CBE—Life Sciences Education
Vol. 15, 1–14, Summer 2016

*Research Methods*

# A Primer for Developing Measures of Science Content Knowledge for Small-Scale Research and Instructional Use

**Kristin M. Bass,*† Dina Drits-Esser,‡ and Louisa A. Stark‡**

†Rockman et al, San Francisco, CA 94105; ‡Genetic Science Learning Center, University of Utah, Salt Lake City, UT 84102

Submitted July 5, 2015; Revised February 3, 2016; Accepted February 4, 2016
Monitoring Editor: Ross Nehm

The credibility of conclusions made about the effectiveness of educational interventions depends greatly on the quality of the assessments used to measure learning gains. This essay, intended for faculty involved in small-scale projects, courses, or educational research, provides a step-by-step guide to the process of developing, scoring, and validating high-quality content knowledge assessments. We illustrate our discussion with examples from our assessments of high school students' understanding of concepts in cell biology and epigenetics. Throughout, we emphasize the iterative nature of the development process, the importance of creating instruments aligned to the learning goals of an intervention or curricula, and the importance of collaborating with other content and measurement specialists along the way.

## INTRODUCTION

The quality of a scientific research project often depends on the availability of appropriately sensitive instruments. Advances in measurement, from nanoparticle biosensors (Howes *et al*., 2014) to disease diagnostics (Vogel, 2014) to digital PCR (Perkel, 2014), are critical for discoveries in basic and applied biological research. High-quality instruments are equally important in biology education. A well-conceived instrument that measures content knowledge is essential for making claims about the effectiveness of a new educational intervention and can provide data for curriculum or program improvement. It can be quite challenging, however, to identify instruments that are suitable for an intervention, to adapt existing items, or to create new ones altogether. Assessing the validity and reliability of these tools adds a level of complexity to the mix, especially within the budget constraints of small-scale projects.

Further, while science, technology, engineering, and mathematics faculty members are most familiar with assessment-related terms such as "student learning outcomes" and "summative assessment," they are much less knowledgeable about issues related to the validity of interpretations of tests, surveys, or assessment items (Hanauer and Bauerle, 2015).[1] They consequently write and use assessment items in courses they teach without undertaking a development process to determine how well their items measure the intended content. This approach may or may not be adequate for assigning student grades *and* is insufficient for studies that seek to determine the efficacy of new interventions. In fact, a recent analysis of evaluations of educational innovations in genetics and bioinformatics found that less than 10% of published articles contained any information about the reliability or validity of study instruments (Campbell and Nehm, 2013). Challenges may also arise when using published instruments or questions from item banks to evaluate learning. In these cases, the items may not be tightly aligned with the learning goals and objectives of courses, curricula, or other educational interventions. Further, if many of the

[1]We define and discuss many common assessment terms, including reliability, validity, and analytical versus holistic scoring. Readers less familiar with core terminology may also wish to consult Rudner (1994) and the glossary in the *Standards for Educational and Psychological Testing* (AERA *et al.*, 2014).

CBE Life Sci Educ June 1, 2016 15:rm2
DOI:10.1187/cbe.15-07-0142
*Address correspondence to: Kristin M. Bass (kristin@rockman.com).

items on an instrument are too easy or too difficult for the majority of study participants, little can be learned about the intervention's efficacy.

This essay is an introduction to developing content knowledge assessments that will be used on a relatively small scale, as opposed to larger-scale assessment programs intended to reach a large audience (e.g., the National Assessment of Educational Progress [NAEP] or the Assessment of Higher Education Learning Outcomes) and that therefore require more advanced psychometric analyses. We, the authors, are two educational psychologists (K.B., D.D.-E.) and an evolutionary biologist (L.S.), all with extensive experience in program evaluation. The first two authors of the paper have worked closely together as external and internal evaluators for the University of Utah's Genetic Science Learning Center (GSLC) to carry out research studies on GSLC curricula and rigorous project evaluations on GSLC programs. We have geared this paper toward faculty who are involved in creating and evaluating curricula or other educational interventions (either for their own postsecondary courses or for a K–12 audience) and individuals who work with external program evaluators and wish to increase their understanding of instrumentation.

We first provide a brief overview of the process of instrument construction. We follow with a more detailed discussion of each step in the process, illustrated by examples from our assessments of high school students' knowledge of cell biology and epigenetics (Drits-Esser *et al*., 2014). In particular, we emphasize how to create instruments aligned to the learning goals of an intervention or curricula and how to determine whether the items are valid for assessing the content they are intended to measure and in the contexts in which they will be used.

## A GENERAL INSTRUMENT-DEVELOPMENT PROCESS

Measurement, whether in biology or education, involves making quantifiable inferences from observable evidence. Gel electrophoresis, for example, determines the relative length of DNA strands. It is not possible to see the strands directly; rather, biologists infer the length of strands based on the distance they migrate in the gel, as visualized by staining. Likewise, educators cannot see knowledge growth at any kind of direct, neurological level. Instead, they have to draw conclusions about what individuals know and can do through their performance on observable tasks. Crafting those tasks or items is both an art and a science, requiring detailed knowledge of the content being assessed, the population being measured, the contexts in which the instrument will be used, and the range of items or tasks that might appropriately elicit the content being tested.

A 2001 National Research Council (NRC) report entitled *Knowing What Students Know* and several reports and journals that followed (NRC, 2006, 2014; Songer and Ruiz-Primo, 2012) represent assessment design as a triangle with three vertices: cognition, observation, and interpretation (Figure 1). The cognition vertex refers to research on how students learn a topic and enables researchers or instructors to identify the precise "targets of inference" (NRC, 2001, p. 62) they wish to measure. The observation component of the

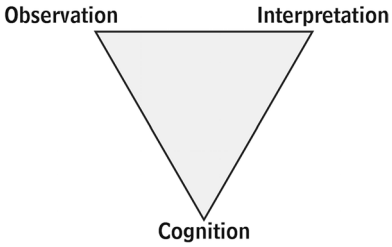


**Figure 1.** The assessment triangle (NRC, 2001, p. 44). Reprinted with permission from the National Academies Press, Copyright 2001, National Academy of Sciences.

triangle describes the tasks designed to draw out students' knowledge and skills and differentiate students based on their levels of understanding and ability. The interpretation vertex refers to the reasoning and analytical tools used to make inferences about latent cognition from the gathered observations. This process utilizes statistical models with large-scale assessments but often has a qualitative component in smaller-scale settings (NRC, 2001, 2014).

The strength of the assessment triangle is that it concretely yet simply illustrates the notion of assessment as a way of reasoning from evidence (NRC, 2001, 2014). The alignment of research on student learning with observations of performance is made possible by interpretive tools and frameworks. In this model, assessment is not just referring to test items or formats, but denotes a process of making quality inferences about student ability using data from a carefully constructed sample of items or tasks (Campbell and Nehm, 2013). It requires researchers or instructors to think explicitly about how they will use the data from the items they have written or selected to draw logical conclusions about what students have learned from a course or intervention.

The coordination of the triangle's three elements is part of what makes developing assessments so difficult, especially in topic areas for which models of learning have not yet been fully specified. A special issue of the *Journal of Research in Science Teaching* contains several case studies of how researchers have attended to the cognition, observation, and inference vertices in their construction of instruments (Songer and Ruiz-Primo, 2012). We believe that it is also possible for novice developers to apply the assessment triangle to their work and describe a process for doing this.

Designing an instrument to measure content knowledge involves four basic steps: 1) identifying the concepts to be measured (also known as construct identification), 2) selecting or writing assessment items, 3) creating a scoring system, and 4) reviewing and validating items. The first step draws developers' attention to the cognition component of the assessment triangle, while the second step focuses on methods of observing student knowledge and skills. The third and fourth steps address the interpretation of student performance, or the match between the observed evidence and the instrument's intent. This multistep, construct-centered process is not entirely linear. Instrument developers frequently iterate between steps, for example, using the results of their validation studies to redesign their items or scoring, or using research on the constructs being assessed to inform the interpretation of student work (NRC, 2001). Throughout these steps, developers also need to attend to the connections between cognitive constructs, student observations, and

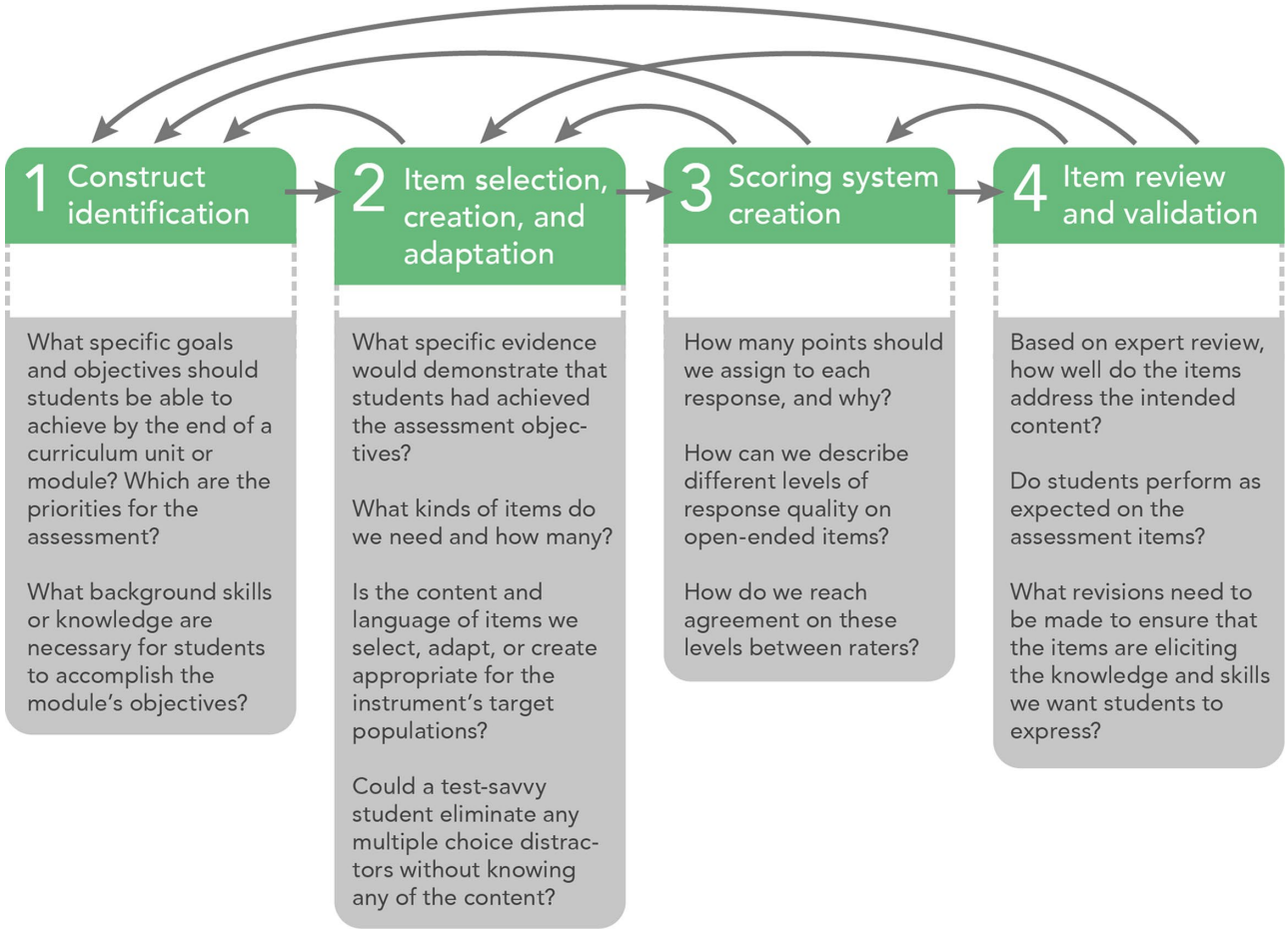


**Figure 2.** Overview of an instrument-development process, including guiding questions for each step. These lists contain examples of the kinds of questions that could be asked and are not meant to be inclusive.

the interpretive methods that will enable them to make appropriate claims with their instruments. Figure 2 illustrates these four steps and lists some questions to consider during the development process.

As we expand on each of the four steps, we provide examples from our process of developing pre and post content knowledge assessments to measure high school biology students' learning gains from two curriculum supplements developed by the GSLC. We used these instruments in small-scale randomized controlled trials that compared the GSLC-developed supplements with other materials that addressed the same learning goals (Drits-Esser *et al.*, 2014). The GSLC develops curriculum supplements that are freely available on its Learn.Genetics and Teach.Genetics websites. The materials include 1) interactive, multimedia learning experiences; 2) short movies; 3) three-dimensional animations; 4) "Learn More" web pages; 5) hands-on models; 6) paper-based learning activities; and 7) graphic organizers for students to use with the online materials. Each supplement addresses several broad learning goals with each learning experience typically focusing on one learning objective. Our primer discusses content measures for the *Amazing Cells* (GSLC, 2008a,b) and *Epigenetics* (GSLC, 2009a,b) curriculum supplements.

### *Step 1. Construct Identification*

It is common to construct content knowledge assessments by reviewing the curriculum or program and diving straight into writing items that match the content of the individual learning activities. That approach misses a critical step. Reviewing a curriculum's overall goals and objectives as well as applicable research on student learning first allows you to design a more cohesive set of questions that address the most important, enduring ideas.

A learning goal is an outcome, broadly stated, that students are expected to accomplish by the end of a unit, module, or intervention. An objective describes the specific knowledge and skills needed to achieve that goal. Objectives are commonly, though not always, phrased as action statements (e.g., explain, interpret, apply; Wiggins and McTighe, 2005). While there can be many objectives for a given goal, some may be more important to know than others. Useful resources for narrowing down which concepts to measure include curriculum developers, content experts, and state and national K–12 science frameworks and standards (e.g., College Board Standards for College Success [College Board, 2009], the Framework for K–12 Science Education [NRC, 2012], the Next Generation Science Standards [NGSS Lead States, 2013], the NAEP Science Framework [National

Assessment Governing Board, 2015]), and undergraduate biology concept inventories (Garvin-Doxas *et al.*, 2007).

In addition, before beginning a new unit or content area, it can be useful to assess the concepts students should already know and/or the misconceptions that may interfere with their learning (Sadler *et al.*, 2012; NRC, 2014). Along with items about the content students are about to learn, this information provides a baseline from which to measure progress. If students do not achieve the intended learning objectives, these data can provide insights on why this may be the case.

As we planned our *Amazing Cells* and *Epigenetics* assessments, we worked with the GSLC curriculum developers to prioritize the learning objectives for the supplements. For example, the overarching goals of the *Epigenetics* supplement were for students to understand 1) what epigenetics is and 2) the relationship between epigenetics and the environment. There were six to eight more specific, measurable learning objectives within each goal. We asked the lead curriculum developer to differentiate the main ideas of the supplement from the ones that were simply "nice to know." Within the two goals, she identified four key objectives for us to measure:

1. The epigenome influences gene expression.
2. Signals from the environment influence gene expression by acting on the epigenome. The epigenome helps cells "remember" the sum total of signals they have received that affect gene expression.
3. Epigenetics can lead to differences in genetically identical twins.
4. Factors from the environment such as diet, physical activity, and stress influence the epigenome.

It is often helpful to begin at the end and map backward from the ultimate goal of what is to be learned to students' prior knowledge. As we articulated the objectives we would measure, we therefore considered the knowledge students had or would need to understand the epigenetics materials. High school students commonly assume that "a gene is a trait or that DNA produces proteins" (Elrod, 2007, p. 2). Students who approached the supplement materials with such misconceptions might have difficulty grasping the subsequent content. On the other hand, the materials might clear up some of those misconceptions, so we wanted to make sure to capture any potential learning gains. We consequently decided to assess students' understanding of the relationship between DNA, genes, and proteins in addition to the supplement's main objectives.

After narrowing down the objectives you want to measure (or even during this process), it is important to identify the kinds of observable evidence you will collect to demonstrate that the objectives have been met. This involves creating or selecting test items.

### *Step 2. Item Parameters, Selection, and Creation*

***Item Parameters.*** There are many ways to assess a given objective. We will discuss two of the most common item formats in content assessments: multiple-choice (MC) and open-ended items (also known as constructed responses, such as short answer, essays, and fill-in-the-blank). These formats have several inherent advantages and drawbacks. For example, MC items are straightforward to score and employ a response format that is quite familiar to most students (Simkin and Kuechler, 2005). However, while MC items can be constructed to access a range of levels of students' cognition, they also have certain limitations, including the difficulty and time investment of developing quality items, potential false indication of students' knowledge and understanding, and potential demographic biases associated with performance (e.g., gender, ethnicity, socioeconomic status). Further, these questions may limit the potential to assess students' ability to think creatively and to organize and synthesize information (Simkin and Kuechler, 2005) and may perpetuate the idea that scientific investigation has only one correct answer. Open-ended items provide for more nuance and variability in student responses, which can access divergence in students' thought processes, and are more likely to capture conceptual learning (Martinez, 1999). However, grading may be time intensive, subjective, and highly sensitive to variation across raters (Simkin and Kuechler, 2005; Stanger-Hall, 2012).

Other options for measuring science knowledge include performance-based assessments (i.e., hands-on activities requiring students to conduct all or part of a science investigation to demonstrate their knowledge and skills [NRC, 2014]), problem scenarios posed within educational computer games (Hickey *et al.*, 2009), or analysis of students' science laboratory notebooks (Baxter *et al.*, 2001; Wallert and Provost, 2014). In some cases, these other types of assessments are more difficult to design and to score; however, they may access students' cognition in ways that MC or short-essay items cannot. We nevertheless focus on these two latter question formats in this article, because they are most familiar to faculty and students and therefore appropriate for this primer.

Before we began drafting instruments for the *Amazing Cells* and *Epigenetics* supplements, we first had to set some parameters for the number of items we needed. We had to be considerate of the amount of time teachers had to administer the instrument, the length of the intervention, and the age of the participating students. Further, we decided that we generally wanted at least three items per objective; three is the minimum number of replications recommended in scientific measurement, and we felt the same standards were suitable here.

We ultimately decided that the instruments should take students approximately 15–20 min to complete, given that we were evaluating high school students' performance on a two-day intervention. Assuming approximately 1 min per MC item and at least 5–7 min for essays, we planned for roughly 12–18 items divided equally among assessment objectives. We also knew that, to arrive at our target number of items, we would have to come up with at least twice as many at the beginning of our process; experience had shown that we would probably eliminate half of our items during internal and external review processes (for instance, because an item did not address the most important content in the curriculum supplement).

It is also important to consider the relative difficulty of the items you wish to develop. Asking questions that require a range of thinking skills will help you better evaluate where students fall on a continuum of learning for each objective. For example, the Trends in International Math and Science Study (TIMMS; Mullis *et al.*, 2009), a survey of students in the elementary and secondary grades, classifies items into

**Table 1.** Examples of publicly available instrument databases with life science items

| | |
|---|---|
| K–12 | American Association for the Advancement of Science (AAAS) Project 2061 Science Assessment: http://assessment.aaas.org |
| | Misconceptions-oriented Standards-based Resources for Teachers (MOSART): www.cfa.harvard.edu/smgphp/mosart/testinventory_2.html |
| | National Assessment of Educational Progress (NAEP): http://nces.ed.gov/nationsreportcard/itmrlsx |
| | Trends in International Mathematics and Science Study (TIMSS): www.timss.org |
| Higher education | American Society of Microbiology's list of concept inventories: www.facultyprograms.org/index.php/resources/concept-inventories |
| | Conceptual Inventories in Biology: http://saber-biologyeducationresearch.wikispaces.com/Concept±Assessments-Biology |

three cognitive domains: knowing (recall and vocabulary), applying (making connections between concepts), and reasoning (using content and process knowledge to solve problems and construct explanations). You may wish to include similar kinds of lower- and higher-order thinking items in your content knowledge instruments. Keep in mind that item format is not necessarily related to cognitive domain, nor is item difficulty necessarily related to either item format or cognitive domain. MC questions can assess reasoning as well as recall, while essays and performance-based assessments may inadvertently elicit knowledge of facts instead of their application (Baxter and Glaser, 1998; NRC, 2014).

***Item Selection and Adaptation.*** In selecting our items, we first reviewed existing instrument databases, including the ones listed in Table 1. We also looked at the Genetics Literacy Assessment Instrument (GLAI; Bowling *et al.*, 2008), a concept inventory intended for undergraduate non–science majors, which we thought might be suitable for the *Epigenetics* supplement. Several of the items addressed the objectives we intended to measure, but the reading level was too difficult for high school students. We knew that, by simplifying some of the language, we would make the items more valid for measuring our particular population. We also recognized that, if we altered the items, we could not claim that our instrument had the same reliability and validity characteristics of the GLAI as published, nor would we be able to compare the performance of our students with that of students in other studies conducted with that instrument. Because the purpose of our instrument was to evaluate the effectiveness of the GSLC's curricular materials, we decided that suitability for the student population being assessed outweighed any interest in generalizing to other groups.

We adjusted the items mainly by shortening sentences or phrases. Table 2 displays one of the original GLAI items and our adaptation. We simplified the phrase "expression of his or her genes" to "gene expression," but expanded the phrase "lasting until adulthood" to "stopping when a person reaches adulthood." We added italics to focus attention on the key differences between answer choices. Additionally, we looked for any terms that high school students might not understand and that would not be addressed in the *Epigenetics* materials. For instance, in this item, we removed the term "menopause," since we thought it might not be familiar to some high school students, and it was not essential to understanding the concept we intended to test. We also divided the question into two parts, the first of which required a *yes* or *no* answer. Our desire to retain the integrity of the original GLAI item inadvertently made the single *no* option much more conspicuous. It is generally advisable to have even numbers of options with a particular stem, so test-takers cannot easily eliminate a choice. We will discuss this issue in further detail in the next section. Nevertheless, we have chosen to share this item to highlight some of the trade-offs encountered with revising items and the need for as many iterations of development, discussion, and revision as time will allow.

***Item Creation.*** We often could not find enough existing items for an objective and had to create our own. In these cases, we looked to the curriculum supplements for inspiration and drafted items in accordance with recommended item-writing guidelines, such as keeping the text simple to minimize the time needed to read items, using positive—not negative—phrasing, and avoiding options such as "I don't know" (Haladyna *et al.*, 2002; American Association for the

**Table 2.** Adaptation of a Genetics Literacy Assessment Instrument item

| Original | Modified |
|---|---|
| At what times during an individual's life does the environment influence the expression of his or her genes? | Can the environment influence gene expression? If so, during which times in an individual's life? |
| A. Beginning at conception and lasting throughout life. | A. Yes, beginning at *conception* and lasting throughout life. |
| B. Beginning at birth and lasting throughout life. | B. Yes, beginning at *birth* and lasting throughout life. |
| C. Beginning at birth and lasting until adulthood. | C. Yes, beginning at birth and stopping when a person reaches adulthood. |
| D. Occurring only during key stages of life such as puberty and menopause. | D. Yes, but only during key stages of life such as puberty. |
| E. Environment has little or no effect on how genes are expressed. | E. No, the environment has little or no effect on gene expression. |
| Correct answer: "A." | |

**Table 3.** Tips for writing MC questions[a]

**General**

- Test for important or significant information (base each question on student learning objective of the lesson, not trivial information).
- Be sure the item would be comprehensible to your students.
  - Avoid unfamiliar vocabulary that is not defined and that is not related to the learning goal.
  - Avoid complex sentences.
  - Avoid words and phrases with confusing or ambiguous meanings.
- Items should have only one right answer.
- Use present tense and active voice.
- Minimize the time required to read each question.

**Stem**

- Include the central idea to avoid repetition in answer choices.
- Keep sentences brief and straightforward with a simple phrase structure and no additional clauses.
- Word positively—avoid negative phrasing.
- Avoid phrasing "all of the following except" or "which of the following is false."

**Answer choices**

- Link one or more of the distractors to misconceptions related to the key idea.
- Each answer choice should be a single word or phrase or a single sentence (*keep options short*).
- Keep all options homogeneous in content.
- Keep answer choice length similar.
- Avoid "all of the above."
- Avoid "none of the above."
- Avoid "I don't know."
- Include from three to five options for each question.
- Keep options independent; options should not be overlapping.
- Phrase options positively, not negatively.
- Avoid distractors that can clue test-wise examinees (e.g., absurd options, formal prompts, or overly specific or overly general clues).
- Avoid giving clues through the use of faulty grammatical construction.

**Illustrations**

- Keep illustrations simple and to the point.
- Illustrations should facilitate the understanding of what is being asked.
- Include the same information in the text and the illustration.

[a] Compiled from sources through AAAS Project 2061 (2011) and adapted with permission. The Supplemental Material contains an expanded list of guidelines.

Advancement of Science [AAAS] Project 2061, 2011). It is also important to include vocabulary that students will have learned in class but to try to avoid jargon that students may not understand. This is especially critical when considering the diverse cultural backgrounds or levels of English proficiency of the students who may be taking the assessments. See Table 3 for examples of item-writing guidelines. In addition, the Supplemental Material contains a more extensive list that the AAAS has compiled.

In the *Amazing Cells* supplement, we wanted to assess the objective "Cells respond differently to signals depending on cell and signal type." In one of the Interactive Explore activities designed to teach this objective, students drag icons representing cell signals to different kinds of cells (e.g., photoreceptor, skin cancer cell) and receive feedback on each cell's reaction. We constructed a question that resembled this activity, albeit with a different cell signal and type than had appeared in the supplement (Figure 3).

This item illustrates a few common MC-writing guidelines. We defined unfamiliar vocabulary, stating explicitly that cytokine was a type of cell signal. We also kept the answer choices approximately the same length and made the content and grammatical structure as homogeneous as possible. In addition, we had pairs of choices with similar wording: two stating that the liver cell would respond in the same way to something, and two stating that the liver cell would respond in a different way. If we had an uneven number of

Cytokine is a type of cell signal. How might a liver cell respond to cytokine?

**A.** Probably in different ways depending on how far the cytokine travels.

**B.** Probably in the same way that a blood cell will respond to cytokine.

**C.** Probably in a different way than a blood cell will respond to cytokine.

**D.** Probably in the same way the liver cell would respond to a nitric oxide cell signal.

**Figure 3.** Sample assessment item written for the *Amazing Cells* supplement (correct answer: "C").

**Table 4.** Partially correct answers to "What does the epigenome do?," from initial pilot test

| Rubric category | Examples from students' responses |
|---|---|
| The epigenome is involved in gene expression. | The genome is the full genetic information of a human. The epigenome is what tells those genes how to be expressed. |
| The epigenome reacts to the environment. | The epigenome is sitting above the genome and changes chemical signals according to environmental factors. |
| The epigenome influences traits. | The epigenome decides what traits you get. |
| The epigenome reacts to signals. | The epigenome arranges in response to signals. |

"same way" and "different way" options, it might have cued test-savvy students to eliminate the choice that did not resemble the others.

A complementary aspect of designing items is determining how to quantify students' responses. This is discussed next.

### *Step 3. Item-Scoring Systems*

Different from a simple answer key, a scoring system establishes clear, consistent criteria for the number of points to be assigned to responses of varying quality. This section describes some strategies for creating rubrics for open-ended written responses. Keep in mind, however, that it is possible to give full or partial credit to any type of item, including MC questions (Briggs *et al.*, 2006; Hadenfeldt *et al.*, 2013).

Rubric development should happen concurrently with item construction. When drafting an open-ended item, you should establish criteria for a complete and correct answer, and speculate on the kinds of responses that would demonstrate partial understanding of the targeted idea. The levels of a rubric, or the scores that are assigned, can be refined by reviewing and categorizing actual student work. This "top-down, bottom-up" process (Chi, 1997) ensures that the final rubric accounts for a priori expectations for what an item should measure while being sensitive to the realities of student performance.

Designers must make a number of decisions to produce a quality rubric. One choice is establishing the number of criteria upon which to rate responses. It may be possible to score a response using a holistic rubric that evaluates the overall quality of an answer. An analytical rubric, which rates performance on several components (e.g., the quality of claim, evidence and reasoning in a scientific argument; McNeill and Krajcik, 2011), requires more time to develop and use but permits greater precision in measuring knowledge and skills.

A related rubric design issue is identifying the most important content for a response that will receive high scores (as opposed to ideas that would be useful for a student to include, but not necessary; Arter and McTighe, 2001). It can therefore be beneficial to revisit the specific goals and objectives you wish to assess and ask how well each detail in a rubric addresses those objectives. Rubric creators must also justify the inclusion of quantitative and qualitative criteria for judging performance. The number of examples or pieces of evidence in a response may not be as important as the quality of those examples. It is important to differentiate what can be counted from what actually "counts."

Finally, a rubric should be so clear and comprehensive that all raters can use it to agree on scores. The best way to evaluate this is to allocate enough time in the assessment-development process to train raters and establish interrater reliability or consistency.[2] It is usually easier to get good interrater reliability using analytical rubrics than holistic ones, so that is another factor to take into consideration when deciding which type to use. When pilot testing a rubric for short open-ended responses, we recommend that raters score a random sample of 10–20 answers and then compare scores. You may choose to establish exact agreement or allow some degree of variation depending on the number of categories in the rubric, the time and resources available for scoring, and the magnitude of the consequences of the scoring decisions (e.g., high-stakes college admissions or teacher promotion; Stemler and Tsai, 2008). There are also different methods for calculating interrater reliability, ranging from percentage of agreement (the most intuitive to interpret) to Pearson's $r$ correlations, Cohen's kappa, and intraclass correlation coefficients (ICCs). Interested readers can consult Stemler and Tsai (2008) and Hallgren (2012) for formulas, along with SPSS and R syntax for these calculations. Rather than explaining the merits and limitations of each of these statistics, we wish to emphasize the general value of using multiple raters to inform instrument design. Discussions about scoring disagreements can lead to changes in rater training or a clarification of rubric criteria, as we will illustrate in an example from the *Epigenetics* assessment.

As part of developing holistic rubrics for open-ended *Epigenetics* items, we conducted a pilot test of the entire assessment with 83 high school students who had just completed the GSLC *Epigenetics* supplement. This allowed us to determine whether the items were eliciting the answers we expected (a point on which we will elaborate in the next section), and helped us generate initial ideas for the scoring rubric.

One of the questions we piloted was, "What does the epigenome do?," which, while broad, related directly to the goals and objectives of the supplement. We expected that students would be able to explain that the epigenome controls gene expression by turning genes on and off. We assumed that we would have at least two categories of partially correct responses but were not sure what the criteria would be for those levels. As we analyzed the pilot data, we determined that students' partially correct answers generally fell into one of four themes or categories (Table 4).

[2]Reliability refers to the ability of an instrument to produce the same results in different situations. It can be assessed with different types of data, each related to a particular source of error or inconsistency in measurement. In this paper, we discuss strategies for establishing agreement between raters. Reliability can also refer to the consistency of items to measure the same underlying idea (internal consistency) or of an instrument to generate similar responses over time (test–retest reliability) (Cook and Beckman, 2006; Lovelace and Brickman, 2013). A consideration of these issues is beyond the scope of this article.

In internal memos, we asked ourselves, "Are all of the partially correct categories the same quality, or are some more correct than others?" We decided that two of the categories demonstrated a more accurate, albeit still incomplete, understanding of the supplement's objectives than the other two categories. We therefore divided the partially correct answers into two levels. Level 1 consisted of answers about the epigenome influencing traits and reacting to signals, and Level 2 contained answers mentioning that the epigenome is involved in gene expression or reacts to the environment. For the curriculum field test, we developed a four-level rubric giving three points to responses that completely and correctly answered the question, one or two points for answers that met the partially correct criteria we had established, and zero points to answers with significant misconceptions.

We continued to adjust the partially correct rubric categories through conversations about our ratings. For instance, we initially disagreed on how to rate responses that indicated only that "the epigenome controls the genome." One rater wanted to give these answers a zero, because they repeated parts of the question, while the other rater thought the answers deserved a one, because they were more accurate than the other answers she had scored as zeros. We ultimately accepted the latter rater's argument and modified the rubric accordingly. Once we felt confident about our rubric, we assigned one person to score all of the responses to that question from our field test and another to score a random 25% sample. We then calculated ICCs to evaluate interrater reliability (Drits-Esser *et al.*, 2014).

For the *Epigenetics* assessment, we codeveloped our rubrics jointly and then evaluated interrater reliability. Alternatively, a researcher may construct a rubric independently and train another individual to serve as a second rater. Rubric clarity is particularly important in these cases, since the rater may be less familiar with the assessment context and the data at hand. An effective rubric should contain definitions of each level that specify the depth of understanding and even the terminology a student should provide to receive credit for that level (Allen and Tanner, 2006). Rubrics should also contain two or more examples per level to reinforce the definitions (Arter and McTighe, 2001).

Even the most detailed rubrics may not be able to accommodate all possible responses. Borderline cases are inevitable. We recommend reviewing some ambiguous answers during rater training and justifying why each response should fall into one level or another. Such conversations may cause you to further clarify the difference between levels (Moskal and Leydens, 2000). If there is a sufficient number of borderline answers with common characteristics (e.g., 10% of your sample or more), you may wish to create a new rubric level entirely.

Finally, keep in mind that while we have been discussing interrater reliability to this point, intrarater reliability is equally important. It is good practice to rescore a small sample (10%) of your own ratings to ensure you have been consistent all the way through. It is common to experience "rater drift" and become more lenient or stringent over time; rescoring some answers or comparing the first answers you scored with the last helps you determine whether you need to adjust some of your ratings.

Once you have assessment items, it is time to validate them. This process utilizes several lines of evidence to build a credible argument about the appropriateness of your items for the context in which you will use them.

### *Step 4. Item Review and Validation*

Validity is defined by the Standards for Educational and Psychological Testing as "the degree to which accumulated evidence and theory support a specific interpretation of test scores for a given use of a test" (American Educational Research Association [AERA] *et al.*, 2014, p. 225). In other words, how well are you measuring what you intended to measure? To what extent can you justifiably use instrument scores to infer what students know at a particular time point or evaluate the efficacy of an intervention?

Validity is not a property of a test but of its interpretation. While many granting agencies require the use of "valid and reliable" instruments, that request is a bit misleading. An instrument may be valid for one context but not another, much like a drug may be indicated for one condition or population but lacks evidence from clinical trials to support other uses. As noted earlier, we commonly encounter this issue when we use or adapt items from concept inventories or other instruments that have been validated for college undergraduates but not younger students.

It would be convenient to think about validity as a single quantifiable indicator, much like information about reliability is reported as correlations or alpha coefficients. However, validity is not a numerical rating per se, but rather an argument that can be supported with different kinds of evidence (Kane, 2013). It is helpful to conceptualize validity as a way to empirically test the claims you are making about your instrument. Think back to your original plans and remind yourself of what you wanted to measure and why. Then collect the most feasible and convincing evidence you can (within your time and budget constraints) to support your interpretations of student knowledge, based on the data you have gathered from your items. In this manner, validation brings the instrument-development process full circle.

If validation is represented as an argumentation process, the question then becomes what sources of evidence might be appropriate to justify various claims. The Standards for Educational and Psychological Testing (AERA *et al.*, 2014) identify five sources: 1) test content, 2) response processes, 3) internal structure, 4) relationships to other variables (a.k.a. external structure), and 5) consequences. Note that it is not necessary to collect data from all five categories. Rather, the amount or type of evidence you need depends on the stakes attached to the instrument and the breadth of its use, as well as the availability of relevant data. We will discuss two types of validity evidence: *test content*, which considers how well items represent a concept or domain; and *response processes*, which examine the reasoning students apply to test answers. We determined that these sources were the most important for our early-stage pilot work, since they could directly inform item revisions.

It is also possible to validate assessments using evidence based on *internal structure* (i.e., the associations between test items, as they relate to the measure's intended constructs), *relationships with other variables* (i.e., the relationship between students' scores on the assessment you have developed to performance on established measures of similar or different constructs), and *consequences* (i.e., the effect of test scores on

positive social outcomes such as improvements in teaching practice or negative outcomes such as cheating). Readers are encouraged to consult Campbell and Nehm (2013), Reeves and Marbach-Ad (2016), and the Testing Standards for a more extended discussion of validity and suggested methodologies for gathering each type of evidence.

***Validation Evidence Based on Test Content.*** The wording, content coverage, and format of a test and the conditions for its administration and scoring are all considered elements of test content (AERA *et al.*, 2014). These should be reviewed internally and externally throughout the assessment design process. We routinely sent drafts of our *Amazing Cells* and *Epigenetics* items to the science content specialists and educators at the GSLC who had developed the two curriculum supplements. For instance, we sent a memo listing several questions we wanted the *Amazing Cells* developers to consider in their review:

- Do the items accurately represent the content covered in the *Amazing Cells* learning objectives? Are there any items in which the content is incorrect or the answer choices are confusing?
- Do the items represent the cognitive skills required for the unit (e.g., memorization, application, synthesis)? Do any of the items seem to be too easy or too hard?
- How well do you think students will be able to answer these questions—will they understand this metaphor or that vocabulary word? Is this question too long, or that one too short?

The reviewers responded with comments about the accuracy of the content and clarifications on the supplement's goals. One of the objectives for *Amazing Cells* was that "cells communicate by sending and receiving signals." One of the educators noted that her "intent was to depict signals as being diverse" and that most of the questions we had designed "aim to narrowly define signals as only being chemical." We consequently constructed more items that addressed the general function of cell signals (i.e., to send messages inside the body) rather than representing signals as one specific form or another.

While we determined that the supplement developers would be the most appropriate individuals to review our items, faculty and others with expertise in a particular field are an obvious resource for external review (Kalas *et al.*, 2013; Perez *et al.*, 2013; Deane *et al.*, 2014). Perez *et al.* (2013), for example, asked reviewers, "Is the correct answer accurate given the scenario?" and "Do any of the other answers strike you as correct?" (p. 671). K–12 teachers or science specialists may also be recruited to evaluate the appropriateness of item content and wording for the students they teach (Fives *et al.*, 2014). They may point out, for example, vocabulary that their students may not understand or scenarios with which students may not be familiar. It is also a good idea to get students' perspectives on items directly, as we will explain in the next section.

***Validation Evidence Based on Response Processes.*** Working directly with the intended test takers helps assessment designers understand whether items actually require the content knowledge and reasoning skills they are supposed to elicit. Students may be able to answer an MC item correctly or eliminate distractors because of their familiarity with general test-taking strategies but not the content being tested. On the other hand, students may know the content but answer a question incorrectly because they apply additional information to their response.

A good example of the latter point comes from an evaluation of an NAEP hands-on performance assessment (Bass *et al.*, 2002). In that study, interviewers prompted fourth-grade students to talk about what they were thinking as they conducted an experiment on sinking and floating. Students observed how high a pencil floated in freshwater and a salt solution, then repeated the procedure to identify the composition of an unknown "mystery water." At the end of the activity, students answered a question designed to test their ability to apply what they had observed: "When people are swimming, is it easier for them to stay afloat in the ocean or in a freshwater lake? Explain your answer" (O'Sullivan *et al.*, 1997, p. 45). Much to the interviewers' surprise, some students—who had stated moments before that a pencil floats higher in saltwater than fresh—said that it was easier to float in a lake. Their rationale? The ocean has waves and sharks, not to mention the fact that salt water stings when it gets in your nose or eyes. It can be awfully hard to float in such challenging conditions. Researchers interpreted these responses to mean that even though students had the knowledge they needed to answer the question, they were considering other factors in their response besides the concentration of salt in a body of water (Bass *et al.*, 2002). While this example is humorous, it illustrates the serious point that test takers can select answer items for very different reasons than you might expect. It is always a good idea to pilot test your assessments to uncover any unintended interpretations.

In our projects with the GSLC, we used two different strategies for examining students' response processes. For *Amazing Cells*, we used a technique called a cognitive interview with three 10th-grade students who had not studied cell biology and one 11th-grade student who had studied cell biology. In a cognitive interview, respondents are asked to share what they are thinking as they answer a test item. They may also be asked follow-up questions about their answers (Hamilton *et al.*, 1997). We suggest intentionally selecting students from diverse backgrounds and differing achievement levels for cognitive interviews, since this will provide useful information in adjusting items and removing jargon. Further, all students should have at least some knowledge of the content being tested. This will allow you to understand the types of responses that the items will evoke for a typical test taker in the population of interest.

We identified 11 MC questions that we had drafted for which we wanted feedback. Many of these items contained some scientific terminology that was not covered in the supplement or used metaphors for cell structure and function that we wanted to vet. We used a standardized cognitive interview protocol (Horizon Research, 2009), which asked students why they chose a particular answer and what they thought of the other answer choices. We followed up with questions about whether any of the words or visuals were confusing or might confuse other students. The latter question is especially valuable, because it enables students to "save face" if they do not know something.

We eliminated or edited several questions based on our interviews. We found that a student answered one question

**Table 5.** Revision of a cell communication item based on cognitive interview results (change in *italics*)

| Original | Modified |
|---|---|
| Diseases such as diabetes and multiple sclerosis occur when there is a breakdown in:<br>A. cell communication<br>B. cell motion<br>C. cell division<br>D. cell differentiation | Diseases such as diabetes and multiple sclerosis occur when there is a breakdown in:<br>A. cell communication<br>B. cell motion<br>C. cell division<br>D. *cell growth* |
| Correct answer: "A." | |

about cell communication disruption (Table 5) correctly because she had recognized a pattern: "most of the questions so far have been on cell communication," leading her to the decision that diseases such as diabetes and multiple sclerosis must be caused by breakdowns in cell communication. We chose to retain this question for the final assessment, but placed it at the beginning of the test so that students could not use information from the other items to inform their response. We also changed one of the distractors from "cell differentiation" to "cell division," because some students were not familiar with the former term and eliminated that choice simply because they did not know what it was.

The second strategy we used to examine students' responses to the items was classroom-level pilot testing, in which students completed a full-length test and provided feedback about items they did not understand. Compared with cognitive interviews, which we conducted with students individually, classroom-level testing allowed us to obtain information from a larger number of students in a relatively short amount of time. In planning our pilot tests, we had to make several decisions about the type of information we wanted to extract from the data. We had to determine the number of students needed to obtain reliable data, and we had to determine the level of student exposure to the content in school before the pilot test. We also needed to decide on the most appropriate student demographics for our needs (e.g., grade, ethnicity, gender, percent free or reduced lunch).

For *Amazing Cells*, we conducted the pilot test with 79 students in three 10th-grade biology classrooms, using items we had refined with data from the cognitive interviews. We piloted with students who had not previously studied cell biology in school in order to identify any items that could be answered with little or no prior knowledge about the topic; these items were eliminated from our final assessments. For *Epigenetics*, we chose to test with 83 biology students from three classrooms who had just completed the GSLC *Epigenetics* supplement in order to identify and eliminate items that were too difficult (i.e., that the majority of students could not answer correctly even after exposure to the material). We determined that three classes' worth of data should produce adequate variation in response patterns while keeping the qualitative data to a manageable level. Additionally, our priority was testing with the appropriate grade level rather than testing with students of certain ethnicities or other demographic indicators. If we had been interested in identifying items that might be biased toward different groups, we might have increased our sample size and performed specialized analyses such as differential item functioning (Osterlind and Everson, 2009).

Knowing that we would refine our items based on data from the pilot—or eliminate questions entirely—we administered tests with 23 items for *Amazing Cells* and 12 items for *Epigenetics*, more than we expected to appear on the final tests (the final *Amazing Cells* tests had 16 MC and open-ended items, and the final *Epigenetics* tests had 8 MC and open-ended items). We applied basic item analyses to students' responses. To evaluate item difficulty, for instance, we ran frequencies for the percentage of correct answers and for each incorrect distractor students selected. It is generally recommended to use items that 30–80% of respondents answer correctly, especially if you hope to measure change from pretest to posttest (Kehoe, 1995). We could also have examined item discrimination statistics, which compare a student's performance on a single item with his or her total test score. In other words, does a student who gets a particular item correct also score high on the rest of the test, or is it possible to get an item correct but score poorly on the other questions? A correlation of 0.15 or less indicates that an item does not effectively delineate between high- and low-performing students and should be eliminated (Kehoe, 1995).

Both item difficulty and item discrimination statistics can be calculated with most standard statistics packages or spreadsheet tools and are therefore recommended for simple item analyses. Researchers or instructors constructing high-stakes measures administered to large numbers of students may wish to consult with a psychometrician to perform more complex analyses utilizing Rasch modeling and generalizability theory. Rasch modeling enables test developers to estimate the difficulty of different items on the same continuous scale and construct multiple assessment forms (Wilson, 2005; Bond and Fox, 2007), while generalizability theory can be used to estimate the number of items, raters, and testing occasions needed to minimize measurement error and obtain an optimally reliable estimate of performance (Shavelson and Webb, 1991).

Table 6 provides examples of two *Epigenetics* items we refined based on feedback from the pilot test. In example 1, we replaced one item about the relationship between DNA, traits, and proteins with another item about the function of DNA. Only 24.7% of the pilot students answered this item correctly, slightly under our threshold of 30%. We had expected a higher percentage of correct responses, given that students had just completed the *Epigenetics* supplement. On further reflection, however, we decided that the concept of traits was not heavily emphasized in the supplement. This fact, compounded with the complexity of the answer choices (each of which compared three different terms), led us to generate another, simpler item to assess students' understanding of the central dogma.

Example 2 illustrates our revision of the open-ended question we discussed in the previous scoring section. We not only developed a rubric that was sensitive to students' responses, but adjusted the item to better elicit the information we wanted. We realized from the responses that we were more interested in knowing about what the epigenome does, which is a higher-level concept, rather than simply what it is. We revised the question accordingly.

**Table 6.** Revision of two *Epigenetics* items based on classroom pilot test results[a]

| Pilot test | Final test |
|---|---|
| 1. What is the relationship between genes and traits?[b] | 1. Which of the following statements about DNA or genes is the most accurate? |
| A. Genes code for DNA. DNA is responsible for individual traits. | A. DNA provides the instructions for making proteins. |
| B. Genes code for proteins. Proteins are responsible for individual traits. | B. Genes provide the instructions for making DNA. |
| C. Genes code for chromosomes. Chromosomes are responsible for individual traits. | C. DNA provides the instructions for making carbohydrates. |
| D. Genes code for carbohydrates. Carbohydrates are responsible for individual traits. | D. Carbohydrates provide the instructions for making DNA. |
| E. Genes are not related to traits. The environment is primarily responsible for individual traits. | E. Proteins provide the instructions for making genes. |
| 2. What does the epigenome do? | 2. Explain (a) *what* the epigenome does to the genome, and (b) *how* it does it. |

[a] Correct answers for item 1: pilot: "B"; final: "A."

[b] Adapted from the Genetics Literacy Assessment instrument (Bowling *et al*., 2008).

After we had used the feedback from the cognitive interviews and classroom pilot tests to revise individual items, we still had to assemble those items into the final pre and post assessments to be used in the curriculum field test. In this final step, we again used strategic placement of items, careful alignment to the supplement objectives, and adherence to time constraints in the classroom to compile an appropriate test. We also checked the distribution of correct answers to ensure that there were no systematic patterns (e.g., B, C, D as correct answers to successive questions), nor a preponderance of one correct letter choice over the others.

For all of our validation procedures, including the cognitive interviews and our pilot testing, we obtained human subjects research approval from our institution. It is important to note that both your instrument validation studies and your field research or evaluation studies may be considered human subjects research, depending on the steps you take (National Science Foundation [NSF], 2015; U.S. Department of Health and Human Services, 2015). Therefore, research approval may need to be obtained from your institution's institutional review board (IRB), the committee that oversees research involving human subjects. Before starting your project, you will need to check with your institution on the requirements for conducting research and publishing the results.

## SUMMARY AND CONCLUSIONS

The findings from an evaluation of a K–12 educational intervention or the claims made about student performance in an undergraduate or graduate course depend greatly on the quality of the instruments used to assess learning. In this essay, we outlined an iterative four-step process for developing and validating items for small-scale, low-stakes research and instructional contexts. First, identify the intervention's or curriculum's broad learning goals. Then, outline and prioritize the specific learning objectives you wish to measure. Second, seek out instruments that address those objectives and/or draft new items. Consider how well the items fit the knowledge and reading level of your target audience and conform to recommended item-writing guidelines (Haladyna *et al.*, 2002). Sometimes you may have to modify existing items. Third, as you prepare items, consider the number of points to assign to each answer and establish clear, reliable scoring criteria for open-ended questions. Fourth, validate your instruments using a variety of data sources, including expert review, cognitive interviews, and testing in classrooms. Collectively, these data provide the backing for warranted arguments supporting the instruments' interpretation and use (DeBarger *et al*., 2013; Kane, 2013).

In our experience, we have learned three particularly salient lessons about creating measures.

1. *Plan your development efforts by clearly describing what you are measuring and the context in which you are collecting your measurements*. It can be tempting to rush into writing items without explicitly articulating your instrument targets, but resist this inclination. The more work you can do up front to describe your intended instrument, the more efficient you can be in the later phases of development. In fact, one of the authors (K.B.) often organizes her thoughts on an instrument cover sheet (see the Supplemental Material) before she begins researching or drafting items. As we have also demonstrated throughout this article, the prior knowledge of the individuals taking the assessments can have a significant influence on their interpretation of items. Reminding yourself of this audience up front can help you construct your questions accordingly. Finally, you may want to think about the time and resources available for data analysis, as this can influence your decision to use easily scored MC items or more labor-intensive open-ended or other types of items.
2. *Leverage a community of experts to help design and review your assessments*. Assessment development is typically not an individual enterprise, but a community affair. We encourage (when possible) collaborations between assessment developers and curriculum designers to prioritize the content to be assessed and to draft items. If needed, consult with content experts to make sure the items accurately represent the ideas the instrument is intended to measure, educators to ensure that the wording of the items is appropriate for the students who will be taking the assessment,

and assessment-takers to identify any unanticipated misinterpretations of test questions. It may also be beneficial to work with psychometricians, external evaluators, or other researchers with expertise in assessment development to add rigor to your work. Some universities have teaching and learning centers with staff members who may be able to help you develop new assessments or improve existing ones. We encourage you to consult these and any other experts you may be able to access.

3. *Allocate more time than you expect*. Developing assessments is not a one-time, one-sitting process, but requires multiple rounds of planning, drafting, review, and revision. The amount of time required to construct an instrument varies based on a) the complexity and number of constructs to be measured, b) the availability of existing instruments, c) the necessity of securing IRB approval to pilot test items, and d) the ease or difficulty of recruiting students or classes who can participate in pilot testing within your preferred time frame.

We believe that investing in constructing quality measures is essential for advancing biology education research and improving the pedagogy of science faculty. As noted in the *Introduction*, vetted assessments are critical for identifying evidence-based curricula and practices that improve student learning. Research and evaluation studies that do not use validated measures are limited in the strength of the claims they can make about the efficacy of the materials or program they investigated. As funding agencies increasingly require projects to demonstrate broad societal impact (NSF, 2007), plans for research and evaluation with thoroughly crafted instruments can influence panelists' decisions to recommend a proposal for funding. Moreover, faculty whose career advancement, retention, or promotion depends in part on demonstrating advancements in education need to provide rigorous evidence of their accomplishments. Data from instruments shown to be valid for their intended purpose can carry significant weight in this regard. Finally, assessment quality is one of the factors considered by manuscript reviewers, since it is part of judging the approach used to study a teaching practice or evaluate an intervention.

If the steps outlined in this paper initially seem daunting, start by asking yourself two questions: "What exactly do I want students to know and do by the end of this course/set of curricular materials?" and "What evidence would convince me that students have learned the core objectives?" Your answers will help you begin to think about the questions you should be asking in your assessments.

You can also take small steps to improve your existing instruments. For example, begin by reviewing an MC test you have used in a recent course. Look at the percentage of students who got each item correct. Were there any items that you thought more students would answer correctly? Were there items that you thought would be difficult and that you expected *fewer* students to answer correctly? For open-ended questions, you could take a similar tally of scores. If you find that more than 80% or less than 30% of the students got full credit on any test question, those items may be too easy or difficult for you to say anything meaningful about learning.

Pick five to seven items that seem problematic and decide how you will handle them. Is the concept addressed by the question important to assess? If so, modify it. If not, replace it. Use the strategies in this primer to guide your revisions and consider reaching out to colleagues for assistance. Incorporate the revised items into a new assessment (even a comprehensive final exam) and see how students perform. You might even include the old items on the same test for comparison. Are you now more confident about what you can say about what students have learned?

We have found that assessment development requires creativity, collaboration, and persistence, much like many other scientific endeavors. With time and experience, the process can become easier, though it will always involve challenging decisions about constructs, observations, and interpretations. We are confident that these efforts are worthwhile. Engaging in an evidence-based process of assessment item development can go a long way toward improving the contributions from small-scale biology education projects.

## ACKNOWLEDGMENTS

We thank Molly Malone, the GSLC's senior education specialist, who helped us prioritize the constructs we measured in our assessments and reviewed drafts of items; Sheila Homburger, the GSLC's science content manager, who also reviewed drafts of items; the students who participated in cognitive interviews; the teachers who allowed us to pilot test the instruments in their classrooms; *LSE* monitoring editor Dr. Ross Nehm and the reviewers who provided valuable feedback on the manuscript; and Dr. George Bass for his advice on a portion of the article. The project described was supported by award number R25RR023288 from the National Center for Research Resources (NCRR). The content is solely the responsibility of the authors and does not necessarily represent the official views of the NCRR or the National Institutes of Health.

## REFERENCES

Allen D, Tanner K (2006). Rubrics: tools for making learning goals and evaluation criteria explicit for both teacher and learners. CBE Life Sci Educ *5*, 197–203.

American Association for the Advancement of Science Project 2061 (2011). Developing and using assessments aligned to science learning goals. Workshop held 12–14 October 2011, in Washington, DC.

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

Arter J, McTighe J (2001). Scoring Rubrics in the Classroom: Using Performance Criteria for Assessing and Improving Student Performance, Thousand Oaks, CA: Corwin Press.

Bass KM, Magone ME, Glaser R (2002). Informing the Design of Performance Assessments Using a Content-Process Analysis of Two NAEP Science Tasks, CSE Technical Report 564, Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing.

Baxter GP, Bass KM, Glaser R (2001). Notebook writing in three fifth grade science classrooms. Elem School J *102*, 123–140.

Baxter GP, Glaser R (1998). Investigating the cognitive complexity of science assessments. Educ Meas *17*, 37–45.

Bond TG, Fox CM (2007). Applying the Rasch Model: Fundamental Measurement in the Human Sciences, 2nd ed., New York: Taylor & Francis.

Bowling B, Acra E, Wang L, Myers MF, Dean GE, Markle GC, Moskalik CL, Huether CA (2008). Development and evaluation

of a genetics literacy assessment instrument for undergraduates. Genetics *178*, 15–22.

Briggs DC, Alonzo AC, Schwab C, Wilson M (2006). Diagnostic assessment with ordered multiple-choice items. Educ Assess *11*, 33–63.

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

Chi MTH (1997). Quantifying qualitative analyses of verbal data: a practical guide. J Learn Sci *6*, 271–315.

College Board (2009). College Board Standards for College Success (Science), New York: College Board.

Cook DA, Beckman TJ (2006). Current concepts in validity and reliability for psychometric instruments: theory and application. Am J Med *19*, 166.e7.

Deane T, Nomme K, Jeffery E, Pollock C, Birol G (2014). Development of the Biological Experimental Design Concept Inventory (BEDCI). CBE Life Sci Educ *13*, 540–551.

DeBarger AH, Penuel WR, Harris CJ (2013). Designing NGSS Assessments to Evaluate the Efficacy of Curriculum Interventions, Austin, TX: Educational Testing Service. www.ets.org/Media/Research/pdf/debarger-penuel-harris.pdf (accessed 29 March 2016).

Drits-Esser D, Bass KM, Stark LA (2014). Using small-scale randomized controlled trials to evaluate the efficacy of new curricular materials. CBE Life Sci Educ *13*, 593–601.

Elrod S (2007). Genetics Concepts Inventory. bioliteracy.colorado.edu/Readings/papersSubmittedPDF/Elrod.pdf (accessed 28 January 2016).

Fives H, Huebner W, Birnbaum AS, Nicolich M (2014). Developing a measure of scientific literacy for middle school students. Sci Educ *98*, 549–580.

Garvin-Doxas K, Klymkowsky KM, Elrod S (2007). Building, using, and maximizing the impact of concept inventories in the biological sciences: report on a National Science Foundation–sponsored conference on the construction of concept inventories in the biological sciences. CBE Life Sci Educ *6*, 277–282.

Genetic Science Learning Center (GSLC) (2008a). Learn.Genetics: Amazing Cells. http://learn.genetics.utah.edu/content/cells (accessed 18 January 2015).

GSLC (2008b). Teach.Genetics: Amazing Cells. http://teach.genetics.utah.edu/content/begin/cells (accessed 18 January 2015).

GSLC (2009a). Learn.Genetics: Epigenetics. http://learn.geneticsutah.edu/content/epigenetics (accessed 18 January 2015).

GSLC (2009b). Teach.Genetics: Epigenetics. http://teach.genetics.utah.edu/content/epigenetics (accessed 18 January 2015).

Hadenfeldt JC, Bernholt S, Liu X, Neumann K, Parchmann I (2013). Using ordered multiple-choice items to assess students' understanding of the structure and composition of matter. J Chem Educ *90*, 1602–1608.

Haladyna TM, Dowling SM, Rodriguez MC (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ *15*, 309–334.

Hallgren KA (2012). Computing inter-rater reliability for observational data: an overview and tutorial. Tutor Quant Methods Psychol *8*, 23–34.

Hamilton LS, Nussbaum EM, Snow RE (1997). Interview procedures for validating science assessments. Appl Meas Educ *10*, 181–200.

Hanauer DI, Bauerle C (2015). The Faculty Self-Reported Assessment Survey (FRAS): differentiating faculty knowledge and experience in assessment. CBE Life Sci Educ *14*, ar17.

Hickey DT, Ingram-Goble AA, Jameson EM (2009). Designing assessments and assessing designs in virtual educational environments. J Sci Educ and Tech *18*, 187–208.

Horizon Research (2009). Interview protocol for student assessment items, Washington, DC: ATLAST Student Writing Workshop, 24–25 January 2009.

Howes PD, Chandrawati R, Stevens MM (2014). Colloidal nanoparticles as advanced biological sensors. Science *346*, 1247390.

Kalas P, O'Neill A, Pollock C, Birol G (2013). Development of a meiosis concept inventory. CBE Life Sci Educ *12*, 655–664.

Kane MT (2013). Validating the interpretations and uses of test scores. J Educ Meas *50*, 1–73.

Kehoe J (1995). Basic item analysis for multiple-choice tests. Pract Assess Res Eval *4*. http://PAREonline.net/getvn.asp?v=4&n=10 (accessed 20 March 2015).

Lovelace M, Brickman P (2013). Best practices for measuring students' attitudes toward learning science. CBE Life Sci Educ *12*, 606–617.

Martinez ME (1999). Cognition and the question of test item format. Educ Psychol *34*, 207–218.

McNeill KL, Krajcik J (2011). Supporting Grade 5–8 Students in Constructing Explanations in Science: The Claim, Evidence and Reasoning Framework for Talk and Writing, New York: Pearson Allyn & Bacon.

Moskal B, Leydens J (2000). Scoring rubric development: validity and reliability. Pract Assess Res Eval *7*. http://PAREonline.net/getvn.asp?v=7&n=10 (accessed 14 October 2015).

Mullis IVS, Martin MO, Ruddock GJ, O'Sullivan CY, Preuschoff C (2009). TIMSS 2011 Assessment Frameworks, Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Lynch School of Education, Boston College.

National Assessment Governing Board (2015). Science Framework for the 2015 National Assessment of Educational Progress, Washington, DC.

National Research Council (NRC) (2001). Knowing What Students Know: The Science and Design of Educational Assessment, Committee on the Foundations of Assessment, Washington, DC: National Academies Press.

NRC (2006). Systems for State Science Assessment, Committee on Test Design for K–12 Science Achievement, Washington, DC: National Academies Press.

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

NRC (2014). Developing Assessments for the Next Generation Science Standards. Committee on Developing Assessments of Science Proficiency in K–12, Washington, DC: National Academies Press.

National Science Foundation (NSF) (2007). Broader Impacts Review Criterion. www.nsf.gov/pubs/2007/nsf07046/nsf07046.jsp (accessed 7 September 2015).

NSF (2015). Human Subjects. www.nsf.gov/bfa/dias/policy/human.jsp (accessed 26 August 2015).

Next Generation Science Standards Lead States (2013). Next Generation Science Standards: For States, by States, Washington, DC: National Academies Press.

Osterlind SJ, Everson HT (2009). Differential Item Functioning, 2nd ed., Thousand Oaks, CA: Sage.

O'Sullivan CY, Reese CM, Mazzeo J (1997). NAEP 1996 Science Report Card for the Nation and the States: Findings from the National Assessment of Educational Progress, Washington, DC: National Center for Education Statistics.

Perez KE, Hiatt A, Davis GK, Trujillo C, French DP, Terry M, Price RM (2013). The EvoDevoCI: a concept inventory for gauging students' understanding of evolutionary developmental biology. CBE Life Sci Educ *12*, 665–675.

Perkel JM (2014). Life science technologies: the digital PCR revolution. Science *344*, 212–214.

Reeves TD, Marbach-Ad G (2016). Contemporary test validity in theory and practice: a primer for discipline-based education researchers. CBE Life Sci Educ *15*, rm1.

Rudner LM (1994). Questions to ask when evaluating tests. Pract Assess Res Eval *4*. http://PAREonline.net/getvn.asp?v=4&n=2 (accessed 2 February 2016).

Sadler PM, Coyle H, Smith NC, Miller J, Mintzes J, Tanner K, Murray J (2012). Assessing the life science knowledge of students and teachers represented by the K–8 national science standards. CBE Life Sci Educ *12*, 553–575.

Shavelson RJ, Webb NM (1991). Generalizability Theory: A Primer, Newbury Park, CA: Sage.

Simkin MG, Kuechler WL (2005). Multiple choice tests and student understanding: what is the connection? Decis Sci J Innovative Educ *3*, 73–97.

Songer NB, Ruiz-Primo RA (2012). Assessment and science education: our essential new priority? J Res Sci Teach *49*, 683–690.

Stanger-Hall KF (2012). Multiple-choice exams: an obstacle for higher-level thinking in introductory science classes. CBE Life Sci Educ *11*, 294–306.

Stemler SE, Tsai J (2008). Best practices in interrater reliability: three common approaches. In: Best Practices in Quantitative Methods, ed. J Osbourne, Thousand Oaks, CA: Sage, 29–49.

U.S. Department of Health & Human Services (2015). Human Subject Regulations Decision Charts. www.hhs.gov/ohrp/policy/checklists/decisioncharts.html (accessed 26 August 2015).

Vogel G (2014). Testing new Ebola tests. Science *345*, 1549–1550.

Wallert MA, Provost JJ (2014). Integrating standard operating procedures and industry notebook standards to evaluate students in laboratory courses. Biochem Mol Biol Educ *42*, 41–49.

Wiggins G, McTighe J (2005). Understanding by Design, 2nd ed., Alexandria, VA: Association for Supervision and Curriculum Development.

Wilson MR (2005). Constructing Measures: An Item Response Modeling Approach, Mahwah, NJ: Erlbaum.