CBE—Life Sciences Education
Vol. 12, 530–541, Fall 2013

Article

A Critical Analysis of Assessment Quality in Genomics and Bioinformatics Education Research

Chad E. Campbell* and Ross H. Nehm†

*School of Teaching and Learning, Ohio State University, Columbus, OH 43210; †Center for Science and Mathematics Education and Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794

Submitted June 13, 2012; Revised January 21, 2013; Accepted February 4, 2013
Monitoring Editor: Clarissa Ann Dirks

The growing importance of genomics and bioinformatics methods and paradigms in biology has been accompanied by an explosion of new curricula and pedagogies. An important question to ask about these educational innovations is whether they are having a meaningful impact on students' knowledge, attitudes, or skills. Although assessments are necessary tools for answering this question, their outputs are dependent on their quality. Our study 1) reviews the central importance of reliability and construct validity evidence in the development and evaluation of science assessments and 2) examines the extent to which published assessments in genomics and bioinformatics education (GBE) have been developed using such evidence. We identified 95 GBE articles (out of 226) that contained claims of knowledge increases, affective changes, or skill acquisition. We found that 1) the purpose of most of these studies was to assess summative learning gains associated with curricular change at the undergraduate level, and 2) a minority (<10%) of studies provided any reliability or validity evidence, and only one study out of the 95 sampled mentioned *both* validity and reliability. Our findings raise concerns about the quality of evidence derived from these instruments. We end with recommendations for improving assessment quality in GBE.

INTRODUCTION

Complex scientific challenges, including escalating global climate change, overpopulation, environmental degradation, and the emergence of new pathogens, demand new types of scientific responses(American Association for the Advancement of Science [AAAS], 2009). In *A New Biology for the 21st Century* (National Research Council [NRC], 2009), several recommendations were proposed to meet these global challenges, including 1) placing a priority on the development of information technologies for use in scientific discovery and 2) devoting resources to the creation of interdisciplinary curricula and professional training opportunities in order to foster cooperation among scientists, engineers, and computer scientists (NRC, 2009). Furthermore, *New Biology* (NRC, 2009) emphasizes that substantial changes are required of biology curricula, and that new methods, tools, and conceptual paradigms are urgently needed for the teaching and learning of such approaches.

The fields of bioinformatics and genomics are emblematic of the "new biology": they weave together aspects of computer science, information technology, and large-scale life sciences research in innovative and integrative ways. In an effort to meet the need for new interdisciplinary curricula and professional training opportunities in genomics and bioinformatics, educational reforms have taken place at the secondary, undergraduate, and graduate levels (Wefer and Anderson, 2008; Haury and Nehm, 2012). The task of educational reform in genomics and bioinformatics education (GBE) is particularly challenging, given that new technologies, empirical discoveries, and new research areas are emerging continuously (e.g., pharmacogenomics, proteomics,

DOI: 10.1187/cbe.12-06-0073
Address correspondence to: Chad E. Campbell (campbell.742@osu.edu).
Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the view of the National Science Foundation.

metabolomics, and biopathway modeling; Curioso *et al.*, 2008).

Since 1995, there has been a substantial increase in publications attempting to evaluate the educational impact of GBE labs, modules, workshops, and curricula, with more than 200 studies published thus far (Campbell and Nehm, 2011). To our knowledge, no work has attempted to take stock of this effort or to evaluate the quality of the tools that the academic community has built to measure what students are learning in these new modules, classes, and degree programs. Given that a central goal of discipline-based educational research (DBER) is to establish evidence-based teaching practices to improve learning outcomes (NRC, 2012), the quality of the evidence used in such efforts is paramount. Indeed, using results from assessment tools that have not been shown to generate valid and reliable inferences 1) is at odds with the principles of scientific research in education (NRC, 2002); 2) goes against the educational assessment standards (American Educational Research Association [AERA] *et al.*, 1999; Brennan, 2006); and 3) risks the use of faulty or misleading information to guide educational evaluation and reform. Therefore, an important question to ask is whether the assessment tools and instruments being used in the GBE literature meet the quality control benchmarks for measurement established by the educational research community (i.e., AERA *et al.*, 1999) and whether educational reform in this area is proceeding with robust, evidence-based claims. Simply put, "Anybody can develop and distribute a test, but whether the scores on that test are meaningful and useful is the question to answer" (Cizek, 2007, p. 20).

In line with the NRC (2012) recommendations, our study explores the quality of the evidence produced thus far in GBE through an evaluation of the educational assessments used to generate evidence about instructional efficacy. Our study specifically reviews the central importance of reliability and construct validity evidence in the development and evaluation of science assessments and the implications of our findings for evidence-based practice in GBE.

Before reviewing the quality of educational assessment tools produced in GBE, we first introduce readers to our framework for educational assessment. Regardless of whether the reader is a practitioner only interested in using assessments that have been developed by others, or a biology educator interested in developing new assessments, issues of assessment quality must be understood in order to choose or develop tools that are capable of generating robust and valuable information about student learning (be it cognitive, affective, or psychomotor). Indeed, evidence-based educational practice does not require just any form of evidence; rather, it requires robust evidence derived from particular sources using particular approaches. Assessments that generate faulty or misleading evidence can mislead the most well-intentioned educational reform efforts.

FRAMEWORK FOR INSTRUMENT DEVELOPMENT AND EVALUATION

Our framework for assessing the measurement quality of assessments is aligned with the concepts of construct validity and reliability established in *Educational Measurement* (Brennan, 2006) and the most recent *Standards for Educational and Psychological Testing* (AERA *et al.*, 1999) and by Messick (1989, 1995). It is important to emphasize that, in terms of educational testing, "reliability" and "validity" do *not* refer to the properties of a test, but rather the inferences derived from the scores that the test produces. A simple example may help to clarify this confusing but important distinction. In journal articles (and conferences), scholars often write (and say): "this test is valid and reliable." But such language is often used as shorthand for the statement "the scores generated from the instrument produce valid and reliable inferences under particular circumstances." Tests themselves do not carry the properties of reliability or validity wherever and whenever they are used; rather, the contexts of test use, coupled with inferences about how the test functions to produce scores, are central aspects of validity and reliability.

For example, many science educators have been known to say that the Force Concept Inventory (Hestenes *et al.*, 1992) is "a valid and reliable instrument." If, for example, this FCI is administered to a group of elementary school students, the scores derived from it are unlikely to be reliable (or valid), because some conditions and contexts invalidate many inferences (e.g., the test takers could comprehend the items, etc.). Thus, it is not appropriate to refer to a *test* as reliable (or valid). This misconception may explain why researchers have been known to defend their use of a test in a new context or population using the argument that "the instrument has been shown to be valid and reliable." But again, a test that has been shown to generate reliable inferences about undergraduate science majors, for example, may not generate reliable inferences for nonscience majors. Furthermore, establishing validity and reliability evidence "is an ongoing process that requires gathering and synthesizing evidence. Evidence should continually be gathered to support or refute what is being claimed about the meaning of a test score" (Cizek, 2007, p. 20).

Given the importance of validity and reliability in science assessment, how does one determine whether test scores generate valid and reliable inferences? Fortunately, most of the devices that we rely on every day must meet particular validity and reliability standards before they can be sold or used. When it comes to educational measurements, rigorous standards and guiding documents have also been established. A foundational source for established perspectives on validity and reliability (and how they should be operationalized) is *Educational Measurement*. For the past 60 yr, the American Council on Education and the National Council on Measurement in Education (NCME) have produced several editions of this book (e.g., Brennan, 2006). In terms of measurement standards, the AERA, the American Psychological Association (APA), and the National Council on Measurement in Education have jointly produced several versions of the *Standards for Educational and Psychological Testing* (most recently, AERA *et al.*, 1999). These guiding documents and standards provide very rigorous, technical, and detailed quality control criteria for test development, evaluation, and use. While different standards of evidence are obviously required for different types of tests (e.g., the Graduate Record Exam vs. a concept inventory), *all* tests, whether for the purposes of classroom or large-scale assessment, should provide *some* form of validity and reliability evidence.

Table 1. Sources of evidence to consider when establishing or evaluating construct validity[a]

Source of validity evidence	Answers this question	Methodology example(s)[b]	Example-related measurement standard(s)[c]
A. Content	Does the assessment appropriately represent the specified knowledge domain?	Delphi Study; textbook analyses; expert survey; Rasch analysis	1.6
B. Substantive	Are the thinking processes thought to be used to answer the items the ones that were actually used?	"Think aloud" interviews during problem solving; cognitive task analysis	1.8
C. Internal structure	Do the items capture one dimension or construct?	Factor analysis; Rasch analysis	1.11
D. External structure	Does the construct represented in the assessment align with expected external patterns of association (convergent and/or discriminant)?	Correlation coefficients	1.14
E. Generalization	Are the scores derived from an assessment meaningful across populations and learning contexts?	Analyses of performance across a diversity of contexts (e.g., ethnicity, socioeconomic status, etc.); differential item functioning	1.5
F. Consequences	In what ways might the scores derived from the assessment lead to positive or negative consequences?	Studying the types of social consequences produced as a result of using test scores (e.g., passing a class, graduating from a program).	1.22 and 1.24

[a]Modified from Messick (1995) and Nitko and Brookhart (2010).

[b]Methodology examples are based on both the classical test theory and item-response theory. For more information about these perspectives, their implicit assumptions, and how they may be used to gather validity and reliability evidence, see chapters in *Educational Measurement* (Brennan, 2006) and the *Handbook of Test Development* (Downing and Haladyna, 2006).

[c]From AERA *et al.* (1999).

Validity

Validity, in an educational sense, refers to issues relating to whether you are measuring what you are claiming to be measuring. This brings up a host of questions, such as: What, exactly, is genomics understanding, and how do I know that I am really measuring it? What, exactly, are inquiry skills, and am I really measuring them? There are many aspects of validity that need to be considered in educational measurement; Messick (1995) identifies six such aspects in his unified model of construct validity. These six aspects include: 1) the intended content coverage (what, exactly is the test supposed to measure, and is it, in fact, measuring it? (see Table 1A); 2) the cognitive processes thought to be used to answer an item (Table 1B); 3) how well the questions align with one another or "hang together" (Table 1C); 4) how well the assessment scores align with other assessments attempting to measure the same or similar constructs (Table 1D); 5) the generalizability of score inferences produced by the assessment (Table 1E); and 6) the potential consequences of using scores derived from the assessment (Table 1F; Nitko and Brookhart, 2010). These (and many other sources of evidence not emphasized by Messick) may be required depending on the purpose of the assessment tool.

Content Validity Evidence (Table 1A). Content evidence pertains to determining how well the items on an assessment match up with the target domain or construct (e.g., "natural selection"). Whether the intended construct to be measured has been defined by 1) collaborating with experts in the field of interest; 2) referencing framework standards; 3) analyzing textbook content; or 4) adhering to a school curriculum that directly influences the design of a program or course, this evidence must consist of 1) how relevant a particular topic (e.g., transcription) is to the intended construct (e.g., genetics) and 2) how well the chosen topics represent the construct as a whole (i.e., content coverage; Messick, 1999). In general, a researcher must provide validity evidence for the criteria used to include or exclude content topics in the assessment tool (Liu, 2010; Nitko and Brookhart, 2010).

It is particularly important that the items developed to measure carefully specified aspects of a topic or construct do not overrepresent or underrepresent it; indeed, these are two major threats to content validity (Messick, 1995). If a test intended to measure knowledge of natural selection asks students questions about plate tectonics, for example, the test will not generate valid inferences about natural selection knowledge (domain-irrelevant measurement variance would be introduced). If a test about natural selection does not include questions about differential survival and reproduction, the test would not generate valid inferences about natural selection knowledge due to content *underrepresentation* (the measure does not capture all of the central aspects of natural selection knowledge). Thus, careful attention to what should (and should not) be included relative to the specified focus of the test is central to test design and validity inferences (Brennan, 2006; Downing, 2006; Haladyna, 2006; Kane, 2006; Liu, 2010; Nitko and Brookhart, 2010).

Substantive Validity Evidence (Table 1B). To obtain substantive validity evidence, test developers must first consider what thinking processes and skills are necessary in order to

complete tasks within the construct. Evidence must be provided to show that the tasks on an assessment engage these processes. For example, a test developer might believe that one of the multiple-choice items on an assessment requires higher-order thinking processes in order to be solved. Assuming that students answer this item correctly, the test developer might infer that students used higher-order thinking processes to solve the problem. However, without obtaining substantive validity evidence, it is unclear whether the students used test-taking skills to solve the problem (such as noticing that the correct option contained more scientific jargon than the other options), or whether they actually engaged in the expected cognitive processes. One approach for gathering substantive validity evidence is to have students participate in "think aloud" interviews, in which they verbally communicate their thinking processes as they complete assessment tasks (Messick, 1995; Nitko and Brookhart, 2010).

Internal Structure Validity Evidence (Table 1C). Internal structure evidence helps to build a case that each of the individual questions on an assessment aligns with the overall target of the assessment. Statistically speaking, we are interested here in how well the questions correlate with one another. Items that show a strong correlation suggest that they are measuring the same thing (e.g., knowledge, response processes, etc.), while those with weak correlations suggest that they measure different things. The overall goal is to ensure that all of the items contribute to the differentiation or diagnosis of student cognitive/affective/psychomotor knowledge along a single dimension. This is important, as a typical characteristic of an assessment tool is to independently measure one dimension or construct (Brennan, 2006; Nitko and Brookhart, 2010).

External Structure Validity Evidence (Table 1D). While internal structure evidence looks at how well items on an assessment hang together, external structure evidence pertains to whether or not measurement scores appropriately correlate with other measures based on the expected relationship between the constructs aligned with those measures. "That is, the constructs represented in the assessment should rationally account for the external pattern of correlations" (Messick, 1995, p. 746). One can provide external structure evidence by comparing the scores on the assessment with other tests that cover the same target domain; this is often referred to as providing convergent validity evidence (e.g., see Nehm and Schonfeld, 2008). For example, two different concept inventories attempting to measure the same construct (e.g., genetics: Genetics Concept Inventory [Elrod, 2007] and the Genetics Concept Assessment [Smith *et al*., 2008]) would be expected to generate strong and significant associations.

In contrast, one must also consider discriminant evidence. This source of evidence helps to establish that the assessment is *not* measuring things that we claim it is not measuring (e.g., reading ability). Such evidence is important, as we would not expect unrelated domains to be linked. If, for example, one found that seemingly unrelated measures showed strong associations, it would likely mean that something has gone wrong with our construct definition or item design, or that both tests are measuring a shared, underlying construct. For example, it is possible that assessment scores derived from a verbal paragraph analysis are strongly correlated with scores derived from an assessment of mathematical word problems; in this case, it may be that *reading comprehension* is in fact the construct that is being captured by these assessments. In sum, convergent and discriminant evidence are commonly used to build a validity argument for an instrument prior to its use in educational research.

Generalization Validity Evidence (Table 1E). When an assessment is created, researchers attempt to validate the inferences derived from the assessment scores. These inferences are validated in a specific context: that is, how, where, to whom, and under what conditions the assessment was administered (e.g., timed, untimed, in-class, homework, etc.). Many contextual factors can affect validity inferences. If the context changes (e.g., from administering to high school students rather than undergraduates), the validity of the inferences may not apply (unless claims of equivalency have been supported with evidence). Inferences may be generalized when the validation process has been examined under many different circumstances or with a very large and diverse population (Brennan, 2006; Kane, 2006; Nitko and Brookhart, 2010).

While it is not necessary to validate inferences beyond your own research population, it is necessary to report the characteristics of that population and the circumstances under which the assessment tool was administered (e.g., a timed assessment or one in which incentives were offered) in order for other researchers to determine whether such inferences may generalize to their particular sample of participants. Assessments do *not* need to be designed to provide *generalizable* inferences; rather, it is crucial that they provide valid and reliable inferences for the population they were designed to assess. Problems arise when educators assume inferences will generalize across contexts and populations.

Consequential Validity Evidence (Table 1F). One must also consider what effects the use of scores derived from an assessment will have on both those who administer the tool and those who take it. The ease of use, instructional benefits, consequences for students, and other mitigating factors should be considered both in the design and in the use of an assessment. Indeed, 1) consequences should be matched to the design and purpose of an assessment tool and 2) attempts should be made to anticipate any unintended consequences of using the test and the scores derived from it. The consequences of the assessment can help dictate how, when, and where it should appropriately be used (Haladyna, 2006; Nitko and Brookhart, 2010). Nevertheless, the consequences of using scores derived from assessments are subjective judgments. That is, how significant is it if a formative assessment tells a teacher that students do *not* have a misconception, when in fact they do, and the students never learned the concept? How significant is it if a student knows two more concepts than an assessment indicates and receives a lower course grade? The severity of score-use consequences is not always clear. In our view, we should work to ensure that accurate information informs all of our instructional decisions (be they formative or summative).

Construct Validity Evidence. Once data have been gathered for the above six sources of evidence, this evidence needs to be examined as a collective whole to determine whether the inferences made based on the assessment scores have construct validity. It is not guaranteed that all of the sources

of evidence will agree with one another, nor is it necessary for them to do so. If construct validity has been established, then—and only then—can we ensure that the inferences we make based on the assessment are valid. Indeed, if only one source of evidence (e.g., content validity) is used to develop and validate an assessment, and inferences about learning gains are made, one is assuming that the other sources of evidence do not contradict these findings. Recent emphasis on evidence-based practice is consonant with the idea that diverse forms of corroborating validity evidence should inform the instrument development.

Reliability

Reliability is one of the most important aspects of measurement (Brennan, 2006). In a practical sense, reliability confronts us every day of our lives: Does the gas gauge in our car consistently indicate how much gas is present in the tank? Does a thermometer consistently represent how cold it is outside? If the instruments used to produce these measures (e.g., gas gauge, thermometer) did not provide us with consistent information or outputs, they would be of little use. In an analogous way, unreliable test scores might at one point indicate that a group of students knew a lot, when in fact they knew very little, and at another point indicate that these same students knew a little, when they in fact knew a lot. In educational testing, reliability may be briefly defined as the consistency across replications of a measurement procedure (Brennan, 2001). These replications can refer to multiple administrations of the same assessment, the administration of two different, but equivalent, assessments, or the scoring of the assessment by multiple graders (Liu, 2010).

There are many different ways to establish whether test scores generate reliable inferences; this depends on what type of consistency is of concern. One can look at the reliability of inferences made over a span of time (stability) by testing a sample of students once, waiting a period of time, and retesting (Table 2A). Alternate forms of reliability can be tested by using parallel versions of the same assessment, or if the use of parallel versions is not appropriate, by splitting the item set into equal parts and examining the correlation between the two outputs (Table 2B). The internal consistency of the items can be determined by analyzing item-response patterns. Internal consistency can be quantified as coefficients (such as Kuder-Richardson or Cronbach's alpha), which are equivalent to the mean of all possible split-half reliability estimates (Table 2C). Finally, for open-response assessments in particular, the reliability of the assessment *raters* is also important and can be assessed through the use of interrater reliability metrics to determine whether the graders appear to be using the same criteria when scoring the assessment (Table 2D). In sum, "reliability" never refers to a single value, as reliability is an umbrella term encompassing several different types of evidence.

Reliability, Validity, and GBE Assessments

Our review of some of the most basic validity and reliability concepts has attempted to demonstrate their unavoidable importance to all forms of scientific research in education (NRC, 2002) and scientific teaching (i.e., using appropriately gathered evidence to inform instructional practices). While the characteristics we have reviewed are important factors that should be considered in all types of assessments, it is beyond the scope of this study to determine whether the assessments in the articles we reviewed are capable of producing valid and reliable inferences in the populations and contexts in which they were used. Rather, we assessed whether any of the articles attempted to establish validity or reliability evidence prior to the administration of their assessments.

METHODOLOGY

The scope of our literature review and analysis of GBE was 1) to examine the characteristics of the educational research that assessed learning targets in GBE and 2) to determine the degree to which evidence-based measurement standards are being followed (see *Framework for Instrument Development and Evaluation* above).

To obtain a representative sample of the research in GBE, we used a university database system to conduct a keyword search for research articles published between 1995 and 2010. A search using "bioinformatics education" and "genomics education" produced 226 articles, of which, 95 were given further attention due to their use of assessments to evaluate knowledge increases, affective changes, or skill acquisition. Examples of these evaluations would include sentences such

Table 2. Sources of reliability evidence to consider when creating or evaluating an assessment[a]

Source of reliability evidence	Answers this question	Methodology example(s)[b]	Related measurement standard(s)[c]
A. Stability	How consistent are scores from one administration of the assessment to another?	Stability coefficient	2.4
B. Alternate forms	Are scores comparable when using similar items to assess the same construct?	Spearman-Brown double length formula: split half	2.4
C. Internal consistency	To what extent do the items on an assessment correlate with one another?	Coefficient alpha or Kuder-Richardson 20	2.4
D. Reliability of raters	Is the assessment scored consistently by different raters?	Cohen's or Fleiss's kappa	2.10

[a]Modified from Nitko and Brookhart (2010).

[a]See Table 1, footnote b.

[b]Examples from AERA *et al*. (1999).

as, "Assessment results showed that students gained an understanding of the Web-based databases and tools and enjoyed the investigatory nature of the lab" (Bednarski *et al*., 2005, p. 207); "The students performed well on independent problem sets, and their feedback about the module was generally positive" (Magee *et al*., 2001, p. 855); and "In comparison with traditional in-person teaching labs, students preferred the virtual lab by a factor of two" (Weisman, 2010, p. 4). An example of an article that would not be included in our analysis would be "Guidelines for Establishing Undergraduate Bioinformatics Courses" (Cohen, 2003). This article performs no assessment of learning targets and only provides information on how to design a bioinformatics course. (A document referencing the 95 articles that we reviewed, along with the classification of these articles, may be found in the Supplemental Materials.)

Methodologically, a modified version of content analysis (Krippendorff, 2003) was used to analyze the 95 articles. In particular, the abstract, methodology, results, and discussion sections were read in order to determine what curricular and educational levels (scale) were addressed and what learning goals (affective, cognitive, and psychomotor) were targeted. These categorizations were made to not only examine the types of assessment work that has been completed, but also to examine whether assessment quality varies among educational levels and targets. A scoring rubric was used to identify and tally facets of measurement quality, including the key indicators of validity and reliability evidence. The coding of articles for rubric features was also examined for scoring reliability (see *Interrater Reliability* below).

Interrater Reliability

Two human raters with graduate degrees in the biological sciences performed the content analysis of the articles in order to ensure scoring reliability. A common approach for measuring the level of agreement between raters is through the use of Cohen's kappa statistic. Values for this reliability measure range between 0.0 and 1.0, with scores > 0.81 considered to be at an "almost perfect" level of agreement (Landis and Koch, 1977). When there was a scoring disagreement between the raters, a consensus score was achieved through deliberation; the consensus scores were then used for our analyses. We performed the kappa calculation in SPSS version 19.0 (IBM; 2012).

Scale and Learning Targets

To determine at which educational scales and learning targets GBE assessments have (and have not) been completed, and to determine whether assessment quality differed across these scales and targets, we categorized the assessment articles into one of five curricular levels: program, course, lesson, professional, or resource development. Articles that discussed program development assessed learning targets over multiple courses. Those that discussed course development assessed learning targets in a single course. Articles categorized as lesson development assessed learning targets through interventions lasting one or more days during a course. Those that analyzed professional development assessed learning targets after some form of professional education class, workshop, or course. Finally, those articles that discussed GBE resources examined the use of public, Web-based databases or tools (such as BLAST, online virtual labs, or other learning tools).

Articles were also categorized based on the educational level that was evaluated, specifically, secondary, undergraduate, or graduate levels. Those articles that assessed the secondary level had a focus on grades 10–12; articles classified as undergraduate focused on college or university classes or programs designed to culminate in a baccalaureate degree; articles classified as graduate focused on classes or programs culminating in a master's or doctoral degree.

All articles in our sample were also analyzed to identify the learning targets (affective, cognitive, or psychomotor) that were the focus of the studies. The affective target is associated with how students feel about or value something; some examples would be students' interests in genomics or whether they value the material taught in particular classes. The cognitive target is associated with intellectual knowledge and concepts, for example learning the composition of a genome or the sequence of a gene. Finally, the psychomotor target is associated with procedures or tasks, such as learning how to use a microscope or how to access information in the National Center for Biotechnology Information (NCBI) database.

Validity and Reliability

Regardless of educational scale or learning target, the focus of our investigation was examining the presence of validity and reliability evidence in GBE. To make our content analysis as encompassing as possible, we used keyword searches in these articles for "validity," "reliability," "trustworthiness," "verify," "consistency," and "interrater agreement" (as well as variations thereof). As in our content analysis, the paragraphs containing keywords were examined for validity and reliability evidence. Further, 20 randomly selected papers were read in full to verify that our keyword search did not miss any information on validation, reliability, or quality control information.

RESULTS

Interrater Reliability

For the two raters, coding of the papers in the sample using the content analysis rubric produced kappa scores of 0.920 ($n = 28$ observations $\times$ 4 categories $\times$ 7 articles), which exceeded the "almost perfect" target value of 0.81.

Educational Scale, Learning Targets, and Assessment Purposes

We initially categorized the GBE assessment studies in our sample into educational scale and learning targets, because we considered the possibility that validity and reliability evidence might differ among these categories. Our analysis of the 95 articles indicated that, in terms of instructional scale, 50 (52.63%) assessed aspects of newly created or modified courses, 26 (27.37%) assessed newly created or modified lessons, nine (9.47%) assessed unique resources, seven (7.37%) assessed newly created or modified programs, and three (3.16%) assessed professional development workshops. In terms of educational level, our analysis revealed that 83 articles clearly fit our classification criteria, but 12 did not.

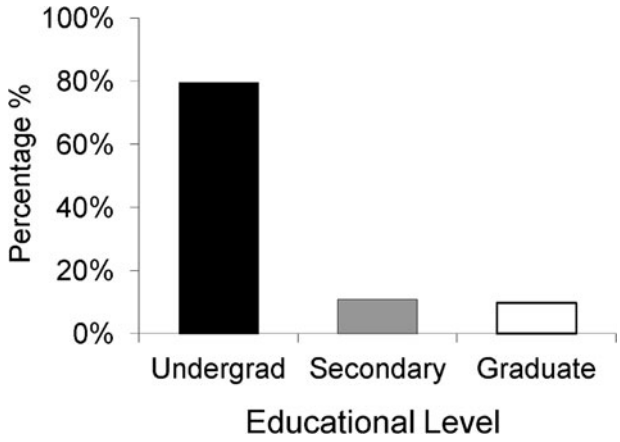

Figure 1. Percentage of articles that evaluated each educational level (secondary, undergraduate, and graduate) in all articles ($n = 83$).

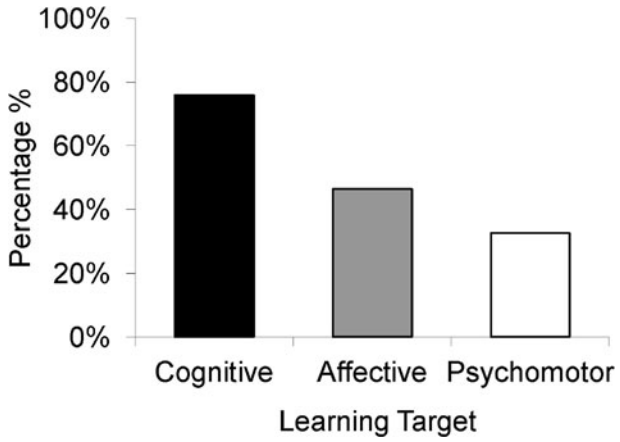

Figure 2. Percentage of articles that assessed each learning target (cognitive, affective, and psychomotor) in all articles ($n = 95$).

Specifically, nine papers (10.84%) focused on the secondary level, 66 (79.52%) focused on the undergraduate level, and eight papers (9.64%) focused on the graduate level. The 12 articles that did not fit our coding scheme focused on resources or professional development that could theoretically be used at more than one educational level.

The categorization of papers by educational level was also examined by curricular level (Figure 1). The analysis of articles that assessed new or modified courses indicated that five (10%) articles focused on the secondary level, 39 (78%) focused on the undergraduate level, and six (12%) focused on the graduate level. Those articles that evaluated new or modified lessons revealed that three (11.50%) assessed the secondary level, 21 (80.81%) assessed the undergraduate level, and two (7.69%) assessed the graduate level. The remaining curricular categories (program, professional, and resource) are not shown, as these particular categories were very sparsely populated.

The analysis of the papers by learning target revealed that 74 papers (77.9%) assessed cognitive targets, 47 papers (49.5%) assessed affective targets, and 32 papers (33.7%) assessed psychomotor targets (Figure 2). This tally produced a total of 153 assessed learning targets in 95 papers (i.e., many authors chose to analyze multiple targets in their studies).

The categorization of papers by learning targets was also analyzed by curricular level. Examining only those articles that assessed new or modified courses in GBE, 41 (82%) assessed cognitive targets, 23 (46%) assessed affective targets, and 15 (30%) assessed psychomotor targets. When we focused on those articles that assessed new or modified lessons in GBE, we found 19 (73.08%) assessed cognitive targets, 15 (57.7%) assessed affective targets, and 12 (46.2%) assessed psychomotor targets. Similar to the analysis at the educational level, the remaining categories were not shown, because the number of cases in each category was extremely low.

Finally, we examined the purposes of the assessments in our sample (which relates to assessment consequences; Table 1F). The 95 publications used summative assessments to make claims about student learning; that is, at the termination of learning events, the assessments attempted to measure cognitive factors, such as comprehending the significance of differential gene expression in microarray data; affective factors, such as whether or not students enjoyed learning about genomics and bioinformatics or whether genomics and bioinformatics were considered as career options; and procedural skills, such as identifying a gene from a DNA fragment on the NCBI website.

Validity and Reliability Evidence

Of the 95 articles containing measurement tools or instruments, our content analysis and search for the key terms "validity," "reliability," "trustworthiness," "verify," "consistency," and "interrater agreement" found only seven articles (7.4%; Table 3) that contained language associated with measurement quality benchmarks and standards. Five of the seven articles made reference to validity (for specific validity sources and methodologies, see *Discussion*), and three articles made reference to reliability (specifically through interrater agreement). It must also be noted, however, that in these seven papers, only three used multiple assessment tools to establish convergent validity evidence (Table 1D). Nevertheless, validity and reliability evidence for each of the assessments used to establish convergent validity evidence was lacking. Of the 95 papers we examined, the only study that mentioned both validity and reliability evidence was Gelbart *et al*. (2009). Overall, our analysis revealed that the vast majority of GBE studies, regardless of educational scale or learning target, lacked reference to the most basic quality control criteria.

Table 3. Articles containing keywords pertaining to education validity and reliability

Author(s)	Valid	Interrater agreement
Chapman *et al*., 2006	X	—
Gelbart *et al*., 2009	X	X
Herron, 2009	—	X
Howard *et al*., 2007	X	—
McEwen *et al*., 2009	X	—
Wefer and Anderson, 2008	X	—
White *et al*., 2002	—	X

Finally, we examined the articles in our sample to determine whether an attempt was made to explicitly delineate the construct that the assessment was built to represent, or whether sources of validity evidence were integrated into a cohesive model. Even though a few articles did present one or two sources of validity evidence (specifically, one article discussed content evidence and four articles included external evidence), none of the studies in our sample provided sufficient evidence to support inferences made about presumptive constructs. Furthermore, not only was there little evidence to support construct validity, there was also no explicit evidence presented that defined the underlying construct intended to be measured. Finally, almost no attention was directed at what content, skills, and dispositions *should* be assessed.

DISCUSSION

Genomics and bioinformatics are prime examples of interdisciplinary fields with considerable potential for solving some of the world's most pressing problems (NRC, 2009). There is an increasing need for educating students about genomics and bioinformatics knowledge and associated tools (e.g., online databases). Since 1995, many educational innovations have been developed for students at multiple educational levels to learn about genomics and bioinformatics. To ensure that these learning opportunities are achieving their intended goals, researchers must perform efficacy studies and use evidence to support evaluative judgments (Dolan, 2008). While many researchers have indeed made an attempt at assessing these new learning opportunities, there has been little work done to determine whether these assessments meet the quality control benchmarks set forth by the educational measurement community, or whether they have produced evidence appropriate for guiding educational reform (NRC, 2012).

Validity and Reliability Evidence

While <10% of articles contained any reference to the establishment of validity and reliability evidence, the seven papers that did contain some evidential data cannot automatically be assumed to have provided sufficient evidence or to have collected that evidence in an appropriate manner. Therefore, it is necessary to discuss not only what types of evidence were (or were not) presented in each article, but also how that evidence was collected.

Content Validity Evidence (Table 1A). The article by Gelbart *et al*. (2009) is the only study that includes content validity evidence. Their quantitative analysis involved a questionnaire for which the content was reviewed by two researchers in the field. It is alarming that no other assessments established this type of evidence, as without it, there is no way to determine whether the items on the assessment are relevant to the construct intended to be measured.

External Validity Evidence (Table 1D). Chapman *et al*. (2006), Howard *et al*. (2007), Wefer and Anderson (2008), and McEwen *et al.* (2009), all used multiple assessments to measure their presumptive constructs. In particular, the articles by Chapman *et al*. (2006), Howard *et al*. (2007), and Wefer and Anderson (2008) used cross-validation to provide convergent external validity evidence. While these authors used multiple assessment tools to measure the same content, this approach only has meaning if the assessments with which you are comparing your own assessment already have supporting validity evidence; there is no mention of such evidence in these studies. While these authors have produced data that support convergent external validity evidence, all the assessment tools may in fact be measuring inappropriate content (i.e., not supported by other experts in the field). Finally, if the same person created these tools, it is likely that any discrepancies or inconsistencies would be present in all the assessment tools, compounding potential problems.

McEwen *et al.* (2009) used a previously validated instrument, the Science Laboratory Environment inventory (SLEI), to assess the psychosocial component of their study. It is unclear whether the SLEI can be generalized to the population they are assessing or whether they made an attempt to test validity inferences with their population. Furthermore, McEwen *et al.* did not provide validity evidence pertaining to the assessment used for the cognitive component of their article.

Construct Validity Evidence. The lack of a clear construct definition was not unique to the 95 articles reviewed in this study; further analysis of our full sample revealed that this large body of work has yet to provide 1) a clearly articulated consensus definition of genomics and/or bioinformatics or 2) a robust conceptualization of content, skills, and dispositions central to the domain (Campbell *et al*., 2012). Clearly, more work is needed in order to establish these domains, so we may discover which facets of GBE are most worthy of assessment.

Reliability Evidence for Responses, Not Scorers (Table 2, A–C). The fact that none of the 95 articles that we studied made any reference to response reliability is concerning (interrater agreement refers to reliability of graders and not to response patterns). In a field in which designing controls and performing replicates of experiments is crucial for building scientific understanding, one might expect that the assessment tools developed by biology educators would also contain reliability measures. It is possible that reliability evidence was in fact gathered, but the teacher-researchers felt it was not necessary to include such information in their articles simply as an artifact of different publishing requirements and expectations between the biological and educational sciences.

While we have alluded to several quality control standards that discipline-based education journals could apply, such standards need to be collaboratively defined in alignment with the goals and expectations of the research community for each publication outlet. Nevertheless, in our view, *some* evidence pertaining to both validity and reliability should be required of all assessment types—formative, summative, cognitive, affective, and psychomotor. In our view, faulty inferences are problematic regardless of whether the assessments that produce them are low-stakes (e.g., formative for the instructor) or high-stakes (e.g., summative for a capstone project needed for graduation).

Interrater Reliability Evidence Only (Table 2D). White *et al.* (2002), Gelbart *et al*. (2009), and Herron (2009) used measurement tools such as midterms, final projects, and surveys to assess students' cognitive, affective, and psychomotor knowledge. While the evidence to support response reliability of

Table 4. Articles mentioning the importance of validity in natural sciences, but not education[a]

Ackovska and Madevska-Bogdanova, 2005	Furge *et al.*, 2009
Almeida *et al.*, 2004	Hingamp *et al.*, 2008
Bergland *et al.*, 2006	Kuldell, 2006
Brame *et al.*, 2008	Lopatto *et al.*, 2008
Buckner *et al.*, 2007	Luo *et al.*, 2007
Butler *et al.*, 2008	Malacinski and Zell, 1995
Centeno *et al.*, 2003	Qin, 2009
Curioso *et al.*, 2008	Rowland-Goldsmith, 2009
Dymond *et al.*, 2009	Shachak *et al.*, 2005
Farh and Lee, 2007	Shaffer *et al.*, 2010

[a]Articles are arranged alphabetically by last name of first author.

these summative assessments is not mentioned, the authors did employ multiple raters to grade them, providing reliability evidence for scoring. However, such information does not address whether these *assessments* display high-quality attributes.

Validity and Reliability Evidence in Genomics and Bioinformatics Scientific Data. During our content analysis, we noticed that, in addition to the seven articles using language associated with measurement quality in their educational research, 20 other articles (Table 4) mentioned the importance of having *validated genomic data*, but did *not* relate such language to their educational research. This finding suggests that the authors value valid and reliable genomic data but they have not transferred this importance to their educational efforts.

Implications of the Results

As Campbell aptly notes: "Why should readers of educational journals accept claims of improved learning without data? Biologists would not accept any new discoveries in research without data to support the claims" (Campbell, 2003, p. 106). This important perspective is nevertheless incomplete and must be taken one step further: data must be evaluated relative to the *quality* of the *tools* that are used to generate it. Evidence-based educational research requires quality *measures* in order to produce quality *data* and quality *inferences* about the data. The entire chain of information production must be considered, not just the *data*.

The findings of our review of the GBE literature are concerning and preclude "evidence-based" or "scientifically based" decision making about how genomics and bioinformatics is best taught and learned. Unfortunately, more than 90% of the peer-reviewed research containing assessment tools about genomics and bioinformatics had no validity or reliability backing. Given that these authors were working toward better scientific understanding of teaching and learning, it is of utmost importance that teacher-researchers in these areas apply the tools, methods, and standards for generating high-quality assessment instruments in order to provide high-quality evidence of teaching and learning within genomics and bioinformatics.

How Does GBE Assessment Quality Compare with Other Fields?

Although a comprehensive review of instrument quality in educational research in general (or biology education in particular) has not been performed, a few noteworthy reviews of science education instruments exist and provide a useful vantage point from which to view our GBE assessment quality findings (Britton and Schneider, 2007; Fraser, 2007; Liu, 2009, 2010). For example, in his review of 50 science education measurement instruments reported in refereed publications since 1990, Liu (2009) found that 100% included validity evidence, and 88% contained reliability evidence. Notably, many instruments contained multiple facets of validity evidence (i.e., those outlined in Table 1, such as content and internal structure validation) and multiple aspects of reliability evidence (e.g., interrater reliability, Cronbach's alpha). Instruments developed for relatively new areas of research, such as creativity in science (Hu and Adey, 2002) and ill-structured problem-solving processes (Shin *et al.*, 2003), for example, contain multiple sources of validity and reliability evidence. In contrast, Nehm (2006) reported on a brief review of 200 evolution education studies and noted an overall lack of rigor relating to assessments, echoing many of the concerns raised in our study of GBE assessments.

Overall, although there is only a small body of prior work on assessment quality to draw upon, it is clear that 1) many instruments developed in the field of science education since 1990 were developed using multiple sources of validity and reliability evidence; 2) relatively new areas of research do not appear to be more likely to lack such evidence; and 3) some domains within science education (e.g., evolution education) appear to suffer from the same limitations as assessment work in GBE. It is clear that broad generalizations about instrument quality cannot be made with confidence.

Looking Forward

Our study has uncovered a concerning lack of quality in educational measurement work in GBE. Our concerns with assessment are hardly new, however. In 2003, for example, Campbell noted: "Although assessment has become a buzz word, it remains a mystery to many of us" (Campbell, 2003). One decade later, it is apparent that the efforts put forth to increase awareness of the importance of high-quality assessments in biology education have had little apparent impact on GBE research. It is our hope that through this article we will bring further awareness of measurement concepts and standards to the GBE community, for it is only through the use of these measurement concepts and standards that reforms within GBE can be made confidently and appropriately. Complying with these standards will also enable researchers and practitioners to know what topics *should* be assessed within GBE and to then use this knowledge to target curricular reform and pedagogical innovation.

Our study also attempted to provide a starting place for those who wish to move the field forward. Given that it can be a daunting task to read and comprehend the quality control benchmarks in full, there are other avenues that one can pursue on the road to high-quality measurement. One approach would be to adopt our framework (summarized in Tables 1 and 2) and its associated set of "best practices." Another approach would be to establish collaborations with faculty

already familiar with educational measurement standards (e.g., educational measurement faculty) or to seek the advice of education faculty members at your, or a nearby, university. If you are not sure whom to contact, you can become engaged in an educational organization, such as the Society for the Advancement of Biology Education Research (SABER), the National Association for Research in Science Teaching (NARST) or AERA, and attend talks and workshops at their annual conferences, such as the workshop entitled "Introduction to Instrument Development and Evaluation in Science Education" at the 2012 NARST conference. One may also apply to an assessment residency with Biology Scholars (a National Science Foundation (NSF)-funded program aimed at improving undergraduate learning through evidence-based assessment).

Finally, there are many books available that review validity and reliability concepts and methods. Nitko and Brookhart's *Educational Assessment of Students* (2010) provides an easy-to-understand introduction to assessment design. The *Handbook of Test Development*, edited by Downing and Haladyna (2006), offers a more comprehensive explanation, with direct references to the standards as well as the mathematical calculations used for many types of validity and reliability evidence. Finally, the National Council on Measurement in Education's *Educational Measurement*, edited byBrennan (2006), offers a complete guide to assessment development, administration, and analysis.

CONCLUSIONS

Genomics and bioinformatics are prime examples of integrative, cross-disciplinary scientific fields emblematic of the future of the biological sciences (NRC, 2009). In line with the recommendations of the authors of *A New Biology For The 21st Century* (NRC, 2009), a large body of work (>200 peer-reviewed studies) has documented the outcomes of educational reform efforts designed to bring life sciences teaching and learning in line with the dramatic scientific developments of the past few decades. Nevertheless, as our analysis of this body of work illustrates, while the efforts put forth may in fact have generated useful outcomes, the majority of findings do not meet the norms of scientific research in education (NRC, 2002) and fail to meet the most basic of educational measurement standards (AERA *et al.*, 1999). Our findings suggest that robust, evidence-based claims are lacking for GBE, weakening efforts to employ scientific teaching in this important area of the life sciences (cf. Handelsman *et al.*, 2006).

While our critical analysis has revealed concerning weaknesses with the educational research that has been completed thus far GBE, we are hopeful that the growing biology education research community (e.g., SABER; see also NRC, 2012) will embrace reform movements emphasizing evidence-based decision making in biology education (e.g., NRC, 2002; AAAS, 2009) and pursue collaborative relationships with the science education and educational measurement communities (e.g., NARST, AERA). We also hope that our review of some of the core aspects of educational measurement and our introduction of key standards documents and resources provides useful avenues for future assessment efforts in GBE. We are confident that robust educational evidence can be established and profitably applied to GBE reform.

ACKNOWLEDGMENTS

We thank several colleagues for helping make our review as constructive as possible. Portions of this work were funded by an NSF Course, Curriculum, and Laboratory Improvement program grant (0837397).

REFERENCES

Ackovska N, Madevska-Bogdanova A (2005). Teaching bioinformatics to computer science students. Int Conf Comput Tool EUROCON *1*, 811–814.

Almeida CA, Tardiff DF, De Luca JP (2004). An introductory bioinformatics exercise to reinforce gene structure and expression and analyze the relationship between gene and protein sequences. Biochem Mol Biol Educ *32*, 239–245.

American Association for the Advancement of Science (2009). Vision and Change in Undergraduate Biology Education. http://visionandchange.org (accessed 15 March 2010).

American Educational Reasearch Association, American Psychological Association, and National Council on Measurement in Education (1999). The Standards for Educational and Psychological Testing, Washington, DC.

Bednarski AE, Elgin SCR, Pakrasi HB (2005). An inquiry into protein structure and genetic disease: introducing undergraduates to bioinformatics in a large introductory course. Cell Biol Educ *4*, 207–220.

Bergland M *et al.* (2006). Exploring biotechnology using case-based multimedia. Am Biol Teach *68*, 81–86.

Brame CJ, Pruitt WM, Robinson LC (2008). A molecular genetics laboratory course applying bioinformatics and cell biology in the context of original research. CBE Life Sci Educ *7*, 410–421.

Brennan RL (2001). Generalizability Theory, New York: Springer.

Brennan RL (2006). Educational Measurement, 4th ed., Lanham, MD: Rowman and Littlefield.

Britton ED, Schneider SA (2007). Large-scale assessments in science education. In: Handbook of Research on Science Education, ed. SK Abell and NG Lederman, Mahwah, NJ: Lawrence Erlbaum, 1007–1040.

Buckner B *et al.* (2007). Involving undergraduates in the annotation and analysis of global gene expression studies: creation of a maize shoot apical meristem expression database. Genetics *176*, 741–747.

Butler PJ, Dong C, Snyder AJ, Jones AD, Sheets ED (2008). Bioengineering and bioinformatics summer institutes: meeting modern challenges in undergraduate summer research. CBE Life Sci Educ *7*, 45–53.

Campbell AM (2003). Public access for teaching genomics, proteomics, and bioinformatics. CBE Life Sci Educ *2*, 98–111.

Campbell C, Nehm RH (2011). Assessing the educational efficacy of genomics and bioinformatics curricula and labs. Paper presented at the Society for the Advancement of Biology Education Research (SABER) conference, 29–30 July 2011, Minneapolis, MN.

Campbell C, Nehm RH, Morton B (2012). Building new assessments for the "New Biology": establishing content validity for a genomics and bioinformatics test.. Proceedings of the National Association for Research in Science Teaching (NARST) Annual Conference, 24–28 March 2012, Indianapolis, IN.

Centeno NB, Villà-Freixa J, Oliva B (2003). Teaching structural bioinformatics at the undergraduate level. Biochem Mol Biol Educ *31*, 386–391.

Chapman BS, Christmann JL, Thatcher EF (2006). Bioinformatics for undergraduates: steps toward a quantitative bioscience curriculum. Biochem Mol Biol Educ *34*, 180–186.

Cizek GJ (2007). Introduction to validity. Presentation to the National Assessment Governing Board, University of North Carolina at Chapel Hill, August 2007, McLean, VA.

Cohen J (2003). Guidelines for establishing undergraduate bioinformatics courses. J Sci Educ Technol *12*, 449–456.

Curioso W, Hansen J, Centurion-Lara A, Garcia P (2008). Evaluation of a joint bioinformatics and medical informatics international course in Peru. BMC Med Educ *8*, 1.

Dolan EL (2008). Education Outreach and Public Engagement, New York: Springer.

Downing SM (2006). Twelve steps for effective test development. In: Handbook of Test Development, ed. SM Downing and TM Haladyna, Mahwah, NJ: Lawrence Erlbaum, 3–25.

Downing SM, Haladyna TM (2006). Handbook of Test Development, Mahwah, NJ: Lawrence Erlbaum.

Dymond JS, Scheifele LZ, Richardson S, Lee P, Chandrasegaran S, Bader JS, Boeke JD (2009). Teaching synthetic biology, bioinformatics and engineering to undergraduates: the interdisciplinary build-a-genome course. Genetics *181*, 13–21.

Elrod S (2007). Genetics Concept Inventory. http://bioliteracy.colorado.edu/Readings/papersSubmittedPDF/Elrod.pdf (accessed 22 May 2013).

Farh L, Le SJ (2008). A project-based assessment for introductory bioinformatics course—an assessment aimed to reinforce students' ability in data analyses interpretation and integration. In: Proceedings of the International Conference on BioMedical Engineering and Informatics BMEI 2008, 832–837.

Fraser BJ (2007). Classroom learning environments. In: Handbook of Research on Science Education, ed. SK Abell and NG Lederman, Mahwah, NJ: Lawrence Erlbaum, 103–124.

Furge LL, Stevens-Truss R, Moore DB, Langeland JA (2009). Vertical and horizontal integration of bioinformatics education. Biochem Mol Biol Educ *37*, 26–36.

Gelbart H, Brill G, Yarden A (2009). The impact of a web-based research simulation in bioinformatics on students' understanding of genetics. Res Sci Educ *39*, 725–751.

Haladyna TM (2006). Roles and importance of validity studies in test development. In: Handbook of Test Development, ed. SM Downing and TM Haladyna, Mahwah, NJ: Lawrence Erlbaum, 739–758.

Handelsman J, Miller S, Pfund C (2006). Scientific Teaching, New York: Freeman.

Haury DL, Nehm RH (2012). The global challenges of genomics education: a path to the future. In: Genomics Applications for the Developing World, ed. KE Nelson and B. Jones-Nelson, New York: Springer, 311–333.

Herron SS (2009). From cookbook to collaborative: transforming a university biology laboratory course. Am Biol Teach *71*, 548–552.

Hestenes D, Wells M, Swackhamer G (1992). Force concept inventory. Phys Teach *30*, 141.

Hingamp P, Brochier C, Talla E, Gautheret D, Thieffry D, Herrmann C (2008). Metagenome annotation using a distributed grid of undergraduate students. PLoS Biol *6*, 296.

Howard DR, Miskowski JA, Grunwald SK, Abler ML (2007). Assessment of a bioinformatics across life science curricula initiative. Biochem Mol Biol Educ *35*, 16–23.

Hu W, Adey P (2002). A scientific creativity test for secondary school students. Int J Sci Educ *24*, 389–403.

Kane M (2006). Content-related validity evidence in test development. In: Handbook of Test Development, ed. SM Downing and TM Haladyna, Mahwah, NJ: Lawrence Erlbaum, 131–153.

Krippendorff KH (2003). Content Analysis: An Introduction to Its Methodology, Thousand Oaks, CA: Sage.

Kuldell NH (2006). How golden is silence? Teaching undergraduates the power and limits of RNA interference. CBE Life Sci Educ *5*, 247–254.

Landis JR, Koch GG (1977). The measurement of observer agreement for categorical data. Biometrics *33*, 159–174.

Liu X (2009). Standardized measurement instruments in science education. In: The World of Science Education: Handbook of Research in North America, ed. W-M Roth and K Tobin, Rotterdam, Netherlands: Sense, 649–677.

Liu X (2010). Using and Developing Measurement Instruments in Science Education: A Rasch Modeling Approach, Charlotte, NC: Information Age.

Lopatto D *et al.* (2008). Genomics education partnership. Science *322*, 684–685.

Luo Y, Gong X, Xu L, Li S (2007). Isolation of RNA and RT-PCR, cloning, and sequencing of noncoding RNAs from fungi. Biochem Mol Biol Educ *35*, 355–358.

Magee J, Gordon JI, Whelan A (2001). Bringing the human genome and the revolution in bioinformatics to the medical school classroom: a case report from Washington University School of Medicine. Acad Med *76*, 852.

Malacinski GM, Zell PW (1995). Learning molecular biology means more than memorizing the formula for tryptophan. J Coll Sci Teach *25*, 198–202.

McEwen LA, Harris DIK, Schmid RF, Vogel J, Western T, Harrison P (2009). Evaluation of the redesign of an undergraduate cell biology course. CBE Life Sci Educ *8*, 72–78.

Messick S (1989). Meaning and values in test validation: the science and ethics of assessment. Educ Res *18*, 5–11.

Messick S (1995). Validity of psychological assessment: validation of inferences from persons' responses and performances as scientific inquiry into score meaning. Am Psychol *50*, 741–749.

Messick S (1999). Assessment in Higher Education: Issues of Access, Quality, Student Development, and Public Policy: A Festschrift in Honor of Warren W. Willingham, Mahwah, NJ: Lawrence Erlbaum.

National Research Council (NRC) (2002). Scientific Research in Education. www.nap.edu/openbook.php?isbn=0309082919 (accessed 15 March 2010).

NRC (2009). A New Biology for the 21st Century: Ensuring the United States Leads the Coming Biology Revolution, Washington, DC: National Academies Press.

NRC (2012). Discipline-Based Education Research: Understanding and Improving Learning in Undergraduate Science and Engineering, Washington, DC: National Academies Press.

Nehm RH (2006). Faith-based evolution education? Bioscience *56*, 638–639.

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.

Nitko AJ, Brookhart SM (2010). Educational Assessment of Students, 6th ed., Upper Saddle River, NJ: Pearson Prentice Hall.

Qin H (2009). Teaching computational thinking through bioinformatics to biology students. In: Proceedings of the 40th ACM Technical Symposium on Computer Science Education, SIGCSE 2009, New York: Association for Computing Machinery, 188–191.

Rowland-Goldsmith M (2009). A new way to introduce microarray technology in a lecture/laboratory setting by studying the evolution of this modern technology. Biochem Mol Biol Educ *37*, 37–43.

Shachak A, Ophir R, Rubin E (2005). Applying instructional design theories to bioinformatics education in microarray analysis and primer design workshops. Cell Biol Educ *4*, 199–206.

Shaffer CD *et al.* (2010). The genomics education partnership: successful integration of research into laboratory classes at a diverse group of undergraduate institutions. CBE Life Sci Educ *9*, 55–69.

Shin N, Jonassen DH, McGee S (2003). Predictors of well-structured and ill-structured problem solving in an astronomy simulation. J Res Sci Teach *40*, 6–33.

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.

Wefer SH, Anderson OR (2008). Identification of students' content mastery and cognitive and affective percepts of a bioinformatics miniunit: a case study with recommendations for teacher education. J Sci Teacher Educ *19*, 355–373.

Weisman D (2010). Incorporating a collaborative web-based virtual laboratory in an undergraduate bioinformatics course. Biochem Mol Biol Educ *38*, 4–9.

White B, Kim S, Sherman K, Weber N (2002). Evaluation of molecular visualization software for teaching protein structure: differing outcomes from lecture and lab. Biochem Mol Biol Educ *30*, 130–136.