g., body of knowledge/skills). With this argument-based approach, claims within the validity argument are substantiated with various forms of relevant evidence. Altogether, the goal of test validation is to accumulate over time a comprehensive body of relevant evidence to support each intended score interpretation within a particular population (i.e., whether the scores should in fact be interpreted to mean what the developer intends them to mean).

CATEGORIES OF TEST VALIDITY EVIDENCE

Historically, test validity theory in the social sciences recognized several categorically different "types" of validity (e.g., "content validity," "criterion validity"). However, contemporary validity theory posits that test validity is a unitary (single) concept. Rather than providing evidence of each "type" of validity, the charge for test developers

[1]A test cannot be "stamped" valid for all purposes and test-taker populations; validity evidence needs to be gathered with respect to all intended instrument uses.

[2]While other key dimensions for evaluating an instrument's quality include reliability (i.e., test score consistency) and utility (i.e., feasibility; AERA, APA, and NCME, 2014), the focus here is on validity only.

[3]While this essay allies with test validity theory as codified in the *Standards for Educational and Psychological Testing* (AERA, APA, and NCME, 2014), the reader will note that there are alternative conceptions of validity as well (Lissitz and Samuelsen, 2007).

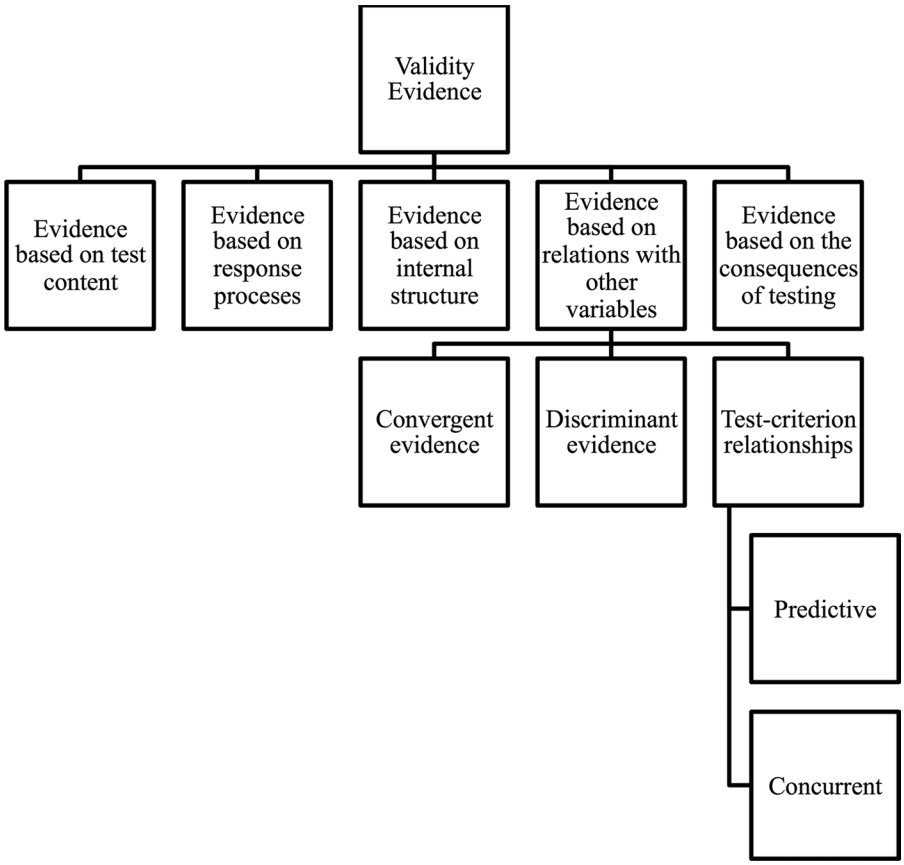

Figure 1. Categories of evidence used to argue for the validity of test score interpretations and uses (AERA, APA, and NCME, 2014).

is to construct a cohesive argument for the validity of test score–based inferences that integrates different forms of validity evidence. The categories of validity evidence include evidence based on test content, evidence based on response processes, evidence based on internal structure, evidence based on relations with other variables, and evidence based on the consequences of testing (AERA, APA, and NCME, 2014). Figure 1 provides a graphical representation of the categories and subcategories of validity evidence.

Validity evidence based on test content concerns "the relationship between the content of a test and the construct it is intended to measure" (AERA, APA, and NCME, 2014, p. 14). Such validity evidence concerns the match between the domain purportedly measured by (e.g., diagnostic microscopy skills) and the content of the test (e.g., the specific slides examined by the test taker). For example, if a test is intended to elicit evidence of students' understanding of the key principles of evolution by means of natural selection (e.g., variation, heredity, differential fitness), the test should fully represent those principles in the sample of behavior it elicits. As a concrete example from the literature, in the development of the Host-Pathogen Interaction (HPI) concept inventory, Marbach-Ad *et al.* (2009) explicitly mapped each test item to one of 13 HPI concepts intended to be assessed by their instrument. Content validity evidence alone is insufficient for establishing a high degree of validity; it should be combined with other forms of evidence to yield a strong evidence-based validity argument marked by relevancy, accuracy, and sufficiency.

In practice, providing validity evidence based on test content involves evaluating and documenting content representativeness. One standard approach to collecting evidence of content representativeness is to submit the test to external systematic review by subject matter–area experts (e.g., biology faculty) and to document such reviews (as well as revisions made on their basis). External reviews focus on the adequacy of the test's overall elicited sample of behavior in representing the domain assessed and any corresponding subdomains, as well as the relevance or irrelevance of particular questions/items to the domain. We refer the reader to Webb (2006) for a comprehensive and sophisticated framework for evaluating different dimensions of domain–test content alignment.

Another approach used to design a test, so as to support and document construct representativeness, is to employ a "table of specifications" (e.g., Fives and DiDonato-Barnes, 2013). A table of specifications (or test blueprint) is a tool for designing a test that classifies test content along two dimensions, a content dimension and a cognitive dimension. The content dimension pertains to the different aspects of the construct one intends to measure. In a classroom setting, aspects of the construct are typically defined by behavioral/instructional objectives (i.e., students will analyze phylogenetic trees). The cognitive dimension represents the level of cognitive processing or thinking called for by test components (e.g., knowledge, comprehension, analysis). Within a table of specifications, one indicates the number/percent of test questions or items for each aspect of the construct at each

Table 1. Example table of specifications for evolution by means of natural selection test showing numbers of test items pertaining to each content area at each cognitive level and total number of items per content area and cognitive level

Content (behavioral objective)	Cognitive process			Total
	Comprehension	Application	Analysis	
1. Students will define evolution by means of natural selection.	1			1
2. Students will define key principles of evolution by means of natural selection (e.g., heredity, differential fitness).	5			5
3. Students will compute measures of absolute and relative fitness.		5		5
4. Students will compare evolution by means of natural selection with earlier evolution theories.			3	3
5. Student will analyze phylogenetic trees.			4	4
Total	6	5	7	18

cognitive level. Often, one also provides a summary measure of the number of items pertaining to each content area (regardless of cognitive demand) and at each cognitive level (regardless of content). Instead of or in addition to the number of items, one can also indicate the number/percent of available points for each content area and cognitive level. Because a table of specifications indicates how test components represent the construct one intends to measure, it serves as one source of validity evidence based on test content. Table 1 presents an example table of specifications for a test concerning the principles of evolution by means of natural selection.

Evidence of validity based on response processes concerns "the fit between the construct and the detailed nature of the performance or response actually engaged in by test takers" (AERA, APA, and NCME, 2014, p. 15). For example, if a test purportedly elicits evidence of undergraduate students' critical evaluative thinking concerning evidence-based scientific arguments, during the test the student should be engaged in the cognitive process of examining argument claims, evidence, and warrants, and the relevance, accuracy, and sufficiency of evidence. Most often one gathers such evidence through respondent think-aloud procedures. During think alouds, respondents verbally explain and rationalize their thought processes and responses concurrently during test completion. One particular method commonly used by professional test vendors to gather response process–based validity evidence is cognitive labs, which involve both concurrent and retrospective verbal reporting by respondents (Willis, 1999; Zucker *et al.*, 2004). As an example from the literature, developers of the HPI concept inventory asked respondents to provide open-ended responses to ensure that their reasons for selecting a particular response option (e.g., "B") were consistent with the developer's intentions, that is, the student indeed held the particular alternative conception presented in response option B (Marbach-Ad *et al.*, 2009). Think alouds are formalized via structured protocols, and the elicited think-aloud data are recorded, transcribed, analyzed, and interpreted to shed light on validity.

Evidence based on internal structure concerns "the degree to which the relationships among test item and test components conform to the construct on which the proposed test score interpretations are based" (AERA, APA, and NCME, 2014, p. 16).[4] For instance, suppose a professor plans to teach one topic (eukaryotes) using small-group active-learning instruction and another topic (prokaryotes) through lecture instruction; and he or she wants to make within-class comparisons of the effectiveness of these methods. As an outcome measure, a test may be designed to support inferences about the two specific aspects of biology content (e.g., characteristics of prokaryotic and eukaryotic cells). Collection of evidence based on internal structure seeks to confirm empirically whether the scores reflect the (in this case two) distinct domains targeted by the test (Messick, 1995). In practice, one can formally establish the fidelity of test scores to their theorized internal structure through methodological techniques such as factor analysis, item response theory, and Rasch modeling (Harman, 1960; Rasch, 1960; Embretson and Reise, 2013). With factor analysis, for example, item intercorrelations are analyzed to determine whether particular item responses cluster together (i.e., whether scores from components of the test related to one aspect of the domain [e.g., questions about prokaryotes] are more interrelated with one another than they are with scores derived from other components of the test [e.g., questions about eukaryotes]).

Item response theory and Rasch models hypothesize that the probability of a particular response to a test item is a function of the respondent's ability (in terms of what is being measured) and characteristics of the item (e.g., difficulty, discrimination, pseudo-guessing). Examining test score internal structure with such models involves examining whether such model-based predictions bear out in the observed data. There are a variety of such models for use with test questions with different (or different combinations of) response formats such as the Rasch rating-scale model (Andrich, 1978) and the Rasch partial-credit Rasch model (Wright and Masters, 1982).

Validity evidence based on relations with other variables concerns "the relationship of test scores to variables external to the test" (AERA, APA, and NCME, 2014, p. 16). The collection of this form of validity evidence centers on examining how test scores are related to both measures of the same or similar constructs and measures of distinct and different constructs (i.e., respectively termed "convergent validity" and "discriminant validity"[5] evidence). In other words, such evidence pertains to how scores relate to other variables as would be theoretically expected. For example,

[4]This source of evidence has been termed "substantive validity" (Messick, 1995).

if a new self-report instrument purports to measure experimental design skills, scores should correlate highly with an existing measure of experimental design ability such as the EDAT (Sirum and Humburg, 2011). On the other hand, scores derived from this self-report instrument should be considerably less correlated or uncorrelated with scores from a personality measure such as the Minnesota Multiphasic Personality Inventory (Greene, 2000). As another discipline-based education research example, Nehm and Schonfeld (2008) collected discriminant validity evidence by relating scores from both the Conceptual Inventory of Natural Section (CINS) and the Open Response Instrument (ORI), which both purport to assess understanding of and conceptions concerning natural selection, and a geology test of knowledge about rocks.

A subcategory of evidence based on relations with other variables is evidence related to test-criterion relationships, which concerns how test scores are related to some other nontest indicator or outcome either at the same time (so-called concurrent validity evidence) or in the future (so-called predictive validity evidence). For instance, developers of a new biostatistics test might examine how scores from the test correlate as expected with professor ability judgments or mathematics course grade point average at the same point in time; alternatively, the developer might follow tested individuals over time to examine how scores relate to the probability of successfully completing biostatistics course work. As another example, given prior research on self-efficacy, scores from instruments that probe teaching self-efficacy should be related to respondents' levels of teacher training and experience (Prieto and Altmaier, 1994; Prieto and Meyers, 1999).

Examination of how test scores are related or not to other variables as expected is often associational in nature (e.g., correlational analysis). There are also two other specific methods for eliciting such validity evidence. The first is to examine score differences between theoretically different groups (e.g., whether scientists' and nonscientists' scores from an experimental design test differ systematically on average)—the "known groups method." The second is to examine whether scores increase or decrease as expected in response to an intervention (Hattie and Cooksey, 1984; AERA, APA, and NCME, 2014). For example, Marbach-Ad *et al.* (2009, 2010) examined HPI concept inventory score differences between majors and nonmajors and students in introductory and upper-level courses. To inform the collection of validity evidence based on relations with other variables, individuals should consult the literature to formulate a theory around how good measures of the construct should relate to different variables. One should also note that the quality of such validity evidence hinges on the quality (e.g., validity) of measures of external variables.

Finally, validity evidence based on the consequences of testing concerns the "soundness of proposed interpretations [of test scores] for their intended uses" (AERA, APA, and NCME, 2014, p. 19) and the value implications and social consequences of testing (Messick, 1994, 1995). Such evidence pertains to both the intended and unintended effects of test score interpretation and use (Linn, 1991; Messick, 1995). Example intended consequences of test use would include motivating students, better-targeted instruction, and populating a special program with only those students who are in need of the program (if those are the intended purposes of test use). An example of an unintended consequence of test use would be significant reduction in instructional time because of overly time-consuming test administration (assuming, of course, that this would not be a desired outcome) or drop out of particular student populations because of an excessively difficult test administered early in a course. In K–12 settings, a classic example of an unintended consequence of testing is the "narrowing of the curriculum" that occurred as a result of the No Child Left Behind Act testing regime; when faced with annual tests focused only on particular content areas (i.e., English/language arts and mathematics), schools and teachers focused more on tested content and less on nontested content such as science, social studies, art, and music (e.g., Berliner, 2011). Evidence based on the consequences of a test is often gathered via surveys, interviews, and focus groups administered with test users.

[5]This is not to be confused with item discrimination, a test item property pertaining to how an item's scores relate to overall test performance.

TEST VALIDITY ARGUMENT EXAMPLE

In this section, we provide an example validity argument for a test designed to elicit evidence of students' skills in analyzing the elements of evidence-based scientific arguments. This hypothetical test presents text-based arguments concerning scientific topics (e.g., global climate change, natural selection) to students, who then directly interact with the texts to identify their elements (i.e., claims, reasons, and warrants). The test is intended to support inferences about 1) students' overall evidence-based science argument-element analysis skills; 2) students' skills in identifying *particular* evidence-based science argument elements (e.g., claims); and 3) errors made when students identify *particular* argument elements (e.g., evidence). Additionally, the test is intended to 4) support instructional decision-making to improve science teaching and learning. The validity argument claims undergirding this example test's score interpretations and uses (and the categories of validity evidence advanced to substantiate each) are shown in Table 2.

ANALYSIS OF CINS VALIDITY EVIDENCE

The example validity argument provided in the preceding section was intended to model the validity argument formulation process for readers who intend to *develop* an instrument. However, in many cases, an existing instrument (or one of several existing instruments) needs to be *selected* for use in one's context. The use of an existing instrument for research or practice requires thoughtful analysis of extant validity evidence available for an instrument's score interpretations and uses. Therefore, in this section, we use validity theory as outlined in the *Standards for Educational and Psychological Testing* to analyze the validity evidence for a particular instrument, the CINS.

As reported in Anderson *et al*. (2002), the CINS is purported to measure "conceptual understanding of natural selection" (as well as alternative conceptions of particular relevant ideas diagnostically) in undergraduate non–biology

Table 2. Example validity argument and validation approach for a test of students' ability to analyze the elements of evidence-based scientific arguments showing argument claims and subclaims concerning the validity of the intended test score interpretations and uses and relevant validity evidence used to substantiate those claims

Validity argument claims and subclaims	Relevant validity evidence based on
1. The overall score represents a student's current level of argument-element analysis skills, because:	–
a single higher-order construct (i.e., argument-element analysis skill) underlies all item responses.	Internal structure
the overall score is distinct from content knowledge and thinking dispositions.	Relations with other variables
the items represent a variety of arguments and argument elements.	Test content
items engage respondents in the cognitive process of argument-element analysis.	Response processes
the overall score is highly related to other argument analysis measures and less related to content knowledge and thinking disposition measures.	Relations with other variables
2. A subscore (e.g., claim identification) represents a student's current level of argument-element identification skill, because:	–
each subscore is distinct from other subscores and the total score (the internal structure is multidimensional and hierarchical).	Internal structure
the items represent a variety of arguments and particular argument elements (e.g., claims).	Test content
the subscore is distinct from content knowledge and thinking dispositions.	Relations with other variables
items engage respondents in the cognitive process of identifying a particular element argument (e.g., claims).	Response processes
subscores are highly related to other argument analysis measures and less related to content knowledge and thinking disposition measures.	Relations with other variables
3. Error indicators can be interpreted to represent students' current errors made in identifying particular argument elements, because when students misclassify an element in the task, they are making cognitive errors.	Response processes
4. Use of the test will facilitate improved argument instruction and student learning, because:	–
teachers report that the test is useful and easy to use and have positive attitudes toward it.	Consequences of testing
teachers report using the test to improve argument instruction.	Consequences of testing
teachers report that the provided information is timely.	Consequences of testing
teachers learn about argumentation with test use.	Consequences of testing
students learn about argumentation with test use	Consequences of testing
any unintended consequences of test use do not outweigh intended consequences.	Consequences of testing

majors before instruction (p. 953). In their initial publication of the instrument, the authors supplied several forms of validity evidence for the intended score interpretations and uses. In terms of validity evidence related to test content, the authors argued that test content was aligned with Mayr's (1982) five facts and three inferences about evolution by means of natural selection, and two other key concepts, the origin of variation and the origin of species. Two test items were mapped to each of these 10 concepts. Similarly, distractor (incorrect) multiple-choice responses were based on theory and research about students' nonscientific, or alternative, conceptions of these ideas. Content-related validity evidence was also provided through reviews of test items by biology professors.

Evidence based on test-taker response processes was elicited through cognitive interviews (think alouds) conducted with a small sample of students (Anderson *et al.*, 2002). The authors provided validity evidence based on internal structure using principal components analysis, which is similar to factor analysis. In terms of validity evidence based on test-score relations with other variables, the authors examined correlations between CINS scores and scores derived from interviews. While Anderson and colleagues did note that a paper and pencil–based test would be more logistically feasible than interview-based assessment methods, validity evidence based on the consequences of testing was not formally provided.

Anderson *et al.*'s (2002) paper did present a variety of forms of validity evidence concerning the CINS instrument. However, and underscoring the continuous nature of test validation, subsequent work has built upon their work and provided additional evidence. For example, in light of concerns that the primary earlier source of validity evidence was correlations between CINS scores and scores based on oral interviews in a very small sample, Nehm and Schonfeld (2008) provided additional validity evidence based on relations with other variables. For example, Nehm and Schonfeld (2008) examined CINS score relations with two other instruments purported to assess the same construct (convergent validity evidence) and with a measure of an unrelated construct (discriminant validity evidence). Nehm and Schonfeld also expanded the body of CINS validity evidence based on internal structure by analyzing data using the Rasch model. The authors' reporting of CINS administration times similarly shed light on possible consequences of testing. The evolution of validity evidence for the CINS noted here certainly speaks to the iterative and ongoing nature of instrument validation processes. With this in mind, future work might examine CINS scores' internal structure vis-à-vis diagnostic classification models (see Rupp and Templin, 2008), since CINS is purportedly a diagnostic test.

TEST VALIDITY THREATS

The two primary threats to test score validity are construct underrepresentation and construct-irrelevant variance. Construct underrepresentation is "the degree to which a test fails to capture important aspects of the construct"

(AERA, APA, and NCME, 2014; p. 12). In other words, construct underrepresentation involves failing to elicit a representative sample of behavior from test takers (e.g., responses to multiple-choice questions) relative to the universe of possible relevant behaviors that might be observed. While it is neither necessary nor feasible to ask respondents to engage in every single possible relevant behavior, it is crucial that the behavior sampled by the test is sufficiently representative of the construct at large. If a test does not fully and adequately sample behavior related to the targeted domain, the test score's meaning in actuality would be narrower than is intended.

Content underrepresentation can be mitigated by initiating test design with a thorough analysis and conception of the domain targeted by the test (Mislevy *et al.*, 2003; Opfer *et al.*, 2012). Knowledge of the construct, and variables that are related or not related to the construct, can also inform the validation process (Mislevy *et al.*, 2003). Beginning test design with a thorough conception of the construct one intends to measure is analogous to the course design approach known as "backward design"; with backward design one first identifies what one wants students to know/be able to do *after* instruction (learning outcomes) and then designs a course to get students there (Wiggins and McTighe, 2005). Other strategies to promote construct representation include building a test based on a table of specifications; submitting a text to external expert content review (as both noted above); and employing a sufficient number of test items to ensure good representation of domain content.

Besides construct representation, the other primary threat to test score validity is construct-irrelevant variance—"the degree to which test scores are affected by processes that are extraneous to the test's intended purpose" (AERA, APA, and NCME, 2014, p. 12). Construct-irrelevant variance is test score variation caused *systematically* by factors other than (or in addition to) those intended; in other words, some part of the reason why one received a "high" or "low" score is due to irrelevant reasons. Two common examples of this are: English skills affecting test scores for non–native English speakers on tests written in English; and computer skills affecting test scores for tests administered via computer. Another example would be if items on a science teaching self-efficacy self-report instrument are written so generally that the scores represent not science teaching–specific self-efficacy but self-efficacy in general. It is critical to mitigate such threats through test design processes (e.g., minimizing test linguistic load). One can often identify potential threats in the course of a thorough analysis of the construct/domain done at early design stages. During test validation one should also disconfirm such threats wherein scores are driven by irrelevant factors; practitioners often conduct factor, correlational, and differential item functioning analyses toward this end.

SUMMARY

Systematic research on postsecondary science teaching and learning and evaluation of local innovations by practitioners hinges on the availability and use of sound instrumentation. Unfortunately, the field of discipline-based education research lacks sufficient existing and high-quality instruments for use in all of these efforts (Opfer *et al.*, 2012; Singer *et al.*, 2012; Campbell and Nehm, 2013). DBERs and practitioners furthermore do not typically have formal training that equips them to evaluate and select existing instruments or develop and validate their own instruments for needed purposes. This essay reviewed contemporary test validity and validation theory for DBERs and practitioners in hopes of equipping them with such knowledge.[6]

This essay was chiefly intended for two audiences: 1) those who will develop new instruments; and 2) those who will evaluate and select from among existing instruments. Here, we summarily denote the implications of this essay for members of these two populations. First, it behooves those developing and publishing their own instruments to explicitly frame, construct, and report an evidence-based validity argument for their proposed instruments' intended score interpretations and uses. This argument should rely on multiple forms of validity evidence and specify the test-taker and user populations for which that argument pertains. If faced with space constraints in journal articles, test manuals or technical reports can be written to detail such validity evidence and made available to the scholarly community.

Like any argument, an evidence-based argument formulated during test validation should be characterized by relevancy, accuracy, and sufficiency. As such, validity arguments should be held up to scientific scrutiny *before* a test's operational use. The quality of a validity argument hinges on a number of factors discussed in this essay. Examples include the alignment of the validity argument claims with intended test score interpretations and uses; the representativeness of the samples from which validity evidence is gathered to the intended test-taker population; the relevance of the expertise held by content reviewers; and the technical quality of external measures. A final point to emphasize is that validation is an ongoing process; additional validity evidence may need to be gathered as theory concerning a construct evolves or as the community seeks to use an instrument with new populations.

Second, potential test users should be critical in their evaluation of existing instruments, and should not merely assume a strong validity argument exists for an instrument's score interpretations and uses with a particular population. Potential users should look to the instrumentation (or methods) sections of published articles for key information, such as whether the test was developed based on a sound theoretical conception of construct, whether the test underwent external content review, and whether scores correlate with other measures as they theoretically should, among other things. One may have to contact an author for such information. Altogether, such practices should advance the quality of measurement within the realm of discipline-based education research.

[6]While our focus is on instruments comprising sets of questions or items intended to elicit evidence of a particular construct or constructs, many of the ideas here apply also to questionnaire (survey) validation. For example, the developer of a questionnaire may interrogate how respondents interpret and formulate a response to a particular question as validity evidence based on response processes.

ACKNOWLEDGMENTS

The authors thank Drs. Beth Schussler and Ross Nehm and two anonymous reviewers for their constructive feedback on an earlier version of this article.

REFERENCES

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

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education (2014). Standards for Educational and Psychological Testing, Washington, DC.

Anderson DL, Fisher KM, Norman GJ (2002). Development and evaluation of the conceptual inventory of natural science. J Res Sci Teach *39*, 952–978.

Andrich DA (1978). A rating formulation for ordered response categories. Psychometrika *43*, 561–573.

Angoff WH (1988). Validity: an evolving concept. In: Test Validity, ed. H Wainer and H Braun, Hillsdale, NJ: Erlbaum, 19–32.

Auchincloss LC, Laursen SL, Branchaw JL, Eagan K, Graham M, Hanauer DI, Lawrie G, McLinn CM, Palaez N, Rowland S, *et al.* (2014). Assessment of course-based undergraduate research experiences: a meeting report. CBE Life Sci Educ *13*, 29–40.

Berliner D (2011). Rational responses to high stakes testing: the case of curriculum narrowing and the harm that follows. Cambridge J Educ *41*, 287–302.

Bollen KA (2002). Latent variables in psychology and the social sciences. Annu Rev Psychol *53*, 605–634.

Boyer Commission on Educating Undergraduates in the Research University (1998). Reinventing Undergraduate Education: A Blueprint for America's Research Universities, Stony Brook: State University of New York.

Brownell SE, Tanner KD (2012). Barriers to faculty pedagogical change: lack of training, time, incentives, and … tensions with professional identity. CBE Life Sci Educ *11*, 339–346.

Campbell CE, Nehm RH (2013). A critical analysis of assessment quality in genomics and bioinformatics education research. CBE Life Sci Educ *12*, 530–541.

DeChenne SE, Enochs LG, Needham M (2012). Science, technology, engineering, and mathematics graduate teaching assistants teaching self-efficacy. J Scholarship Teach Learn *12*, 102–123.

Denofrio LA, Russell B, Lopatto D, Lu Y (2007). Linking student interests to science curricula. Science *318*, 1872–1873.

Embretson SE, Reise SP (2013). Item Response Theory, Mahwah, NJ: Lawrence Erlbaum Associates.

Fives H, DiDonato-Barnes N (2013). Classroom test construction: the power of a table of specifications. Pract Assess Res Eval *18*, 2–7.

Freeman S, Eddy SL, McDonough M, Smith MK, Wenderoth MP, Okoroafor N, Jordt H (2014). Active learning increases student performance in science, engineering, and mathematics. Proc Natl Acad Sci USA *111*, 8410–8415.

Greene RL (2000). The MMPI-2: An Interpretive Manual, Boston, MA: Allyn & Bacon.

Harman HH (1960). Modern Factor Analysis, Oxford, UK: University of Chicago Press.

Hattie J, Cooksey RW (1984). Procedures for assessing the validities of tests using the "known-groups" method. Appl Psychol Meas *8*, 295–305.

Kane MT (1992). An argument-based approach to validity. Psychol Bull *112*, 527–535.

Linn RL (1991). Complex, performance-based assessment: expectations and validation criteria. Educ Researcher *20*, 15–21.

Lissitz RW, Samuelsen K (2007). A suggested change in terminology and emphasis regarding validity and education. Educ Researcher *36*, 437–448.

Marbach-Ad G, Briken V, El-Sayed NM, Frauwirth K, Fredericksen B, Hutcheson S, Gao LY, Joseph S, Lee VT, McIver KS, *et al.* (2009). Assessing student understanding of host pathogen interactions using a concept inventory. J Microbiol Biol Educ *10*, 43–50.

Marbach-Ad G, McAdams K, Benson S, Briken V, Cathcart L, Chase M, El-Sayed N, Frauwirth K, Fredericksen B, Joseph S, *et al.* (2010). A model for using a concept inventory as a tool for students' assessment and faculty professional development. CBE Life Sci Educ *9*, 408–436.

Mayr E (1982). The Growth of Biological Thought: Diversity, Evolution and Inheritance, Cambridge, MA: Harvard University Press.

Messick S (1994). The interplay of evidence and consequences in the validation of performance assessments. Educ Researcher *23*, 13–23.

Messick S (1995). Standards of validity and the validity of standards in performance assessment. Educ Meas *14*, 5–8.

Mislevy RJ, Steinberg LS, Almond RG (2003). On the structure of educational assessments. Measurement *1*, 3–62.

National Research Council (2003). BIO2010: Transforming Undergraduate Education for Future Research Biologists, Washington, DC: National Academies Press.

Nehm RH, Schonfeld IS (2008). Measuring knowledge of natural selection: a comparison of the CINS, an open-response instrument, and an oral interview. J Res Sci Teach *45*, 1131–1160.

Opfer JE, Nehm RH, Ha M (2012). Cognitive foundations for science assessment design: knowing what students know about evolution. J Res Sci Teach *49*, 744–777.

Prieto LR, Altmaier EM (1994). The relationship of prior training and previous teaching experience to self-efficacy among graduate teaching assistants. Res High Educ *35*, 481–497.

Prieto LR, Meyers SA (1999). Effects of training and supervision on the self-efficacy of psychology graduate teaching assistants. Teach Psychol *26*, 264–266.

Rasch G (1960). Probabalistic Models for Some Intelligence and Achievement Tests, Copenhagen: Danish Institute for Educational Research.

Rupp AA, Templin JL (2008). Unique characteristics of diagnostic classification models: a comprehensive review of the current state-of-the-art. Measurement *6*, 219–262.

Schussler EE, Reed Q, Marbach-Ad G, Miller K, Ferzli M (2015). Preparing biology graduate teaching assistants for their roles as instructors: an assessment of institutional approaches. CBE Life Sci Educ *14*, ar31.

Singer SR, Nielsen NR, Schweingruber HA (2012). Discipline-based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering, Washington, DC: National Academies Press.

Sirum K, Humburg J (2011). The Experimental Design Ability Test (EDAT). Bioscene *37*, 8–16.

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

Smith MK, Wood WB, Knight JK (2008). The Genetics Concept Assessment: a new concept inventory for gauging student understanding of genetics. CBE Life Sci Educ 7, 422–430.

Webb N (2006). Identifying content for student achievement tests. In: Handbook of Test Development, ed. SM Downing and TM Haladyna, Mahwah, NJ: Erlbaum, 155–180.

Wiggins GP, McTighe J (2005). Understanding by Design, Alexandria, VA: Association for Supervision and Curriculum Development.

Willis GB (1999). Cognitive Interviewing: A "How To" Guide, Research Triangle Park, NC: Research Triangle Institute.

Wright BD, Masters GN (1982). Rating Scale Analysis, Chicago, IL: MESA.

Zucker S, Sassman S, Case BJ (2004). Cognitive Labs. http://images.pearsonassessments.com/images/tmrs/tmrs_rg/CognitiveLabs.pdf (accessed 29 August 2015).