Adv in Health Sci Educ (2016) 21:221–234
DOI 10.1007/s10459-015-9587-z

REVIEW

Progress testing: critical analysis and suggested practices

Mark Albanese · Susan M. Case

Received: 2 December 2014 / Accepted: 19 January 2015 / Published online: 7 February 2015

Abstract Educators have long lamented the tendency of students to engage in rote memorization in preparation for tests rather than engaging in deep learning where they attempt to gain meaning from their studies. Rote memorization driven by objective exams has been termed a steering effect. Progress testing (PT), in which a comprehensive examination sampling all of medicine is administered repeatedly throughout the entire curriculum, was developed with the stated aim of breaking the steering effect of examinations and of promoting deep learning. PT is an approach historically linked to problem-based learning (PBL) although there is a growing recognition of its applicability more broadly. The purpose of this article is to summarize the salient features of PT drawn from the literature, provide a critical review of these features based upon the same literature and psychometric considerations drawn from the Standards for Educational and Psychological Testing and provide considerations of what should be part of best practices in applying PT from an evidence-based and a psychometric perspective.

Keywords Progress testing · Longitudinal assessment · Formative assessment · High stakes assessment

Educators have long lamented the tendency of students to engage in rote memorization in preparation for tests rather than engaging in deep learning where they attempt to gain meaning

Susan M. Case was formerly with the National Conference of Bar Examiners and before that the National Board of Medical Examiners.

M. Albanese (✉)
University of Wisconsin-Madison Medical School, National Conference of Bar Examiners, 302 South Bedford Street, Madison, WI 53705, USA
e-mail: maalbane@wisc.edu

S. M. Case
Ajijic, Jalisco, Mexico
e-mail: smc53536@gmail.com

from their studies. Rote memorization driven by objective exams has been termed a steering effect (Blake et al. 1996; van der Vleuten et al. 1996; Norman et al. 2010). Progress testing (PT), in which a comprehensive examination sampling all of medicine is administered repeatedly throughout the entire curriculum, was developed with the stated aim of breaking the steering effect of examinations (Blake et al. 1996) and of promoting deep learning. It was developed at about the same time at the University of Maastricht (Maastricht) and the University of Missouri-Kansas City School of Medicine (UMKC) in the early-1970s (van der Vleuten et al. 1996; Arnold and Willoughby 1990). McMaster began using PTs in 1992 (Blake et al. 1996; Norman et al. 2010). Utrecht implemented PT in years 4 and 5 of their 6 year patient/problem-oriented curriculum in 2002–2003 (Rademakers et al. 2005). PT is an approach historically linked to problem-based learning (PBL) although there is a growing recognition of its applicability more broadly (Verhoeven et al. 2005).

The purpose of this article is to summarize the salient features of PT drawn from the literature, provide a critical review of these features based upon the same literature and psychometric considerations drawn from the Standards for Educational and Psychological Testing and provide considerations of what should be part of best practices in applying PT from an evidence-based and a psychometric perspective. We distinguish between uses that are low stakes, medium stakes and high stakes. Low stakes are for student use in self improvement; medium stakes will result in having students engage in remediation, but will not stop their degree progress; while high stakes uses can temporarily stop or terminate student progress toward their degree.

What is a progress test (PT)

General characteristics of PTs

The following qualities are distinctly characteristic of PTs: (1) Assessment of student learning is based upon "end-objectives" of the curriculum (van der Vleuten et al. 1996) or "competencies that students were expected to demonstrate upon graduation" (Arnold and Willoughby 1990) and is therefore decoupled from the specifics of what students have learned or are learning at any given time; (2) tests are created to be so comprehensive as to make it virtually impossible to study for them, especially using rote memorization approaches; (3) Scores on individual test administrations are used for formative assessment not for summative assessment; and (4) whether students are making adequate progress (medium and high stakes decisions) is judged on the basis of accumulated performance over several tests to reduce student concern over performance on any single test.

Implementation features of PTs

Besides these four general characteristics of PTs there are a number of features that are part of various implementations that are often, but not always used.

PT length

PTs generally have been long tests with from 180 multiple choice (MC) items (McMaster) to 400 MC items (UMKC). Maastricht used 250 True/False items until relatively recently converting to MC and Utrecht reported using 40 case scenarios from which 320 open

response items were drawn (Rademakers et al. 2005). McMaster reported allowing 3 h for students to complete their 180 MC item exam, no other article in the literature we reviewed reported the testing time provided.

Directions about guessing

Maastricht and McMaster both discourage students from guessing on questions. McMaster recommends to students that they "attempt to answer only those questions for which they have at least partial knowledge." (p. 1003) (Blake et al. 1996) Maastricht uses a "do not know" option and both McMaster and Maastricht use a scoring formula (also called the correction for guessing) in which wrong answers are penalized to discourage students from guessing.

PT creation and administration

In almost all cases, a new test is created for each test administration and there are from 3 (McMaster and Utrecht) to 4 administrations per year (Maastricht and UMKC). Each of the PTs is produced using the same overall test "blue print", an overall design that specifies the distribution of items among different content areas and other relevant features. In some schools (Maastricht), the PT production process is labor intensive and involves a large number of teaching faculty in the item writing and review processes and the final assembly of the test. A current reference for each item is generally required. Conversely, at McMaster each test is computer-generated from a relatively fixed set of about 3,000 items. One faculty member reviews the test and changes or rejects items.

Similarly, in some schools (Maastricht) scoring the exam also involves faculty input to weigh student feedback about test item quality and psychometric data in determining which items will contribute to final scores. Again, the opposite extreme is likely McMaster, where PTs are machine-generated, and identification of students with problems is done using a variant on norm referenced scoring, where students are flagged at 1.5 and 2 SD below their class mean.

Students across all grades are tested during the same testing window with the same PT. Items are released to students after the examination with detailed feedback and returning student feedback about item quality is often used to determine if some items should be excluded from computing final scores.

PT scoring, scores and score reports

MC, TF and other objective items are generally machine scored. Utrecht's open-ended response items were scored by the faculty who wrote the items (Rademakers et al. 2005). Although successive exams might differ in difficulty from one another, there generally has been no formal attempt to adjust for the varying difficulty and/or variability of scores (a process called equating), although McMaster actually reports student scores on a standardized progress curve to reduce the effect of test difficulty.

The score reports provided to students for feedback vary by the institution and can be quite extensive. At the UMKC, after each test administration, students received a copy of the test items along with the correct answers referenced to current literature as well as a printout of his or her performance on each item, his or her percentage of items correct on each of the major content areas and a comparison of their individual performance to the entire student body at the same educational level. Maastricht included comparative data from a group of physicians who had also taken the PT. McMaster reports the individual's

score, number attempted, correct/attempted, as well as class means for all classes sitting the test.

Summative uses of PT results

How the PTs are used for summative purposes varies by institution, but the results from a single PT administration are not usually used for high stakes decisions. Generally, medium and high stakes decisions about students such as enforced remediation, delayed progress, academic probation or cancellation of student registration are based upon student performance on all PTs administered in a given year. McMaster faculty were particularly wary of using PT data for high stakes decisions about students: "no student passes or fails solely from PT performance, no PT score is reported on record"; and mandated that such decisions had to include information from sources other than the PT (Blake et al. 1996).

PTs have also been used for comparing institutional performance in the Netherlands (Van der Vleuten et al. 2004; Muijtjens et al. 2008) as well as some cross national comparisons (Verhoeven et al. 2005). Instability in comparisons at specific time points across institutions have lead investigators to develop relatively complex cumulative deviation scores to stabilize the results (Muijtjens et al. 2008). However, these approaches have complex properties that need to be taken into consideration to interpret them properly (Albanese 2008).

Resource requirements

The amount of resources required by PTs varies by institution depending on specific implementation policies. Over time, massive item banks have been created as new items accumulate (McMaster accumulated 3,500 items approximately 3 years after implementing their PT (Blake et al. 1996), UMKC (Arnold and Willoughby 1990) and Maastricht (van der Vleuten et al. 1996) had 15,000 or more after 13+ years of implementation). However, in some schools these item banks are only marginally useful because there is an intense development and review process that each item must complete in order to be included on each PT and items are not "grandfathered in" because of previous use. Thus, items that have been used before are often revised before being used again. Arnold and Willoughby (1990) report that beyond extensive time required by faculty for item writing and review, a staff of five managed the UMKC PT process. Van der Vleuten et al. (2004) report that collaboration on the production of PTs with other institutions took time for the new institutions to develop adequate faculty support and infrastructure to meet the demands of the process. The PT at UMKC was maintained for over 30 years, but in 2004, the school replaced it with the Comprehensive Basic Science Examination of the National Board of Medical Examiners (http://research.med.umkc.edu/qpe/, Accessed September 24 2009). The reason for this change was in part the cost of faculty time and five support staff needed for the PT process (Personal communication with Louis Arnold, August 31 2007). If open-ended questions are used, additional faculty time will be necessary for grading (Rademakers et al. 2005). By contrast, the McMaster progress test, which is currently also administered at 3 other schools worldwide, is maintained on a part-time basis by a research coordinator.

It is common for the same PT to be administered to examinees at all levels within a short testing window, requiring significant physical space and staff to proctor the exams. As schools transition to computer administration of examinations, this may create logistical problems for administering a PT simply because all students across all classes are administered the PT at approximately the same time. This requires that there be sufficient numbers of computers to accommodate the large numbers of students or the curriculum

flexibility to stage the examination over an extended period. The testing window varies by school. For example, at UMKC it was an 8 h period from 8 a.m. to 4 p.m. on a Saturday. Doing it on a Saturday solved the curriculum flexibility problem because most classes were not in session (Personal correspondence with Louise Arnold, 4-14-14) A more radical option would be to put the examination on-line and allow students to take it from anywhere that they have web access. This unproctored option would raise potentially significant problems in determining: who is actually providing answers; if answers were individual or "group work"; and exam security. It would probably also be limited to low stakes applications since any higher stakes would make the aforementioned uncertainties intolerable. McMaster has recently spearheaded the International Partnership for Progress Testing (IPPT) that has partnered with 3 schools from North America, Europe and Australia to provide on-line progress testing (ipptx.org, accessed April 13, 2014). Each school determines the administration conditions, so the potential for unproctored administration may be close at hand.

Summary

To summarize, a PT is a comprehensive assessment that is often aimed at curriculum end-state skills and which is generally decoupled from the specifics of what students have learned at the time any given PT is administered. The PT is administered repeatedly throughout medical school, usually 3–4 times per year. A new PT is created with each administration; however, the same exam is administered to students across all classes during a particular testing window. PTs are created to be so comprehensive as to make it virtually impossible to study for them and in preparation for taking the PT, students are often instructed not to attempt to study. When PT scores are used for medium and high stakes purposes, progress is judged on the basis of accumulated performance over several tests to reduce student concern over performance on any single test. PTs can be resource-intensive owing to the need for large testing capacity over a relatively short time period as well as continuous review and updating of references and providing detailed feedback to students.

In the next section, we examine PTs from the perspective of psychometric issues that such an examination presents.

Psychometric issues

This section broadly considers the features of the PT that affect its reliability and validity. Because almost everything about an assessment will affect these qualities, we will more specifically examine the general qualities of PTs from descriptions and data reported in the literature in reference to the Standards for Educational and Psychological Testing (Standards) (1999). The Standards are a joint publication of the National Council on Measurement in Education, the American Educational Research Association and the American Psychological Association and represent best practices in testing drawn from the literature and expertise of noted scholars in the field. The specific issues we will address in this section are: A. Reliability, and B. Validity issues.

Reliability

We examined how scores vary under different conditions and compare the results to established guidelines for how much variability due to error is tolerable for scores to be

useful. Although reliability can be defined in various ways depending upon the different sources of variance taken into account, the most commonly employed estimate for PTs is either alpha internal consistency estimates or test–retest across repeated administrations. Internal consistency estimates such as Cronbach's alpha are dependent upon the test length, score variance and item variance. Score variance and item variance are affected by item difficulty and item discrimination, which in turn are affected by item quality. If an item is poorly written and ambiguous, difficulty will be increased, guessing or the use of the DK response will be increased, and the distinction between relatively knowledgeable and less knowledgeable students will be reduced.

Test length

The appropriate number of questions for a test is at least in part determined by the purpose of the test and the proposed use of test scores. PTs are generally used to provide formative feedback to students and curriculum managers about student progress without the steering effect of traditional exams. Medium and high stakes applications typically combine the results from 3–4 PT administrations.

The test–retest reliability of the McMaster exam ranged between .53 and .64 (Blake et al. 1996). The internal consistency alpha values of the Maastricht exam have ranged between .70 and .80 (van der Vleuten et al. 1996). The extensive length of the exams has probably been a function of their broad content coverage and the need to sample content from such a vast domain. The reliability values reported tend to be relatively modest considering the length of the PTs. However, they are adequate to serve most needs, particularly since medium and high stakes applications have generally been based upon combining results from 3–4 examinations. On the other hand, test length might become an issue when subscores are reported. Subscores are sometimes based upon a relatively small subset of items, which can yield problems with subscore reliability.

Item-types

Multiple choice and true–false items have been most frequently used in PTs. Utrecht employs short answer questions based upon 40 cases (Rademakers et al. 2005). If the PT is designed to test end-state skills, consideration should be given to include items that require application of knowledge within a vignette framework or are complex problem-solving exercises. Such items may be difficult to write, and require training of item writers in the mechanics of writing high-quality questions; training aids are widely available. In all situations, item writers need to have a good understanding of the purpose of the test. This understanding will be enhanced by participating in an interdisciplinary discussion of the issue. Item quality will be enhanced by item-writing training followed by extensive item review, both with regard to the accuracy of the item content as well as an assessment of the extent to which each item fits with the purpose of test. This in-depth item development requires extensive resources, making use of consortia or use of items developed by professional test development agencies worth considering.

Note that it may take students somewhat more time to answer such items. McMaster allowed students 1 min per item for taking PTs. The USMLE, for example, administers 45 questions per hour for Step 2 (80 s per question), which would be considered a graduation-level exam that includes only application of knowledge, vignette-style questions.

Scoring and grading

Scoring multiple choice and true–false exams has generally been a matter of reporting the percentage of items answered correctly. Key scoring issues surround whether or not to correct for chance success (formula scoring), and how to handle the DK response if it is used. In terms of interpreting group performance, applying formula scoring makes sense as it provides a portrayal of students overall performance that controls for the intrinsic likelihood of a random response achieving success. The matter gets more complicated in applying formula scoring to individual student scores since it includes negative points for wrong answers. If formula scoring is going to be used in scores that will reflect individual student performance and contribute to medium or high stakes decisions about the student, there are implications for the directions to students about what to do in the face of incomplete knowledge as well as the potential for student risk-taking propensity to affect scores. This has been a controversial matter that has been researched and debated for over 30 years. We will treat it in more detail in a later section under construct validity.

If items are used that must be scored by humans or that do not involve simple selection of options, then scoring becomes much more complex. Consistency between raters and within raters at multiple times are important considerations. Rater training and monitoring then become significant concerns. Standard 5.6 states that;

> When test scoring involves human judgment, scoring rubrics should specify criteria for scoring. Adherence to established scoring criteria should be monitored and checked regularly. Monitoring procedures should be documented." (1999, P. 65)

Utrecht reported using short answer items that were graded by the individual who wrote the item (Muijtjens et al. 2008). In this type of situation, it is hard to maintain quality control and determine to what extent grader objectivity has been maintained. In addition, if instructors are held accountable for student performance on the PT, having them score the items they wrote for the exam introduces a level of conflict of interest that could jeopardize the integrity of the process. It is generally preferable to have independent graders score such exams.

Scoring is one concern, but creating a grade that is used for summative high stakes decisions about students is another. Initially, McMaster used an innovative regression approach that took into account the longitudinal character of the PT. However, this approach has been criticized because it "could reward poor performance at the very beginning of the course" (p. 225) (McHarg et al. 2005) resulting in larger gain scores. McMaster dropped the regression approach for this reason. McHarg et al. (2005) recommend the three point classification used at Maastrict (unsatisfactory, doubtful, satisfactory). Normative approaches to setting standards have been recommended owing to the instability of passing rates when absolute test standards were applied to PT results over an 8 year period at Maastricht (Verhoeven et al. 1999). While normative approaches may be a practical solution to the instability of passing rates from absolute test standards, there are ways to equate scores that could provide stable passing rates from absolute test standards. At least the possibility for using criterion-based standards as well as use of periodic standard-setting exercises by stake-holders should be considered in planning for a PT.

Another scoring issue relates to comparing results from different PTs across time. If the exam questions are changed from one time to the next, or at any point in the long chain of administrations, the level of difficulty of each exam form is likely to be somewhat different. Growth and exam difficulty differences then become intertwined. This phenomenon has been recognized in the standardized testing industry in the use of parallel forms of exams.

> Standard 13.17 states:
>
> When change or gain scores are used, such scores should be defined and their technical qualities should be reported. The use of change or gain scores presumes the same test or equivalent forms of the test were used and that the test has (or the forms have) not been materially altered between administrations." P. 149 (Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education 1999). "The process of placing scores from such alternate forms on a common scale is called equating." (1999, P. 51). "The goal in equating is for scores to measure the same construct with very similar levels of reliability and conditional standard errors of measurement." (1999, P. 57).

There are many approaches to equating tests, but they usually involve administering the two (or more) forms of the examination to the same group or random subgroups of the examinees and then adjusting the distribution of scores on one exam to match the distribution of the scores on the other(s). The simplest adjustment is to make the mean scores on each form equivalent by adding or subtracting a constant to scores (Kolen and Brennan 1995; Kolen 1988). The important point is that if the exam questions change, it is more than likely that the score distribution will change, and this needs to be considered in interpreting the results from the different exams. Equating PT is especially challenging because students are expected to improve their performance from one administration to the next and each PT is composed of completely new items. As a consequence, changes in PT scores from one administration to the next can be due to either growth in knowledge or test changes or both. The IPPT consortium circumvents this problem by computing standard scores for each examinee based upon the mean and SD from their cohort of students. This removes the test difficulty from the equation, so students are only evaluated by their performance relative to their cohort. Moderate and high risk decisions are then only based upon combining the standard scores from multiple test administrations. While standardizing scores within cohort removes variation in intrinsic differences in test difficulty from scores, it can also remove growth. However, McMaster addresses this problem by reinterpreting standard scores on a hypothetical growth curve based upon an average over numerous administrations.

Score variance

The reliability of PT scores reported in the literature (usually alpha internal consistency values) have generally been relatively low, in the .45–.71 range (Blake et al. 1996; Arnold and Willoughby 1990, McHarg et al. 2005). Such low reliabilities for relatively long, centrally created tests are probably due to the inclusion of items for which students can only guess or select DK. Reliability is greatly affected by score range. If too many examinees perform at chance levels, guessing the correct answer or responding DK, the reliability will be reduced. Similarly, if too many examinees answer almost all the questions correctly because the test is easy, the reliability will be reduced. Generally, the maximum score variance, and therefore the greatest reliability, will be obtained if the test mean percent correct is mid-way between chance performance and 100 % correct. Thus, for a true–false PT with chance performance at 50 %, the test would be maximally discriminating if the mean percent correct for a given examinee population were 75 % [(100 − 50)/2 + 50]. For a 5-option multiple choice item (chance % = 1/5 × 100 = 20 %), it would be 60 % [(100 − 20)/2 + 20]. As the average item difficulty

for an examinee group departs from this point, the score variance tends to decline and reliability estimates tend to decline as well.

For newly entering medical students, performance on a PT is often at chance levels with large numbers of DK responses (Blake et al. (1996) report that McMaster first-year students did not answer 80 % of the items.) Students at Maastricht performed approximately 3 % above chance levels on the first examination they took (Muijtjens et al. 2008). At UMKC, Arnold and Willoughby (1990) reported minimum performance on multiple choice PTs to be 25 %; after adjusting for chance success, this would be from 0 to 6 %, depending on whether there were 4 or 5 options used (and assuming the percentage was based on the total of 400 items, not just those answered). For graduating medical students, performance has the potential to be quite high, which could introduce ceiling effects. However, this has not been the case in studies reported in the literature. Both Blake et al. (1996) and Muijtjens et al. (2008) found that PT scores reached a maximum at less than 50 % for chance corrected percentages, Arnold and Willoughby report maximum percentages for graduating students of 75 %, which represents a chance corrected value of 70 %; values that come closest to meeting what one would expect for students at the highest levels, but not at levels that would raise concerns about ceiling effects. This finding is concerning, implying that the questions are inaccurately targeted (i.e., too advanced, or otherwise not assessing knowledge and skills that should be known by the graduating student), or that the questions are poorly constructed without a clear correct answer.

Final reliability issues

A final issue is that the most important reliability estimate is the one for scores that are used in the high-stakes determinations of students' fate. Low achievement level in entry is expected, however low achievement level at graduation is not. As noted above, the low graduation achievement may be due to either item writers having unrealistically high expectations of student knowledge or to poorly written and ambiguous questions. It is critically important that scores used for making high-stakes decisions meet the highest standards and it is a two part decision. The first part pertains to the reliability of the score used to make the decision and the second part deals with the cut point on the score range that determines pass/failure. With PT, the combined results from an entire year of 3–4 PTs have generally been used to make such decisions. Even though the scores from individual tests may have modest reliabilities, the combination will be much more reliable. However, with the exception of the regression-based Personal Progress Index scores used at McMaster (test–retest values ranged from .53 to .64), the reliability of scores used for grading purposes have generally not been reported in aggregate.

As to the choice of a cut point, one of the few studies that specifically examined methods of standard setting identified a modified Angoff Method as producing results that are most consistent with normative approaches (1 SD below the mean) and yielded a relatively reasonable failure rate of 7.2 % among sixth year medical students (Verhoeven et al. 1999). However, the cut-score found was 41.4 % which for T/F items is less than chance.[1] Methods for setting cut scores and deriving overall grades that meet high psychometric standards remain a fertile area of research.

[1] The Angoff Method asks judges to visualized a hypothetical borderline pass group and then asks them to identify the percentage who would answer the item correctly. Verhoeven specifically instructed the judges not to correct for chance success, so as best we can tell, the percentage was based on the number of items that the borderline group should answer correctly out of the total of 250 T/F items.

Validity

Validity refers to whether the test measures what it is intended to measure and this is often heavily dependent upon how the scores are going to be used, that is, valid for what purpose. Validity is sometimes classified into different types. We will use the classic approach involving: content, predictive and construct validity.

Content validity

Content validity considers the degree to which an examination reflects the desired content domain. The content validity of PTs is considered by some to be its greatest strength. The content domain for the examination is supposed to represent the body of knowledge that students are expected to know upon graduation. This body of knowledge is generally agreed upon by the faculty in the medical school before the PT is created. Linking the examination to such a goal that was developed through such a consensus process has intrinsic appeal and is fundamental to content validity.

Predictive validity

Predictive validity refers to the extent that scores on an examination can predict future performance. Research at McMaster showed correlations between PT results and licensure examinations of 0.60 (Blake et al. 1996) This value was achieved using PT results obtained after 25 months of medical school (3 months before taking the licensure examination); lower values were seen for tests administered in the earlier years of education although a PT administered a year before had a predictive validity of about 0.50. Willoughby et al. correlated the UMKC PT with the NBME, Part I for four separate groups and found values of .32, .38, .59 and .82 (Willoughby et al. 1977). Although the amount of evidence for PT predictive validity is relatively sparse and some is over 30 years old, it has been promising.

Construct validity

Construct validity refers to whether the test is measuring the cognitive process or skill it is intending to measure. There are three issues that pertain to this. First, a PT is intended to measure a goal state, usually what students are expected to achieve at graduation. Achieving faculty consensus on what these expectations are and then stating them in terms that can govern PT construction is no small challenge. These expectations are likely to be a moving target that will need ongoing attention.

The second construct validity issue is that students should show progress on the PT as they progress in their education, ideally starting at chance levels on entry and being at the highest levels upon graduation. The evidence in this regard has been generally supportive.

Willoughby and Hutcheson (1978) reported PT mean percent correct scores at UMKC that increased across the 6 year curriculum from 6.1, 16.13, 30.67, 41.60, 50.93 to 56.00 %, respectively. Muijtjens et al. (2008) found that chance corrected PT scores were <5 % correct at baseline and reached a maximum of about 45 %. Blake et al. (1996) report PT results from three classes of students (1994–1996) at McMaster. They found scores began between 10 and 20 % beyond chance on entry and rose quite linearly until reaching approximately 50 % at the fifth exam administered at approximately the 20th month. After that point, the scores appear to flatten for the remaining three examinations as though reaching a ceiling at 50 %.

The low achievement level on entry is what would be expected and supports the construct validity of the PTs. The low maximum scores found for graduating students are not ideal, but it is generally the case that faculty over-estimate what student performance on examinations will be. This has been found from the time of the earliest attempts at standard setting by Nedelsky and the many different methods that have been used since. Generally, performance standards need to be titrated with actual data to arrive at reasonable expectations. This reinforces McHarg's recommendation to use normative standard setting approaches with PTs (McHarg et al. 2005).

The third issue pertains to avoidance of construct irrelevant variance, defined in the standards as the "degree to which scores are affected by processes that are extraneous to its intended construct." (1999, P. 10). One potential source of construct irrelevant variance can arise from using formula scoring. Several studies in the latter half of the last century argued that student risk aversion, a construct irrelevant characteristic for a PT, plays a role when examinees know that wrong answers will lead to points being subtracted from their score. The studies found that when students were instructed to answer questions they omitted under correction for guessing instructions, they did better than chance (Cross and Frary 1977; Traub et al. 1969; Albanese 1986), This finding was confirmed at Maastricht more recently by Muijtjens et al. (1999). Further, the mechanism underlying the differential performance may be complex. Kruger and Dunning (1999) found that individuals in the bottom 25 % tend to over-estimate their performance on assessments while those in the highest 25 % tend to under-estimate their performance. Because PTs are generally used for low stakes decisions, particularly early in the curriculum, the problems noted with the correction for guessing are probably limited. Low stakes, low risk. However, as scores accumulate and some students find themselves on the low end of the performance spectrum, the PTs will start to play a role in higher stakes decisions. As this higher stakes role starts to take effect, students may perceive the correction for guessing to have more weight and it could affect their test taking strategies. If one uses the correction for guessing in computing scores for students, consideration should be given to the potential to introduce student risk taking propensity as a factor affecting scores (and whether that is desirable) and the instructions to students should make it clear that such a correction will be applied.

Suggested practices

PTs may be extremely useful in keeping students in PBL curricula progressing toward general standards of competency. McMaster found the PT to be a solution to high rates of failure on the Licentiate of Medical Council of Canada (LMCC) toward the end of the second decade of their existence. LMCC board failure rates reached 19 %, more than four times the national average by 1989. They initially adopted the practice test for the LMCC developed by the University of Toronto. The PT was developed as a long-term solution. Failure rates immediately dropped to 5 % and scores on the LMCC continued to climb over the next decade (Norman et al. 2010). Thus, the PTs provided a means of keeping students progressing toward the general competencies of the profession, while still enabling them to have the benefits of the self-directed learning and small group problem solving elements so valued in PBL.

Formula scoring, sometimes used to discourage students from wild guessing, particularly early in the curriculum when they may be able to do little more than guess on PT items has the potential for making student risk taking propensity a factor in scores. Thus,

the correction for guessing should be used with caution, particularly toward points where moderate and/or high stakes decisions may be made.

PTs, are probably most appropriate when used in a curriculum with an emphasis on self-directed learning and little structure with regard to the order in which topics are studied, characteristics that are associated with a traditional problem-based learning curricula. In such curricula, students are expected to show growth over time, but are not expected to demonstrate expertise on specific topics at specific points in time; and total test scores are used for formative feedback, not to direct student learning.

A challenge to using PTs is that creating them can be expensive. New tests are needed for each of the 3–4 administrations in a given year. Consortia have sometimes been developed to share the costs. Examples are the IPPT spearheaded by McMaster, the schools in the Netherlands, and UMKC adoption of the NBME examination. Consortia are able to pool their psychometric expertise to address some of the more complex issues associated with PTs, such as establishing stable passing rates.

However, if the results are neither used to direct student learning nor for summative assessment, one could argue that PTs could be much shorter tests. This is particularly true for beginning students because examinee performance is expected to be near chance levels. One mechanism to develop shorter tests would be through multiple matrix sampling in which students would be subdivided into 3–6 groups and each group administered a different smaller sample of items (fraction equal to the reciprocal of the number of groups). This would allow testing time to be reduced to at most an hour as each student would have to answer one third or fewer items (60–83 items). The key to this approach would be to have a sufficient number of students in each subgroup and a sufficiently reliable subset of items to meet the needs of the examination. For example, if different schools are being compared, without delving into the complexities of equating for different subsets of items, there would need to be enough students in each subgroup at the different schools to have the statistical power needed to viably make the between school comparisons.

Some might argue that having students take a 3 h examination in which they can answer only randomly to items might be a good wakeup call to the reality that they have a lot to learn. While this is a good lesson for students to learn, except for exceedingly slow learners, it could probably be achieved within the first hour with the rest of the time yielding diminishing returns that could be put to better use. Some examinees may resolve this issue themselves by leaving early, but it is not clear if all who should do so actually do. Further, a shorter testing window would allow for alternative curriculum programming. At least having a shorter progress test with fewer items in the initial phases (e.g., first year) is something that could serve for future research.

In schools with a more structured educational setting, traditional PTs seem less appropriate. In these schools, an alternative compromise model might be more appropriate. For example, one could develop a PT for each year of medical school that covers the content which students are expected to learn in that year. Students would then take the Year 1 PT periodically over the course of the first year, and would be expected to show growth but not mastery until the end of the year. The questions would be more appropriate for the level of student knowledge at that point in time; reliability would be higher; and scores would provide a more concrete indication of the amount of knowledge learned that year. In subsequent years, they would be tested with the Year 2 PT, the Year 3 PT, etc. Something similar to this was done at the University of Michigan where students were assessed every 8 weeks on the clinical subjects as a whole, regardless of which clerkships students had completed. By the end of the year, they were expected to have mastered the clerkship material; during the course of the year, they were only expected to demonstrate

growth (Woolliscroft et al. 1995). Case Western Reserve conducted a somewhat similar study in which retired USMLE test material were used to develop and administer six web-based cumulative achievement tests (Swanson et al. 2010). The authors argue that a cumulative achievement test can be an effective complement to progress testing, with the achievement component encouraging students to retain already-covered material, while a progress testing component would assess growth toward the knowledge and skills expected of a graduating student.

Another alternative for LCME accredited schools might be to use the Comprehensive Basic Science Examination (CBSE) of the National Board of Medical Examiners as the UMKC turned to or old forms of the USMLE Step exams in a manner similar to that used at Michigan. Step 1 or the CBSE could be used for the first 2 years and Step 2 for the last 2 years. Both of these exams cover immense amounts of content and are composed of high quality items. The psychometrics of test scores could then be carefully managed by the National Board of Medical Examiners (NBME) and recommendations for the use of scores provided by those with expertise and a vested interest in ensuring proper use.

Summary of suggested practices

The use of PTs has slowly grown since being introduced in the early 1970s at Maastricht and University of Missouri-Kansas City. With recent multi-institutional collaborations on PTs and even global expansion possibilities, and with alternate uses of test scores, we recommend the following as best practices:

1. Review the purpose of your PT program and the uses that will be made of scores, making sure that the psychometric properties are sufficient for those uses. Be especially vigilant not to combine students from different classes in computing reliability and validity estimates as to do so will improperly include growth variance yielding inappropriately inflated values.
2. Consider the consequences of various incentives for performance on the PT, whether yielding inappropriately inflated values intended or not.
3. Pay careful attention to the instructions regarding test preparation and guessing, making sure that they are consistent with the uses of test scores and serve the best interests of students, especially those who are at highest risk for failure.
4. Review the use of resources, including student time to answer large numbers of questions on which they have little knowledge, and consider shorter PTs for students who are early in their education.
5. Review the quality and appropriateness of PT questions in light of the relatively poor performance often found with graduating students, especially if standards set are below chance performance levels.
6. Consider participating in consortia or other means of pooling resources for creating PTs.

References

Albanese, M. A. (1986). The correction for guessing: A further analysis of Angoff and Schrader. *Journal of Educational Measurement, 23*(3), 225–235.

Albanese, M. (2008). Benchmarking progress tests for cross-institutional comparisons: Which road taken makes a difference and all roads have bumps. *Medical Education, 42*, 4–7.

Arnold, L., & Willoughby, T. L. (1990). The quarterly profile examination. *Academic Medicine, 65*, 515–516.

Blake, J. M., Norman, G. R., Keane, D. R., Mueller, C. B., Cunnington, J., & Didyk, N. (1996). Introducing progress testing in McMaster University's problem-based medical curriculum: Psychometric properties and effect on learning. *Academic Medicine, 71*(9), 1002–1007.

Cross, L. H., & Frary, R. B. (1977). An empirical test of Lord's theoretical results regarding formula scoring of multiple choice tests. *Journal of Educational Measurement, 14*(4), 313–321.

Joint Committee on Standards for Educational and Psychological Testing of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.

Kolen, M. J., (1988). Traditional equating methodology. *Educational Measurement: Issues and Practice* 115.

Kolen, M. J., & Brennan, R. L. (1995). *Test equating*. New York: Springer.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121–1134.

McHarg, J., Bradley, P., Chamberlain, S., Ricketts, C., Searle, J., & McLachlan, J. C. (2005). Assessment of progress tests. *Medical Education, 39*, 221–227.

Muijtjens, A. M. M., Schuwirth, L. W. T., Coen-Schotanus, J., Thoben, A. J. N. M., & van der Vleuten, C. P. M. (2008). Benchmarking by cross-institutional comparison of student achievement in a progress test. *Medical Education, 42*, 82–88.

Muijtjens, A. M. M., van Mameren, H., Hoogenboom, R. J. I., Evers, J. L. H., & van der Vleuten, C. P. M. (1999). The effect of a 'don't know' option on test scores: number-right and formula scoring compared. *Medical Education, 33*, 267–275.

Norman, G. R., Neville, A., Blake, J. M., & Mueller, B. (2010a). Assessment steers learning down the right road: Impact of progress testing on licensing examination performance. *Medical Teacher, 32*, 496–499.

Norman, G., Neville, A., Blake, J. M., & Mueller, B. (2010b). Assessment steers learning down the right road: Impact of progress testing on licensing examination performance. *Medical Teacher, 32*(6), 496–499.

Personal communication with Louis Arnold, August 31, 2007.

Rademakers, J., Ten Cate, T. J., & Bär, P. R. (2005). Progress testing with short answer questions. *Medical Teacher, 27*(7), 578–582.

Swanson, D. B., Holtzman, K. Z., & Bulter, A. (2010). Cumulative achievement testing: Progress testing in reverse. *Medical Teacher, 32*(6), 516–520.

Traub, R. E., Hambleton, R. K., & Singh, B. (1969). Effects of promised reward and threatened penalty on performance of a multiple-choice vocabulary test. *Educational and Psychological Measurement, 29*, 847–861.

Van der Vleuten, C. P., Schuwirth, L. W., Muijtjens, A. M., Thoben, A. J., Cohen-Schotanus, J., & van Boven, C. P. (2004). Cross institutional collaboration in assessment: A case on progress testing. *Medical Teacher, 26*(8), 719–725.

Van der Vleuten, C. P. M., Verwijnen, G. M., & Wijnen, H. F. W. (1996). Fifteen years of experience with progress testing in a problem-based learning curriculum. *Medical Teacher, 18*, 102–109.

Verhoeven, B. H., Snellen-Balendong, H. A., Hay, I. T., Boon, J. M., van der Linde, M. J., Blitz-Lindeque, J. J., et al. (2005). The versatility of progress testing assessed in an international context: A start for benchmarking global standardization? *Medical Teacher, 27*(6), 514–520.

Verhoeven, B. H., Van der Steeg, A. F. W., Scherpbier, A. J. J. A., Muijtjens, A. M. M., Verwijnen, G. M., & van der Vleuten, C. P. M. (1999). Reliability and credibility of an Angoff standard setting procedure in progress testing using recent graduates as judges. *Medical Education, 33*, 832–837.

Willoughby, T. L., Dimond, E. G., & Smull, N. W. (1977). Correlation of quarterly profile examination and national board of medical examiner scores. *Educational and Psychological Measurement, 37*, 445–449.

Willoughby, T. L., & Hutcheson, S. J. (1978). Edumetric validity of the quarterly profile examination. *Educational and Psychological Measurement, 38*, 1057–1061.

Woolliscroft, J. O., Swanson, D. B., Case, S. M., & Ripkey, D. R. (1995). Monitoring the effectiveness of the clinical curriculum: Use of a cross-clerkship exam to assess development of diagnostic skills. In A. I. Rothman & R. Cohen (Eds.), *Proceedings of the sixth Ottawa conference on medical education* (pp. 476–478). Toronto: University of Toronto Bookstore Custom Publishing.