Adv Physiol Educ 41: 110–119, 2017;
doi:10.1152/advan.00116.2016.

BEST PRACTICES

Best practices in summative assessment

Jonathan D. Kibble
College of Medicine, University of Central Florida, Orlando, Florida

Submitted 22 July 2016; accepted in final form 9 January 2017

Kibble JD. Best practices in summative assessment. *Adv Physiol Educ* 41: 110–119, 2017; doi:10.1152/advan.00116.2016.—The goal of this review is to highlight key elements underpinning excellent high-stakes summative assessment. This guide is primarily aimed at faculty members with the responsibility of assigning student grades and is intended to be a practical tool to help throughout the process of planning, developing, and deploying tests as well as monitoring their effectiveness. After a brief overview of the criteria for high-quality assessment, the guide runs through best practices for aligning assessment with learning outcomes and compares common testing modalities. Next, the guide discusses the kind of validity evidence needed to support defensible grading of student performance. This review concentrates on how to measure the outcome of student learning; other reviews in this series will expand on the related concepts of formative testing and how to leverage testing for learning.

summative assessment; validity; blueprinting; reliability; generalizability

SUMMATIVE ASSESSMENTS are usually applied at the end of a period of instruction to measure the outcome of student learning. They are high stakes for all concerned, most obviously for the learners who are being judged but also in the sense that the data may be used to drive course improvement, to assess teaching effectiveness, and for program-level assessments such as accreditation. At the other end of the spectrum, we define formative assessments as those intended to enrich the learning process by providing nonjudgmental feedback; they are assessments for learning than assessments of learning (39). Assessment often falls somewhere between these pure summative and formative poles, for example, when grade incentives are provided for assignments or quizzes during a course. Therefore, there is a continuum of summative to formative assessment depending on the primary intended purpose, although feedback to learners should be a common feature.

Both summative and formative testing have important effects on student learning, and careful attention on the selection and deployment of each is needed. It is an age-old axiom that summative assessment drives learning since most college-level students will think hard about strategies to maximize performance. On the other hand, we should not underestimate the value of formative assessment, especially given the recent demonstrations of how powerfully the "testing effect" enhances learning and memory compared with other study methods, such as rereading a text (38). Therefore, just as selection of a summative assessment plan must align with the overall course goals, formative assessment should be an integral part of the instructional plan for a course. The present review will focus on the practical steps needed to build robust tools to measure final learning outcomes from the instructor perspective; leveraging assessment for learning will be the topic of another review in this series.

Criteria for Excellent Assessment

One of the most enduring frameworks to define what makes a good assessment is van der Vleuten's notion of assessment utility, which he defined as the product of reliability, validity, feasibility, cost effectiveness, acceptance, and educational impact (44). Reliability refers to the reproducibility of the measurement; validity asks whether there is a coherent body of evidence supporting the use of the assessment results for their stated purpose, i.e., does the test measure what it purports to? Feasibility and cost effectiveness relate to how realistic tests are in the local context, and acceptance refers to the whether all the stakeholders have regard for the process and the results. Educational impact relates to whether the assessment motivates students to prepare in ways that have educational benefits. Norcini et al. (35) extended this framework to include equivalence and catalytic effect. Equivalence asks if similar results and decisions will occur when tests are used across cycles of testing or in different institutions, and the idea of catalytic effect asks whether the results and feedback from assessment drive future learning forward. We will draw on these frameworks throughout this review to clarify the purpose of various suggestions in an effort to remain evidence based in an area of education where intuition and tradition often exert powerful influence on instructors.

Learning Outcomes and Assessment Planning

My experience has been that subject matter experts naturally tend to start thinking about the content they should teach in a course, then about how they will teach it, and finally about how to assess student learning. As an example, a few years ago, I wrote a review textbook for medical physiology (27) and, looking back, I did not think much about learning outcomes, relying instead on what seemed implicitly clear content the book would need to include. All I really did was create my own synthesis of well-trodden ground, with some multiple-choice practice questions thrown in for assessment. In contrast, shortly afterward, I joined the planning team in a new medical school where we had to decide how discipline-based learning would be incorporated into an integrated curriculum (26). We were now confronted with student learning outcomes that placed at least equal importance on patient care, critical thinking, team skills, communication, information literacy, and professionalism as they did on knowledge of physiology. Therefore, our

Address for reprint requests and other correspondence: J. D. Kibble, Univ. of Central Florida College of Medicine, 6850 Lake Nona Blvd., Orlando, FL 32827-7408 (e-mail: jonathan.kibble@ucf.edu).

assessment plan needed much more than written tests of medical knowledge but also included practical assessments, direct faculty observation of students, peer assessment, projects, portfolios, collaborative writing, and group presentations.

The landmark *Vision and Change* report charting the future of undergraduate biology education makes clear that we should drive our course planning from the intended big-picture learning outcomes (2). By definition, a good learning outcome must be measurable, such that serious thought about summative assessment is needed at the start of the planning process. Learning taxonomies are helpful when developing and matching learning outcomes with assessments. The most commonly used is Bloom's taxonomy (8), which was modified by Anderson and Krathwohl (4) to define six cognitive process dimensions (remembering, understanding, applying, analyzing, evaluating, and creating) that can be applied in four different knowledge dimensions (factual, conceptual, procedural, and metacognitive); an excellent interactive version of this taxonomy with examples is available via the Iowa State University Center for Excellence in Teaching and Learning (9). Crowe et al. (15) have also developed a "Blooming Biology Tool" to assist in aligning learning outcomes and assessment and have presented initial data suggesting improved teaching and learning outcomes.

In medical education, the same mantra of driving curriculum decisions from learning outcomes is an accreditation requirement (29a). Borrowing from medical education, a simple but powerful taxonomy for thinking about assessment is Miller's pyramid (32), which describes levels of learning starting with a knowledge base ("knows") to basic competence of knowing what should be done ("knows how") to being able to demonstrate a skill or behavior under standard conditions ("shows") to actually applying the competencies in a real situation ("does"). A new top layer was recently proposed for Miller's pyramid for individuals in advanced training (e.g., PhD and health professionsals) who have formed a true professional identity and consistently display such values (the person not only "does" but also "is") (16). Table 1 shows common assessment methods using Miller's pyramid and provides commentary on some of their advantages and disadvantages (see also Refs 3, 5b, 17, 29, and 42).

In physiology courses, it is likely that many of our examinations will be of the written type. Although written tests are classified in the "knows" and "knows how" levels, they certainly have the potential to test higher orders of learning. For example, questions that include data or graphical interpretation or that require predictions or decisions address levels of application, analysis, and evaluation. Longer-form written exams, such as essays and projects, can require students to provide a synthesis of multiple sources of information or even demonstrations of creativity that could, for example, be presented orally or as posters and can be collected over time in a portfolio (30). As we move further away from knowledge testing to higher levels where performance observations are needed or collections of work are judged, use of rubrics becomes essential to make clear for students and graders what the standards for accomplishment are (1). A good rubric is a matrix that clearly articulates what the levels of achievement are with clear descriptors of performance levels that meet expectation, exceed expectation, or are below expectation. The syllabus should also indicate how scores are applied to the rubric and how scores from different assessments are combined.

In a previous study (28), I used a modified Bloom's taxonomy to annotate a new question bank that was created by nine faculty members for an upper-division undergraduate human physiology course. Despite setting many learning outcomes for the course beyond the knowledge level, about half the questions developed were still found to be at this most basic level and only ~20% reached the application level. While testing some basic knowledge is a good thing, the data indicated that faculty members often defaulted to testing knowledge, perhaps because such items are easier to develop and grade. For me, this experience underlined how important it is to be intentional about matching assessment to learning outcomes and also the need for faculty development and peer review in the test development process. As an aside, we also discovered that faculty members cannot reliably judge the difficulty of individual items they write; we will be discussing best practices around test construction and standard setting later.

Validity: Meaningful Interpretation of Test Scores

The historical literature on validity is complex, and a search on the topic is likely to yield articles about multiple types of assessment validity. The 21st century consensus definition according to the *The Standards for Educational and Psychological Testing* (2a) is more straightforward: "Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of the test." For example, the Medical College Admissions Test (MCAT) proposes that test scores are a good predictor of whether students will do well in medical training. This unitary idea of validity is also referred to as construct validity. A construct is some postulated attribute of people, assumed to be reflected in their test scores (14); in the MCAT example, the construct is "educational ability."

The negative decisions we make about our students based on test scores can have enormous emotional and financial consequences for them personally, just as our positive decision to graduate or progress a student has consequences for their future teachers, professions, and clients. Therefore, validity becomes the most fundamental consideration once we start building tests for decision making. Validity does not refer to the assessment instrument itself but rather to the scores produced at a given time in a given context and with a given group of students in relation to how those results will be used to make decisions. The modern idea of validity requires that we establish lines of evidence to argue that we can make good decisions based on our test scores. According to the *Standards* (2a), there are five general lines of validity evidence, which are based on the following: test content, response process, internal test structure, relations to other variables, and assessment consequences. In effect, validity is about stating hypotheses about how a test may be used and gathering the evidence to support or refute the hypothesis. For something like the MCAT, huge resources are needed to establish what material should be tested and to develop excellent test items, and equal effort is needed to show that the results indeed predict future outcomes in medical school and beyond. However, even if we are single instructors in a foundation-level course, there are some simple things,

Table 1. *Common assessment tools grouped by Miller's levels of competence*

Assessment Categories and Instruments	Description of Assessment	Advantages	Disadvantages
	Assessment of "knows" and "knows how"		
MCQs	MCQs are a selected response instrument consisting of a stem or case/problem description, a lead-in question, and a list of options. Single-best answer is the most common format.	Efficiently tests a broad range of knowledge and application in a single test; easiest format to produce statistically reliable results; automated marking	Limited cognitive level tested; should not be used to extrapolate what students can show/do; hard to write to avoid technical flaws that allow "test-smart" students to select correct responses; faculty training needed
Short answer	Student answers structured questions with an open-ended response. They are scored against predetermined model answers.	Easy to create items; reasonable content coverage; easier to grade than full essays but still requires significant effort; often valued because the answer is not in front of the student (constructed response instead of selected response)	Require a large number of questions to match MCQs for reliability (e.g., 30-40 structured items); usually test at the same cognitive levels as MCQs but less efficient
Essay/report	Student submits prose in response to a stimulus. They can be scored either with a points system against a rubric or using a global rating.	Easy for faculty members to create; can assess written communication skills; possibility of assessing complex topics and ability to make coherent arguments	Limited representation of content; usually modest reliability and interrater agreement; time consuming for faculty members to grade
Oral exam (viva voce)	One or more examiners ask candidates questions face to face. Questions should be blueprinted; answers should be recorded and graded against a predetermined rubric.	Traditional in some disciplines (notably surgery); valued for demonstration by the candidate of synthesis under pressure but no verified advantages; better used in formative situations	Low reliability; high interrater variation; testing usually limited to knowledge level; prone to unconscious examiner biases
	Assessment of "shows" (demonstrations of performance in simulated setting)		
Laboratory practical/simulated clinical exams	Students are observed performing defined tasks in a specified time and are graded against a standardized checklist. In medical education, the most common example is the objective structured clinical examination.	Fairly authentic situations; tasks or cases can be standardized, allowing more reliable grading	Good reliability requires several stations (usually >10); labor intensive to create, grade, and set standards; expensive to deploy; is a demonstration of student's best effort, not what is done in real practice; context specificity makes it hard to extrapolate skills observed at a given station
	Assessment of "does" (conscious demonstration of performance in a real-world setting)		
Direct observation	Students are observed in a practice situation such as a laboratory or clinic. Rating scales are needed to describe the criteria of interest (e.g., procedural skills or communication skills).	Provide an assessment of what learners do in real situations; easy to administer; can also provide global rating of performance	Requires faculty training to assure reliability; need to have multiple encounters to provide reliable data
Portfolio	A collection of work samples, projects, and evaluations over time to provide evidence of achievement of goals. It should be accompanied by student goal setting and frequent faculty feedback on progress.	A representation of actual performance over time; powerful as a feedback and as a progress monitoring device	Time consuming for students and faculty members; often low acceptance by students; hard to grade reliably and to set standards under high-stakes conditions
Peer assessment	A group of students typically assess each other's work using a rubric or criteria previously determined by faculty. The object of assessment is variable (e.g., project, presentation, or professional behavior).	Encourages student responsibility and ownership; develops student skills of judgment; valuable alternative feedback perspective, especially for teamwork and behavior	Grade inflation likely with less reliability of scores; better suited to formative assessment; reluctance to give negative feedback if not anonymous; requires faculty members to brief students on how to assess and give feedback; should be supervised
Self-assessment	Students make judgments about their own learning, achievements, and learning outcomes, usually according to established criteria.	Encourages goal setting and responsibility; promotes the development of reflective practice	Grade inflation likely with less reliability of scores; requires guided practice to develop self-monitoring skills
360-Degree (multisource feedback)	Surveys completed by several individuals within the candidate's domain of competency, including supervisors, peers, other coworkers, and clients. These are sually targeted at observable behaviors and interpersonal skills.	Authentic assessment in a real-world setting; includes multiple perspectives; provides evidence about behavior and therefore is a powerful feedback tool	Reluctance of evaluators to provide negative feedback of workmates; large number of raters (>10) needed for reliable data; difficult to deploy and collect data
Assessment of "is" (consistent demonstration of expected values, attitudes, and behaviors; fully formed professional identity, e.g., independent scientist or healthcare provider)			
Interviews	A subject-object one-on-one interview to explore professional identity.	Indepth personal exploration	Requires a highly skilled examiner; data from "does" level are a prerequisite
Standardized survey inventories	A new area of research with limited tools available.	Easy to deploy; theoretically grounded	Relies on self-report; not well validated at this time

MCQ, multiple-choice question.

discussed below, that we can do to avert major threats to validity.

Validity Evidence Based on Content

As physiology teachers, the most common type of test we use is probably a written assessment, and for this the most fundamental type of validity evidence relates to test content. There should be documentary evidence of a test outline and plan that shows in detail what topics are tested (specifically, how many test items on each topic), how they relate to the learning outcomes, and what cognitive levels are being tested. An external examiner should be able to look at this document and agree that the test is a representative sample of the domain of interest and appropriately addresses the goals of the course. The notion of whether a test is a representative sample of domain of interest is a basic and critical component of test development. This is sometimes referred to as instructional validity and includes not only a face value judgment of content sampling but consideration of the extent to which instruction was truly provided for the tested content. Errors in this stage of test development are likely to have a major impact on acceptability of test results, discussed further below.

Documentation of the test outline and plan is known as a test blueprint and can be a simple table with a topic or learning outcome in each row, with columns describing other attributes such as cognitive levels (e.g., knowledge or application), competencies (e.g., knowledge or attitude) and where each cell indicates the number of items devoted to that category. The examples shown in Table 2 are for single tests. The syllabus should indicate the overall assessment plan for a course describing the different types of assessment and their relative weighting. Similar tables can be created to show how tests relate to learning outcomes. Coderre (10) has provided some excellent tips on making and using blueprints, among the most important of which is sharing them with stakeholders. Feedback from colleagues in the planning stage can avoid one of the most common validity threats known as construct underrepresentation (18). Construct underrepresentation may mean too few items in particular areas, bias of some kind in the item selection, or a mismatch with learning outcomes. While we will be discussing the importance of test statistics (see *Validity Based on Internal Structure*), the value of expert opinion during test development should not be underestimated since most decisions about testing end up being value laden in some way. The finished blueprint should also be shared with all the teachers in a team-taught course (likely it will serve as their instruction on what items to write) as well as with students so that the process is transparent. In my experience, such an approach goes a long way to assuring a sense of fairness and broad acceptance of the assessment plan (the other dimension of fairness is difficulty level, which is also discussed at greater length later). Coderre points out that it is not enough to prepare a blueprint; there also needs to be follow through to prepare items according to the plan and to audit adherence to the blueprint. It is very helpful to create item banks using a commercial software program, which typically allow for metadata tags to be applied to items, such as what learning outcomes an item addresses as well as linkage to the performance statistics. Once a blueprint is operational, it can also become apparent that the learning objectives do not properly reflect the true relative importance of concepts, which often emerge as the ones most tested, and this can help inform continuous course improvement.

Validity Evidence Based on Response Process

This category is mostly about the integrity of data throughout the testing process. This might seem trivial but I can recall two painful examples that make me attentive this aspect of validity. In my first year as a faculty member, I recall a very distressed student pleading with me to reconsider a grade I had given on an essay. Although initially skeptical, I retrieved her essay from a large stack to discover that the number circled in pen on her script was not the number represented in the grading spreadsheet: a simple clerical error that had caused much

Table 2. *Examples of assessment blueprints at the level of a single test*

A 100-Item Written Physiology Semester Exam

	Body System					
Cognitive Level	Renal	Cardiovascular	Pulmonary	Endocrine	Reproductive	Item Totals
Remembering	1	3	2	2	2	10
Understanding	6	6	7	7	4	30
Applying	6	8	6	5	5	30
Analyzing	6	5	3	3	3	20
Evaluating	1	3	2	3	1	10
Creating	0	0	0	0	0	0
Item Totals	20	25	20	20	15	100

A Clinical Skills Midterm Exam with 12 Stations

	Body System/Topic				
Physician Task/Competency	Musculoskeletal	Cardiovascular	Pulmonary	Neurological	Gastrointestinal
History taking Skills	2	3	2	3	2
Physical exam skills	2	3	2	3	2
Clinical reasoning (data interpretation)	2	3	2	3	2
Plan and management	2	3	2	3	2
Psychosocial issues included	1	2	1	1	1
Lifestyle medicine/prevention	1	2	2	1	1

unnecessary distress. I also once published an electronically rescored multiple-choice final exam to over 300 medical students only to discover that the way I had used the program caused some kind of scoring error. Both errors were caught and corrected but at some cost to all concerned. At my current institution, we have set a policy allowing 1 wk for scores in all courses to be double checked and we conduct data audits before anything is published to students. However, the reality is that individual faculty members are often doing a lot of manual grading under tight deadlines and data processing errors are probably common. Response process validity is best achieved with a quality assurance plan, starting with clear testing instructions to students and practice opportunities for examinees regarding test formats (e.g., are they familiar with computer-based testing programs or how to complete test forms?). There should be a documented process for checking final answer keys and any rescoring procedures when items are removed, a protocol for how data moves between software systems such as from spreadsheets to learning management systems, and an audit of manual data entry. In addition to students experiencing smooth test deployment and scoring processes, the syllabus should explain cases where scores are combined to give composite totals, and score reports should allow students to reproduce final grading data; any rubrics or other forms of rating instrument should be available and explained to students before tests are administered.

Validity Evidence Based on Internal Structure

This relates to the reproducibility (reliability) of test scores and to other statistical or psychometric properties (e.g., item difficulty and discrimination index). My experience has been that this is one of the least appreciated areas in high-stakes classroom testing and one of the most critical types of validity evidence for test scores that will be used to assign grades because it deals with measurement error. The idea of reliability is simply whether test scores are reproducible (41). In classical test theory, it is assumed that the behavior being measured in a person or a group is stable and that an observed test score consists of a true score of the ability or behavior combined with an error score. Sources of combined error include human factors such as level of fatigue or anxiety at the point of testing as well as inherent errors within the measurement tool. To explore good test-retest reproducibility, it is helpful to think about students taking an imaginary parallel test, such that by comparing the results we could determine if students get the same scores and are ranked in the same position in the class and if decisions about pass/fail or grades are the same. In classroom testing, we do not usually have the option of double tests but could, for example, randomly split the test scores in half and compare the two data sets. The most commonly used reliability coefficient is Cronbach's α (13), which takes the idea of splitting a test up to its logical limit by subdividing it the most possible times, i.e., by comparing each item to the rest of the test. Cronbach's α provides a test-retest correlation value between 0 and 1 and is referred to as a measure of internal consistency, with a value of 1.0 indicating a perfectly reproducible test. It is often quoted that $\alpha > 0.9$ is the ideal target for high-stakes tests with a lower limit of 0.7 being acceptable for classroom tests (19, 36).

So why does reliability matter? The value of knowing α is that we can use it to estimate error and explore confidence intervals of scores at different cut levels, using the SE statistic, which is the SD of the error term, calculated as follows:

$$SE = SD_{(x)}\sqrt{1 - \alpha}$$

Table 3 shows sample data showing how the confidence interval for a test score is affected by changing the test reliability from 0.5 to 0.9. Imagine a traditional grading scheme of ABCF, where 90%, 80%, and 70% cut scores are applied, respectively. How should we treat the case of a student with a score of 65%? The data in Table 3 show that if our test reliability is 0.9 or above, then the student's score of 65% is outside a 95% confidence interval and we would probably feel comfortable awarding a failing grade. However, if our test reliability is <0.8 in this case the student has a score within the 95% confidence interval. Would you fail this student? My school currently has an ABCF grading scheme, and most courses unconsciously avoid the conundrum of borderline test failures by including in the overall assessment plan some continuous assessment points or team points that mean in practice a score lower than 65% is likely to be needed to fail a course outright. We have also used standard setting methods to aid with decision making (see below). On the other hand, I have encountered faculty members in the past who have absolute faith that a score of 69% represents a true failure, when they have no knowledge of the margin of error in their examinations. We will next introduce some other basic test statistics and consider how to maximize test reliability.

Any commercially available testing program will calculate Cronbach α for a set of test scores as well as providing some other standard item statistics. Most useful are individual item discrimination indexes that allow faculty members to review the performance of test questions. Item discrimination describes the extent to which success on a given item corresponds to success on the whole test. A discrimination index (there are many) is calculated using equal-sized high- and low-scoring groups on the test. The idea is that if generally strong students get an item correct and weak students get it incorrect, then the item is "discriminating." In practice, for each item, the number of successes by the low-scoring group is subtracted from the number of successes by the high-scoring group and this difference is divided by the size of a group, producing an index ranging from $+1$ to -1. Traditionally, the top and bottom 27% of the class are used, and, generally speaking, item discrimination values of $+0.4$ and above are regarded as high and less than $+0.2$ as low (20). Another approach to discrimination is to calculate the "point-biserial correlation," which is the Pearson correlation between responses to an item and scores on the

Table 3. *Effect of test reliability on the confidence intervals for test scores*

Reliability Coefficient (Cronbach α)	SE, %	95% Confidence Interval, %
0.5	4.2	±8.3
0.6	3.8	±7.4
0.7	3.3	±6.4
0.8	2.7	±5.3
0.9	1.9	±3.7

Note: data are derived from a test with a mean score of 80% and SD of 6%.

whole test and, therefore, also takes a value from +1 to −1. Values of either index that are close to zero or are negative detract from the test reliability since we assume that all items on the test are cooperating to measure the same attributes and such potentially faulty items should be carefully reviewed after the test.

Item analysis is part science and part judgment; items with a high percentage correct will naturally fail to discriminate, but several such items are likely to be intentionally included on a test to gauge basic mastery and these should not be eliminated just because they do not discriminate (although ideally there are not too many such items). At my school, we routinely include opportunity for students to record item challenges using report cards during the test and these can be a great help when combined with item analysis to understand what went wrong if an item performs poorly: was the wording ambiguous? does it conflict with what was taught in some way?

Our ability to maximize reliability of tests comes down to two major factors: *1*) having enough test items and *2*) having high-quality discriminating items. As a guide, if the average item discrimination is around +0.3, then 50–60 items are enough to produce a reliability of around 0.8; if the average item discrimination is +0.2, then 100 items are needed (24), whereas >100 items usually produces only small additional gains in reliability. Item difficulty around 60–70% correct gives the best potential for high discrimination. However, the reality is that there may be limits to the number of "hard" items you can use depending on the grading traditions of your institution. For example, 50% of students with a grade C in a course would likely make acceptance levels for assessment very poor, so there are always trade offs and judgments to be made. Item discrimination levels are also a function of the students; if you teach a course with a wide range of ability levels, this tends to produce high item discrimination, whereas classes such as those with medical students are usually fairly homogeneous in ability levels and there is not much real difference between the top and bottom quartiles in a class.

From the foregoing discussion, we can better appreciate solutions to the two most common threats to validity in faculty-developed tests: namely, construct underrepresentation and construct irrelevant variance (18). Construct underrepresentation is most commonly a problem of too few items in the sample domain but also results from the inclusion of trivial test items, maldistribution of test items across topics, or poor reliability; use of a strong blueprint that broadly samples the domain of interest with enough items to generate reliable measurements and the development of high-quality items using peer review should prevent construct underrepresentation problems. Construct irrelevant variance represents noise in the measurement and increases the error term. Construct irrelevant variance has several sources, but most are again related to poorly constructed items that can be caught in peer review, for example, items may be too hard, too easy, contain trivial details, are culturally insensitive, are biased in some way (e.g., long reading time for second language students), or are off target from learning outcomes. Other examples causing construct irrelevant variance are items that include language cueing test-wise students to the correct answer and guessing from limited option sets. Several studies have shown how peer review can significantly increase item quality and test reliability (31, 33, 46). Another factor that causes construct irrelevant variance is "teaching to the test," which may result in scores that do not accurately reflect what students know or do not know. This is one reason my school does not allow faculty members to give preexam review sessions, which often carry an implicit expectation of clues to the tested content; instead, we invite student questions and provide liberal access to faculty office hours leading up to major tests. Instances of cheating or loss of test security are other examples of construct irrelevant variance in test scores. Downing (18) also notes that indefensible passing scores produce construct irrelevant variance and are a major validity threat, bearing in mind the whole point of validity is to be on solid ground when making decisions from the test outcomes.

Generalizability theory is an alternative to the basic approach to reliability studies offered by classical test theory. Generalizability theory uses an analysis of variance approach, and a generalizability coefficient is calculated as the ratio of wanted variance: total variance; if the only input variable is ranked student scores, this produces the same answer as Cronbach α on a 0–1 scale. However, generalizability theory is much more flexible because the investigator can identify intended facets (factors) of variance such as students, items, or raters, and the analytic approach allows each variance to be examined separately. The statistical tools also allow for a followup decision study that allow estimates to be calculated for how each variable can be manipulated to increase reliability. For example, how many additional items or raters would be needed to reach a reliability of 0.9? If the reader is familiar with doing ANOVA calculations, generalizability coefficient calculations are fairly straightforward and free software tools are available (6).

The most advanced approach to reliability is the use of item response theory. Unless the reader has a statistics background or available expertise, this approach is less accessible and probably only worth pursuing if you are conducting larger-scale exams, perhaps with different test forms or multiple campuses. Classical test theory and generalizability theory are both limited by the inability to separate the effects of test difficulty from student ability. Item response theory is based on setting up probability functions in which the probability of correctly answering an item is a function of student ability. This sigmoidal curve will move position and slope according to item difficulty, discrimination, and guessing effects. Pretest data are needed to execute the mathematical models, making it impractical for most single instructor courses, but supplemental literature is provided for the interested reader (43).

Validity Evidence Based on Relationships to Other Variables

It is often the case for physiology teachers that the main purpose of our judgments about student learning is to determine readiness for future learning, whether it is progressing within an undergraduate program, moving on to further training or to employment. The validity hypothesis (commonly known as predictive validity) is that exam results meaningfully predict that students are ready for this next stage, which is often readily testable by checking on the outcome. For example, in a new medical school, we needed to show that our newly developed internal assessment program would produce meaningful prediction of success on United States Medical Licensing Exam (USMLE) Step 1 (25). This is an example of

convergent validity evidence where we should expect that tools measuring very similar constructs produce similar outcomes. It is also valuable to include comparisons expected to produce divergent outcomes. For example, correlation with physiology exams in our institution is much lower when comparing outcomes with patient interviewing skills or research project performance (unpublished observations). Similarly, in a recent study (12) where we developed a novel assessment to focus on clinical reasoning, the degree of correlation with knowledge testing outcomes was significantly less than previous comparisons between tests of medical knowledge. These kinds of observations, when taken together, give confidence that we are able to make valid measurements of the intended construct.

Analysis of the predictive power of test results is necessarily a long-term project, but we can also look to concurrent tests for validity evidence. In the case of a new school, we elected to run a series of progress tests in parallel with the formal curriculum using the National Board of Medical Examiners Comprehensive Basic Science Exam, which was given five times over a 2-yr period. We were able to correlate results of the internal testing program with these external gold-standard tests as a concurrent outcome to provide validity evidence for our newly developed exams (25). In undergraduate physiology courses, similar data could be obtained by comparing student testing outcomes in parallel courses that are measuring similar constructs; curriculum committees or institutional quality improvement offices can usually offer support for such studies.

Validity Evidence Based on Consequences of Testing

Consequences validity evidence is a relatively new domain but is somewhat analogous to van der Vleuten's consideration of educational impact (44). Cook and Lineberry (11) have recently likened high-stakes summative assessments to medical tests in that they both result in important decisions and actions for the subject; the argument that follows is that neither kind of test should be performed unless the need is justified and benefit clearly exceeds costs. We are asking the following question: "Does the activity of measuring and the subsequent interpretation and application of scores achieve our desired results with few negative side effects?" (11). Consequences evidence considers impacts that may be beneficial or harmful, intended or unintended (2). For example, in my school, we have recently changed the definition of a C grade from "conditional" (meaning a progress committee would review the candidate in detail to determine if remediation should occur) to "unconditional" (meaning the student passes the class without further discussion). After graduating four classes, we were able to model the relationship between the number of C grades and final outcome on USMLE Step 1; while C grades were correlated with lower scores, they did not predict outright failure, and thus our remediation point needed to be revised. An additional concern was that the presence of a conditional C grade evidently produced student distress by generating uncertainty as to whether a student with a C grade would be promoted to the next academic year. The high level of student distress expressed in perception surveys was an example of an unjustified negative side effect, and the overall consequences validity argument indicated a need for change.

In educational scholarship, we are familiar with using final assessment scores as the outcome measure to determine whether instructional interventions have had a positive impact on learning but we rarely think about the impacts of the assessment itself. Cook and Lineberry (11) have proposed a framework that includes assessing impacts first on the examinee: is there evidence that the test itself promotes learning? For example, I have frequently advocated decreasing the number of summative knowledge assessments within courses to allow more time for learning and have never observed any appreciable change in final exam performance when summative quizzes were replaced with formative quizzes (i.e., I found no apparent learning benefit of making midsession quizzes summative). Another student impact to consider is whether there is evidence of improved preparation due to testing; for example, does the presence of a practical exam induce greater time spent practicing skills rather than remembering information? What are the effects on motivation, emotions, and well-being of the summative testing program?

We can also investigate impacts on faculty members. Is there evidence that the curriculum is being improved to address apparent areas of student weakness? Are teachers collaborating more effectively as a result of sharing in the planning and development process? Are scholarly projects emerging? Is there higher status attached to demonstrations of externally validated high-achieving students? How are their emotions and well-being affected by student performance or by resource limitations? Similarly at the program level, evidence of consequences or impacts of testing might relate to allocation of resources or curriculum changes driven by testing outcomes.

A final special case related to consequences validity is the impact of grading classifications. This is most pronounced at the pass/fail cut point such that standard setting requires particular attention, discussed further below. Apart from any practical considerations for repeating or remediating failed courses, there is likely to be a negative impact on self-efficacy and motivation of receiving the label of "failure" (40). In medical education, the issue of what classifications to use for grading is a hot topic, with many schools shifting from traditional letter grade to pass/fail systems (7). Given that grades are used later on to help select graduates for highly competitive postgraduate residency training, the impact of grades is potentially huge. At my school, we are actively reviewing whether to shift to a pass/fail system with concerns that we may be hurting students who are competent to practice but have some C grades compared with students in other schools who have an undifferentiated "pass" grade. Examining the consequences of labels is therefore an important topic, especially remembering the data in Table 3, which demonstrates that the difference between a B and a C could be spurious to begin with!

Setting Standards

The cut points on exams are given special significance that can have major impact on examinees, especially around the pass/fail line. Emphasis on traditional arbitrary numbers like "70% is passing" is rather meaningless unless faculty members justify this special status. At a minimum, it is helpful to maintain a database of examination items over time and to keep record of student performance. This allows some degree of prediction about the likely test outcomes and ability to compare new items with old items. Once a testing database is established, faculty members must make decisions about whether it

will be completely sequestered or not. What degree of postexamination review and feedback will be allowed? Will past examinations be provided for students to review before testing? In my view, the summative testing database should be secure since it takes a lot of faculty effort to create a validated bank that has been subject to peer review and item analysis. It is rarely possible to generate completely new high-quality assessments each year. A secure item database is the bedrock of valid and reliable testing. However, providing students with feedback is also important and can risk the leaking of questions. In most database programs, annotation of items is possible that allows detailed reporting of strengths and weaknesses by topic. In addition to this, our faculty members hold closed-test reviews for the purpose of coaching and apply the same security measures used during examinations. Students who fail tests are allowed to review one on one with a faculty member. We monitor item performance each time an item is reused and watch for trends such as decreasing difficulty and discrimination that suggest an item may be compromised. Practice quizzes should be developed separately from the main summative item bank; they should be used liberally during the learning phase of the course and should include rich feedback.

There are several formal standard setting methods that can provide stronger justification for where cut points are defined (5a, 34). By their nature, standards are an expression of values and all the methods rely on expert judgment in some way. The first step is to select the kind of standard desired. This can be norm referenced to the performance of examinees in a cohort or criterion referenced ahead of time. Norm-referenced standards are most suited to situations like admissions or selecting students for awards, where examinees are being ranked for selection to some category with limited availability. When we are interested in whether students are competent, criterion-referenced standards are more appropriate, although faculty members sometimes gravitate to norming scores as an easy fix when difficulty levels seem wrong after a test. Many formal standard setting methods have been described and validated and have been reviewed in more detail elsewhere (5a, 34). As an example of this type of process, one of the most commonly used is the Anghoff method (5), in which a panel of judges (6–8 judges ideally) are first asked to discuss the characteristics of the borderline (minimally competent) student. The judges then go through the whole test and indicate for each question whether such as student should get it correct or not. The mean of the judge's scores is used as the passing standard. I have used this method many times to check that a test conforms with the institutionally fixed values of passing grades (e.g., 70%) and found it to be remarkably accurate with question histories and often close to the lower limit of the 95% confidence interval of actual student scores on the test. The Anghoff method has the advantage that judges can also comment on the individual items as a check on content validity and item quality. The credibility of passing standards set this way depends on who the judges are, and they should ideally be a diverse group of faculty members with good working knowledge of the curriculum and students. Standard setting represents a gold-standard ideal that is not possible in all situations. However, even having another pair of eyes on the test items in development and a colleague to help review and make decisions during item review and scoring is helpful to strengthen validity.

Other Aspects of Assessment Utility and the Need to Compromise

Fairness of assessment is one aspect not completely addressed through consideration of validity evidence. The *Standards* (2) describe fairness in several ways: lack of bias, equitable treatment of all in the testing process, and equality in outcomes and in opportunities to learn. Assuring that tests are as fair as possible requires a combination of planning and data gathering. Before testing, all examinees should have equal access to learning materials, practice opportunities, and test instructions. In the test development process, there should be an effort to avoid introducing bias. For example, if the construct of interest is knowledge of physiology, then unnecessarily complex language or complex mathematical treatments beyond the prerequisite course level should be avoided. Monitoring differential item functioning between ethnic or other groups is advisable where possible to evaluate possible sources construct irrelevant variance affecting certain groups. Where direct interactions between the examiner and examinee are involved in the testing process, the examiner needs to be particularly conscious of potential bias and introducing construct-irrelevant variance through factors such as undue stress on examinees. For example, I can vividly remember as a student not doing well on a pharmacology oral exam given by two rather angry and probably very tired examiners; only after the encounter did I realize that I knew the correct answers to most of the questions they had asked. In cases where examinees have a learning or physical disability, the law requires that appropriate accommodations are provided and students should be made aware early in the program how to access such services and reminded of the process in each course syllabus.

There are several important elements to judging assessment utility that require qualitative data such as student and faculty surveys, focus groups, or interviews. Student input on the validity of content sampling, quality of items, the difficulty level, and a global sense of acceptance is easily obtained through perception of instruction surveys. Student perception is not the final determinant but is a valuable perspective to consider. Similarly, debrief meetings in team-taught courses can quickly establish the faculty viewpoint. Feasibility and cost effectiveness are areas that faculty members are usually quite vocal about, particularly in relation to the time demands on them for setting, supervising, and grading the assessment. These are very real concerns that often mean compromise is needed. A common problem is the introduction of construct underrepresentation discussed earlier because of feasibility concerns. For example, practical or clinical examinations are resource intensive and often end up with too few stations. Just as learning is contextual, so is assessment, and results from one testing station do not generalize to the whole construct (3, 41). For instance, the ability to solve a problem about cardiac function does not help to determine if the student can solve problems about the gastrointestinal tract, a problem known as case specificity. On the other hand, if we were to cancel the practical examination because of concerns for reliability of data, this could have a disastrous educational impact by leading students to avoid practicing the very skills that are needed to

meet the learning outcomes. In such cases, creative solutions are needed, such as having a shorter practical exam that is extended with supplementary written items to bolster reliability (45). The overall utility considerations for assessment often demand compromise and judgment.

Summary and Practice Points

High-stakes assessment is among the biggest responsibilities we have, given the potential impacts the results have on students from a social, emotional, and financial perspective as well as the long-term impact on our profession and future clients. In summary, some basic elements of practice for excellent assessment are as follows:

- Use backward design that starts by defining the learning outcomes and what types of assessment are most suitable to measure the outcomes.
- Document a testing blueprint that shows what domains will be tested and how this matches the learning outcomes; share the blueprint with all stakeholders.
- Engage as much as possible in faculty peer review during the test development process to avoid introducing construct underrepresentation and construct irrelevant variance.
- Include enough items, and items of high quality, to assure adequate test reliability and defensibility of scores.
- Apply standard setting methods.
- Provide students with clear instructions and practice materials and develop a plan to assure the integrity of data throughout the testing process.
- Monitor the fairness, acceptability, and impact of testing over time with routine surveying of stakeholders and comparison of test scores with other measures of student outcomes.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the author(s).

AUTHOR CONTRIBUTIONS

J.D.K. conceived and designed research; analyzed data; interpreted results of experiments; drafted manuscript; edited and revised manuscript; and approved final version of manuscript.

REFERENCES

1. **Allen D, Tanner K.** Rubrics: tools for making learning goals and evaluation criteria explicit for both teachers and learners. *CBE Life Sci Educ* 5: 197–203, 2006. doi:10.1187/cbe.06-06-0168.
2. **American Association for the Advancement of Science.** *Vision and Change in Undergraduate Biology Education: a Call to Action* (Online) http://visionandchange.org/files/2013/11/aaas-VISchange-web1113.pdf [04 July 2016].

2a. **American Educational Research Association, American Psychological Association, National Council on Measurement in Education**. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association, 1999.

3. **Amin Z, Seng CY, Eng KH (editors).** *Practical Guide to Medical Student Assessment*. Singapore: World Scientific, 2006. doi:10.1142/6109
4. **Anderson LW, Krathwohl D, Cruikshank KA, Mayer RE, Pintrich PR, Raths J, Wittrock MC.** *A Taxonomy for Learning, Teaching, and Assessing: a Revision of Bloom's Taxonomy of Educational Objectives* (Complete Edition). New York: Longman, 2001, p. 508–600.
5. **Anghoff WH.** Scales, norms and equivalent scores. In: *Educational Measurement*, edited by Thorndike RL. Washington, DC: American Council on Education, 1971.

5a. **Ben-David MF.** AMEE Guide No. 18: Standard setting in student assessment. *Med Teach* 22: 120–130, 2000. doi:10.1080/01421590078526.

5b. **Ben-David MF, Davis MH, Harden RM, Howie PW, Ker J, Pippard MJ.** AMEE Medical Education Guide No. 24: Portfolios as a method of student assessment. *Med Teach* 23: 535–551, 2001. doi:10.1080/01421590120090952.

6. **Bloch R, Norman G.** Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. *Med Teach* 34: 960–992, 2012. doi:10.3109/0142159X.2012.703791.
7. **Bloodgood RA, Short JG, Jackson JM, Martindale JR.** A change to pass/fail grading in the first two years at one medical school results in improved psychological well-being. *Acad Med* 84: 655–662, 2009. doi:10.1097/ACM.0b013e31819f6d78.
8. **Bloom BS, Krathwohl DR, Masia BB.** *Taxonomy of Educational Objectives: the Classification of Educational Goals*. New York: McKay, 1956.
9. **Center for Excellence in Learning and Teaching, Iowa State University**. *Revised Bloom's Taxonomy* (online). http://www.celt.iastate.edu/teaching/effective-teaching-practices/revised-blooms-taxonomy [15 July 2016].
10. **Coderre S, Woloschuk W, McLaughlin K.** Twelve tips for blueprinting. *Med Teach* 31: 322–324, 2009. doi:10.1080/01421590802225770.
11. **Cook DA, Lineberry M.** Consequences validity evidence: evaluating the impact of educational assessments. *Acad Med* 91: 785–795, 2016. doi:10.1097/ACM.0000000000001114.
12. **Cramer N, Asmar A, Gorman L, Gros B, Harris D, Howard T, Hussain M, Salazar S, Kibble JD.** Application of a utility analysis to evaluate a novel assessment tool for clinically oriented physiology and pharmacology. *Adv Physiol Educ* 40: 304–312, 2016. doi:10.1152/advan.00140.2015.
13. **Cronbach LJ.** Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297–334, 1951. doi:10.1007/BF02310555.
14. **Cronbach LJ, Meehl PE.** Construct validity in psychological tests. *Psychol Bull* 52: 281–302, 1955. doi:10.1037/h0040957.
15. **Crowe A, Dirks C, Wenderoth MP.** Biology in bloom: implementing Bloom's Taxonomy to enhance student learning in biology. *CBE Life Sci Educ* 7: 368–381, 2008. doi:10.1187/cbe.08-05-0024.
16. **Cruess RL, Cruess SR, Steinert Y.** Amending Miller's pyramid to include professional identity formation. *Acad Med* 91: 180–185, 2016. doi:10.1097/ACM.0000000000000913.
17. **Downing SM.** Assessment of knowledge with written test forms. In: *International Handbook of Research in Medical Education*, edited by Norman GR, van der Vleuten CP, Newble DI. Dordrecht: Kluwer Academic, 2002, p. 647–672. doi:10.1007/978-94-010-0462-6_25
18. **Downing SM.** Validity: on meaningful interpretation of assessment data. *Med Educ* 37: 830–837, 2003. doi:10.1046/j.1365-2923.2003.01594.x.
19. **Downing SM.** Reliability: on the reproducibility of assessment data. *Med Educ* 38: 1006–1012, 2004. doi:10.1111/j.1365-2929.2004.01932.x.
20. **Ebel RL.** Procedures for the analysis of classroom tests. *Educ Psychol Meas* 14: 352–364, 1954. doi:10.1177/001316445401400215.
24. **Hopkins K.** *Educational and Psychological Measurement and Evaluation*. Needham Heights, MA: Allen and Bacon, 1998.
25. **Johnson TR, Khalil MK, Peppler RD, Davey DD, Kibble JD.** Use of the NBME Comprehensive Basic Science Examination as a progress test in the preclerkship curriculum of a new medical school. *Adv Physiol Educ* 38: 315–320, 2014. doi:10.1152/advan.00047.2014.
26. **Khalil MK, Kibble JD.** Faculty reflections on the process of building an integrated preclerkship curriculum: a new school perspective. *Adv Physiol Educ* 38: 199–209, 2014. doi:10.1152/advan.00055.2014.
27. **Kibble JD, Halsey C.** *The Big Picture: Medical Physiology*. New York: McGraw-Hill Professional, 2009.
28. **Kibble JD, Johnson T.** Are faculty predictions or item taxonomies useful for estimating the outcome of multiple-choice examinations? *Adv Physiol Educ* 35: 396–401, 2011. doi:10.1152/advan.00062.2011.
29. **Kramer GA, Albino JE, Andrieu SC, Hendricson WD, Henson L, Horn BD, Neumann LM, Young SK.** Dental student assessment toolbox. *J Dent Educ* 73: 12–35, 2009.

29a. **Liaison Committee for Medical Education**. *Functions and Structure of a Medical School: Standards for Accreditation of Medical Education Programs Leading to the MD Degree* (online) http://lcme.org/publications/ [04 July 2016].

30. **Madden T.** *Supporting Student e-Portfolios: a Physical Sciences Practice Guide*. United Kingdom: The Higher Education Academy Physical Science Center, 2007.

31. **Malau-Aduli BS, Zimitat C.** Peer review improves the quality of MCQ examinations. *Assess Eval High Educ* 37: 919–931, 2011. doi:10.1080/02602938.2011.586991.
32. **Miller GE.** The assessment of clinical skills/competence/performance. *Acad Med* 65, *Suppl*: S63–S67, 1990. doi:10.1097/00001888-199009000-00045.
33. **Naeem N, van der Vleuten C, Alfaris EA.** Faculty development on item writing substantially improves item quality. *Adv Health Sci Educ Theory Pract* 17: 369–376, 2012. doi:10.1007/s10459-011-9315-2.
34. **Norcini JJ.** Setting standards on educational tests. *Med Educ* 37: 464–469, 2003. doi:10.1046/j.1365-2923.2003.01495.x.
35. **Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, Galbraith R, Hays R, Kent A, Perrott V, Roberts T.** Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach* 33: 206–214, 2011. doi:10.3109/0142159X.2011.551559.
36. **Nunnaly J.** *Psychometric Theory*. New York: McGraw-Hill, 1978.
38. **Roediger HL, Karpicke JD.** Test-enhanced learning: taking memory tests improves long-term retention. *Psychol Sci* 17: 249–255, 2006. doi:10.1111/j.1467-9280.2006.01693.x.
39. **Rolfe I, McPherson J.** Formative assessment: how am I doing? *Lancet* 345: 837–839, 1995. doi:10.1016/S0140-6736(95)92968-1.
40. **Schunk DH.** Self-efficacy and academic motivation. *Educ Psychol* 26: 207–231, 1991. doi:10.1080/00461520.1991.9653133.
41. **Schuwirth LW, van der Vleuten CP.** General overview of the theories used in assessment: AMEE Guide No. 57. *Med Teach* 33: 783–797, 2011. doi:10.3109/0142159X.2011.611022.
42. **Shumway JM, Harden RM; Association for Medical Education in Europe.** AMEE Guide No. 25: The assessment of learning outcomes for the competent and reflective physician. *Med Teach* 25: 569–584, 2003. doi:10.1080/0142159032000151907.
43. **Thissen D, Steinberg L.** Item response theory. In: *The SAGE Handbook of Quantitative Methods in Psychology*, edited by Millsap RE, Maydeu-Olivares A. London, UK: SAGE, 2009, p. 148–177. doi:10.4135/9780857020994.n7
44. **Van Der Vleuten CP.** The assessment of professional competence: developments, research and practical implications. *Adv Health Sci Educ Theory Pract* 1: 41–67, 1996. doi:10.1007/BF00596229.
45. **Verhoeven BH, Hamers JG, Scherpbier AJ, Hoogenboom RJ, van der Vleuten CP.** The effect on reliability of adding a separate written assessment component to an objective structured clinical examination. *Med Educ* 34: 525–529, 2000. doi:10.1046/j.1365-2923.2000.00566.x.
46. **Wallach PM, Crespo LM, Holtzman KZ, Galbraith RM, Swanson DB.** Use of a committee review process to improve the quality of course examinations. *Adv Health Sci Educ Theory Pract* 11: 61–68, 2006. doi:10.1007/s10459-004-7515-8.