Medical Licensure Testing

White Paper for the Assessment Review Task Force of the Medical Council of Canada

Draft, May 2011

Krista Breithaupt, PhD.

Table of Contents

Introduction

Public trust of professionals in medicine depends on the belief that those who are admitted to specialties (certification), and whose employment is governed by a licensing body, have met certain minimum requirements. Certified candidates are measured on a set of knowledge and skills that experts in the field deem essential. The minimum standard of proficiency protects the public by withholding licensing from medical practitioners who do not meet the required level of proficiency to practice. Successful demonstration of tested competencies (skills and knowledge) is required for certification and for licensure. In this discussion, the terms certification and licensure are sometimes used interchangeably in discussions of theory and practice for licensure of medical professionals.

In the *Standards for Educational and Psychological Testing* (AERA, APA & NCME, 1999; draft 2011) guidelines are available that describe the responsibilities of those who create and use tests for employment and credentialing. In the *Standards,* testing for the Licentiate of the Medical Council of Canada falls into the category defined as credentialing:

> "Testing used in the processes of licensure and certification, which will here generically be called credentialing, focuses on the applicant's current skill or competency in a specified domain...Although licensure typically involves provision of a credential for entry into an occupation, credentialing programs may exist at varying levels, from novice to expert in a given field. Certification is usually sought voluntarily, although occupations differ in the degree to which obtaining certification influences employability or advancement. Testing is commonly only a part of a credentialing process, which may also include other requirements, such as education or supervised experiences. The *Standards* apply to the use of tests as a component of the broader credentialing process." (AERA, APA & NCME joint committee, Chapter 11, p. 1).

Our discussion will use the terminology of credentialing when discussing the MCC Qualifying Examinations I and II.

Competent candidates must not be deprived of employment. Therefore, fairness to the public and to the candidate must be balanced when decisions are made using test scores. One way to demonstrate that decisions are fair is to examine evidence of appropriate test construction, validity and reliability (measurement error) in the examination process. Validity evidence will usually take into account the test construction and reliability as psychometric properties of scores. However, validity also requires consideration of the broader consequential issues arising from decisions based on test scores. This paper will discuss some particular issues in test construction, test reliability and validity of test scores that are critical for fairness in credentailing testing in medicine. A very brief summary of each of measurement principle will be followed by specific application to the certification of health practitioners.

Test Construction Guidelines

Measurement texts broadly outline the steps involved in constructing an achievement test (e.g. Crocker & Algina, 1985; Worthen, Borg & White, 1993). Guidelines for those who create and use test for employment and certification are available from our professional organizations (e.g. the draft *Standards for Educational and Psychological Testing*, AERA, APA & NCME, in review 2011; *Educational Measurement*, Brennan Ed., 2006) and also from experts in medical education (e.g. Ottawa 2010 Conference Consensus Statements on assessment by Norcini and Prideaux et al; *Medical Teacher,* 2011; *Handbook of Test Development*, Eds. Downing & Haladyna, 2006). While the vocabulary used by measurement experts and assessment professionals in medical education does differ, both groups recognize our legal responsibility for the decisions made using certification and credentialing examination results, and all of our guidelines discuss the basic principles of fairness, reliability and validity.

These three principles are built into the work of the testing program in a variety of ways. First it is necessary to determine which objectives the test is intended to measure, and the definitions for competencies or objectives must be specified. The objectives may be skills, abilities, or knowledge constructs. For example, the criteria for proficiency may be based on agreed benchmarks which represent the skills necessary to progress into advanced training or enter practice. In credentailing testing, these competencies must be developed in a transparent and disciplined process and approved by an appropriate policy body who has direct responsibility over the testing program. A variety of qualitative and quantitative methodologies are acceptable, and the maintenance of currency depends on periodic review of the required blueprint. Raymond & Neustel (in Downing & Haladyna, Eds., 2006) provide a comprehensive overview of methodology and theory used for determining content for credentialing examinations. In many task or practice analysis studies, a survey is piloted and administered to identify the frequency and importance of each activity. There are precise definitions of tasks, cognitive activities and information acted upon when a task analysis survey is constructed to determine what activities are important across a variety of settings.

Interim updates to content can be made with a routine monitoring of objectives within the broader defined blueprint. However, larger studies of the entire domain of knowledge and skill will be needed on a periodic schedule and are normally published and available in advance for a variety of stakeholders outside of the examination program itself. A reasonable convention for credentialing examinations is to conduct a practice analysis or blueprinting study every 5 to 7 years, or in response to important developments in the profession served by the examination. With the results of these studies, and an approved specification for the test, or blueprint, can be drawn that defines relative coverage of each of the objectives.

Next, it is necessary to select assessment formats that are appropriate to measure the abilities, skills and knowledge required for the test. Objectivity in scoring, as well as the time and

cost required for administration are important considerations in developing the performance measurements and experiences required by the test. Writing and assembling items according to the blueprint provides a pilot version of the test that can be administered in a calibration trial. These data can then be used to revise and trim items and revisit decisions such as the appropriate number of test questions and testing time required.

Many statistical techniques are available to determine if test items and test scores have desirable properties (e.g. item statistics, estimates of error in total scores, decision accuracy, dimensionality or latent factor analysis, scale score and item reliabilities). Theorists in educational measurement (e.g. Crocker and Algina, 1986) make a clear distinction between criterion-referenced tests (CRTs), and the methods used to develop and evaluate examinations for used to rank order candidates on ability, identify learning, or inform educational plans (norm-referenced tests, or NRT decisions). When the test scores are used for a mastery decision (e.g. there is a pass or fail associated with a cut score) criterion-referenced estimation of the item and test score properties is appropriate. Criterion-referenced interpretations are made when a selection decision is required (e.g. entry into practice or advanced training). The blueprinting effort and test construction process hinge on the clarity of the defined purpose of the examination, and the specific objectives in the competency domain.

Test Construction for Medical Licensure

Historically, certification has been granted by an organization of practicing professionals who developed, administered and scored their own tests. The development of guidelines for testing practices in psychology and education has led to wide acceptance of basic principles to govern tests for licensure and certification (Haladyna, 1987; Hambleton & Fennesy, 1993; Shimberg, 1981). Legalities of testing for certification have led to recognition of the need for defensible and valid tests (Cavanaugh, 1991). As a result there has been a trend towards national testing in the professions for credentialing, and large psychometric service organizations have developed to oversee much of the testing for professional proficiency.

The objectives of testing for proficiency in professional practice reflect the domain of skills and knowledge deemed necessary for protection of public safety. The titles granted by credentialing convey to the public that practitioners have met certain minimum requirements for proficiency, and control admission of candidates into employment in medicine (Jaeger, 1995). Guidelines such as the *Uniform Guidelines on Employee Selection Procedures* (U.S. Department of Justice, 1978) and the *Standards for Psychological and Educational Testing* (AERA, APA & NCME, 1985; 1999; 2011 draft) each delineate essential and desired characteristics for selection tests. In contrast to achievement testing that places individuals along a wide continuum of ability levels, an NRT use of the test score, credentialing examinations are designed solely to

distinguish between those who are minimally competent and candidates who are not (CRT use of test scores). Identification of the criteria necessary for safe and effective practice must be consistently applied, and consensual for the construction of fair tests.

Task Analysis

The content domain sampled by certification tests will be defined according to requirements for safe and effective practice. These criteria must be explicitly defended and fair to examinees. Requirements for proficiency may not be so stringent that competent practitioners are excluded from professions. Smith & Hambleton (1993) describe credentialing examinations as criterion-referenced tests (CRT's), because decisions resulting from the test are absolute and not relative (the latter are norm-referenced tests or NRT's). This CRT interpretation is critical, as it will drive the definition of the purpose of the licensing examination and must be adhered to in the implementation of the task analysis. The purpose of the examination (or criterion decision) will be the basis for determining an appropriate methodology for the task analysis or blueprinting study, and will be the basis for defining each knowledge, skill and objective statement that might be considered in test construction.

In order to establish what the specifications or criteria for building test forms, the test construction process is often very detailed (e.g. LaDuca, 1994). In this discussion, the terms blueprinting, task analysis or practice analysis are equivalent and refer to the process of defining what should be measured by the test. The blueprint is like a recipe for the set of competencies on the exam. These specifications must be defined clearly enough to guide question or task development and specific enough to support the construction of multiple equivalent forms of the test using different test questions and tasks.

The process of determining what competencies, knowledge and skills are needed for certification exams can involve qualitative and quantitative data collection, and usually requires many iterations. Comprehensive approaches may involve critical incident studies, exhaustive lists of competencies judged for frequency and importance, and several panels of expert judges (Raymond & Neustel, in Downing and Haladyna, 2006). Task analysis that relies on a database of activities that practitioners typically are required to perform may not include critical elements if these occur rarely. For this reason a survey of practitioner's activities may not be sufficient. Similarly, a medical data source such as incidence of presentations, index of diseases, or billing data will be inadequate in determining criticality of some knowledge or skills. The appropriateness and sufficiency of the test content is the primary concern in the blueprint production. Once this is determined, the competencies or objectives identified in the blueprinting or task analysis are explicitly linked to selection or coding of test items. Other aspects of test construction also bear directly on the validity of the test (e.g. item bias, form equating) and will be discussed later.

Item Formats

Many item types may be included in the test, and decisions about which format may be most appropriate normally have a practical and a theoretical basis. It has been argued that the multiple choice format is not adequate to measure complex reasoning, while it may be very efficient to administer and score (Shepard, 1991). Some examples of other kinds of measurable responses might be obtained from 'justified' multiple choice items, multiple selection tasks, true or false items, completion tasks, constructed responses, figural responses and more complex performances supported by computerized administration. With current technology, a wide variety of testing formats are available (Sierci & Zenisky, in Downing & Haladyna, 2006). Items need to be matched to the complexity of the knowledge construct or set of skills tested. For example, candidates may be tested on basic science knowledge using multiple choice tests, but more complex items are often preferred to determine clinical skills. Oral interviews may be used to assess affective or personality traits of examinees. Computerized tests can provide an environment where the examinee may decide which of several sources of information to use for making a treatment decision. The computer administration also allows variations which reflect the consequences of candidates' decisions.

Health practitioners work in a complex environment involving high stakes decisions based on human interactions. For this reason observed tasks, in the form of high-fidelity performance assessment has been common in medicine for several decades (e.g. simulated patient examinations, or objective structured clinical skills examinations). Candidates may be required to demonstrate communication, history taking, physical examination skills, treatment planning, and problem solving in a situation that closely resembles a clinical encounter (Hambleton & Fennesy, 1993; Ross et al., 1996). In all forms of performance assessment there is greater latitude for personal styles in arriving at solutions, at the same time it is more difficult to score these tasks objectively. Performance assessment of clinical skills can involve actors who pose as patients in simulated case rooms. These performances can involve presentation by an actor in a simulated case room, and a trained examiner who may use an objective rating scale or checklist to judge performances. Selection of any test format is made with consideration of the possible contribution of these performances to the reliability and validity of the entire examination process.

The statistical characteristics of many of these item types may be examined based on response data obtained. Classical measurement theory, and more recently, modern methods such as item response theory (IRT) have been used to determine the whether the test and items have desirable characteristics. IRT has been useful in the pre-calibrating new items that are added to an existing test and for comparing scores for individuals who take different forms of a test. IRT provides a family of mathematical models which place the responses to test items and the underling ability estimates for candidates on a common scale. This IRT scaling allows

different test forms to be pre-equated. As with any model for response data, it is necessary for certain assumptions to be met when using different IRT formulae (Smith & Hambleton, 1993). For example, unidimensionality and equal item difficulties are assumed when applying the one parameter logistic (Rasch) model. The basis for selecting a preferred scaling or scoring model will depend on important empirical, theoretical and practical considerations, including the kinds of test items and the service goals of the testing organization (e.g. immediate score reporting, detailed feedback, use of multiple forms, or computerized administration). Each item format and scoring model has some advantages and disadvantages which may contribute to validity of decisions made based on test scores.

Standard Setting

There are three general categories and multiple variations of methods used to determine cutoff scores for proficiency tests (e.g. Cizec, 1996; Crocker & Algina, 1985). Standard-setting techniques find the optimal cutoff score to minimize the risk of certification of unqualified candidates. Some analytic methods popular in the last half-century used judgments of item content. Analytic techniques attributed to Angoff (1971) and to Ebel (1972) require expert judges to determine what percentage minimally qualified candidates would answer each of a set of items correctly. Nedelsky's (1954) method takes into account the number of response options that this candidate would be able to eliminate in a multiple choice item. Several modifications of these methods have been developed also, that provide judges with several rounds of deliberation and discussion to arrive at a recommendation for the passing score. They may also include feedback on the actual examinee performance based on a previous cohort of test takers (e.g. Jaeger, 1995). Other analytic methods take into account the relative importance and difficulty of each task for the minimally qualified candidate (e.g. Nedelsky, 1954). In any standard setting, it is important to justify the selection of expert panels of judges, and also to provide training on the purpose of the exam and associated definition of proficiency. Standard setting is a subjective process, albeit informed by some data, and for this reason it is usually important to use several objective panels of judges to inform a policy decision on the passing score. This use of alternative panels is intended to mitigate possible biasing factors.

Performance-based techniques for setting cutoff scores use examinee responses to items and total scores from a representative sample of test takers. Two popular methods are the borderline method and the comparison groups method. The borderline method tests candidates who are expected to perform at about the level of the cutoff score, and a point (e.g. the mean score) is selected in that distribution of scores as the standard. The comparison groups method uses two distributions of scores from samples that have been previously identified as masters and non-masters (e.g. novice and experienced practitioners). The point of intersection of these two distributions is used to represent the cutoff score. In both of these performance-based

methods the selection of candidates is critical for the resulting standard. Performance based techniques can be improved if they are supplemented analytic or other judgmental methods to allow the policy decision to consider a wider range of information.

Holistic approaches to standard setting involve a panel of experts who make a decision based on the test as a whole. Judges are asked to decide what proportion of the total (correct) score a minimally qualified examinee would be capable of reaching, based on an appraisal of patterns of correct and incorrect responses. Some variations of this method require panelists to examine samples of responses to sets of questions for high, low and moderate proficiency examinees. Research has demonstrated that different standards are selected when each of these methods are used (Mills, 1983). This underlines the essential subjectivity involved in definition of criteria, selection of expert panels, training of judges, and selection of response and item samples (for performance based techniques). It is therefore commonly recommended that pilot analysis and hybrid methods be considered in setting the cutoff score for mastery decisions.

Reliability Theory

Two types of errors of measurement may be identified in examination of item and test scores, random or systematic. The first error is a source of bias in scores and an issue of validity, while the second is defined as measurement error and can be estimated in reliability studies. Reliability estimates allow the test designer to determine the possible size and sources of construct irrelevant variation in test scores. The usual assumption made is that the skill, trait or ability measured is a relatively stable defined quantity during testing. Therefore, variation in obtained scores is usually attributed to sources of error. Examples of methods used to isolate sources of unreliability include test-retest reliability (Pearson correlations) from two test administrations, and internal consistency (Cronbach's alpha) to estimate the amount of shared variance between responses to items from one test administration. Reviews of measurement reliability theory and methods can be found in Crocker & Algina (1985) and Worthen, Borg & White (1993). The reliability of decisions made with CRT's should also be examined for decision consistency (Swaminathan, Hambleton & Algina, 1974). Cohen's Kappa compares the classifications of one set of candidates made with the exam with classifications based on an external criterion. It is generally known that adequate reliability is a necessary but not sufficient condition for test validity. Some appropriate methods for assessing the reliability of licensing tests will be considered next.

Reliability of Medical Licensure Tests

Some problems arise when gathering evidence for the reliability of decisions made using scores from licensing examinations. These arise due to nature of the classification

decisions made, and due to the complexity of the proficiency constructs and items used to assess the required knowledge and skills. Common methods of acquiring reliability evidence (test-retest or agreement with a criterion test) are often not appropriate or feasible for certification examinations in medicine. Re-taking the examination is usually restricted to maintain security, and those who challenge the exam a second time are a subset of candidates who did not succeed in their first attempt. Any conclusions about re-test data is suspect as a result.

Because these tests are developed to protect the public, and not to predict success in practice, any criterion score comparison selected for decision consistency will be insufficient (e.g. success in residency, or client complaints in future practice). This is evident when we consider self-reported knowledge, supervisors' assessments, or other criterion information gained later in practice (e.g. publications, patient outcomes). The literature shows only minimal associations (low correlations) with scores obtained on certification tests (e.g. Meguerditchian, Bordage & Tamblyn, 2011). This is because certification test are designed to distinguish performances at the level of minimum proficiency, not performances among successful or unsuccessful candidates (Shimberg, 1981). Some other issues concerning traditional measures of reliability arise for licensing examinations and will be discussed here.

Re-testing for reliability analysis is also expected to result in very large practice effects which could be expected to lead to inflated statistical estimates. Additionally, as tests and items are considered secure and many versions of licensing exams must be produced it would be rare for identical tests to be administered in a short time period. Medical practitioners who are entering testing for certification or licensure usually comprise a very homogeneous group who has already been selected from a larger pool of applicants to medical schools. This homogeneity and similarity of ability will reduce (attenuate) correlations between equivalent tests. Finally, there is no incentive for those who pass the test to be re-tested, so a re-test design is infeasible for a variety of reasons.

Common measures of the consistency of responses across sets of test questions (homogeneity of response data), depend on the assumption of unidimensionality of the domain tested. This assumption is usually not strictly met by certification tests. Certification tests contain a variety of competencies, knowledge and skills which may be dissimilar (e.g. communication skill versus knowledge of therapeutics). While it may be possible to estimate consistency in sub-domains of a multi-faceted test, interpretations are difficult when responses to items depend on more than one ability. Some researchers suggest that performance assessment using standardized patient simulations violate both the unidimensionality and the stable trait assumptions that forms the basis of much of our reliability theory.

LaDuca (1994) described clinical encounters where responses to items are the result of a combination of six possible care environments and six physician tasks. Freidman (1991) argued that instability in performance is the result of the interplay of cases, raters and examinee

behaviors. Also, the number of performance assessments (cases) included for each administration of the test is limited by time and expense. This effectively reduces simple estimates of consistency by decreasing the relative amount of generic competencies represented across cases, in comparison to specific skill or knowledge for a given case. Ultimately, decisions made based on licensing test scores are dichotomous, and the consistency of classifications made using cases, or total scores is more important than homogeneity of items. Therefore, two kinds of reliability estimates are appropriate when examining licensing exam data.

Decision Consistency

Two kinds of errors can arise from misclassification using a test score, these are non-masters who reach the cutoff score (false positives) and masters who fail to reach the standard (false negatives). The first error is the most serious for certification exams due to the risk of exposing the public to incompetent practitioners. Jaeger (1995) suggests that the highest estimated misclassification rate be reported to reflect the purpose and high stakes of these testing decisions. Specifically, the risk of certification of unqualified practitioners must be minimized (e.g. <10%), even as the likelihood of false negatives may rise. Common indices of decision consistency only summarize the level of agreement, not the severity of each type of error (e.g. Cohen's kappa).

Analyses of score precision based on a single test administration are also available to estimate the reliability of CRT decisions (e.g. Jaeger, 1995; Livingston & Lewis, 1995). Jaeger (1995) describes methods that examine the standard errors around the cutoff score by using a split test. A cutoff score is determined for each half test, and the consistency of classifications made with each half is calculated. A variation might be used in medical credentialing to examine performances on standardized patient cases, using a cut-score for each case. Any statistical estimation method must be selected with consideration of mathematical model assumptions; sometimes these are difficult to meet with complex item formats. Usually, practical considerations will guide the choice of reliability indices.

Several methods for increasing the accuracy of mastery decisions have been described which can decrease the risk of false positive misclassification decisions for certification exams. Increasing the sensitivity of scores in the region of the cutoff will generally lead to greater accuracy in CRT decisions. This can be accomplished using IRT estimates for item properties, and maximizing reliability in the zone of the passing score. Similarly, test reliability specifications can be used to build and judge the quality of test forms using classical analysis of the item and form properties.

As mentioned above, some methods for minimizing classification errors include selecting items with difficulty values close to the standard set for the total test, and increasing the numbers of items with good discrimination between masters and non-masters (Hambleton & Fennesy,

1995). Generally, low statistical estimates for reliability are found for complex item types, such as constructed response items and performance assessments. These items may be desirable for proficiency examinations because they offer increased authenticity over traditional objectively scored item types. For this reason, it is often argued the classical models and interpretations for reliability are not appropriate. Freidman summarizes this problem: "For the sake of reliability, medical educators may be measuring complex behaviors using simple tools" (Freidman, 1991 p392). In the case where performances are judged by experts, a true estimate of error will be the extent to which different judges arrive at the same rating for a single examinee (inter-rater agreement). While some other common approaches to establishing reliability may not be useful for certification testing, decision consistency and inter-rater reliability are critical for CRT's that use constructed responses or performance items.

Reliability of Scoring Procedures

The most effective method of increasing agreement among raters who must judge performances is by clearly defining the scoring criteria and enhancing training for examiners. This may be accomplished in orientations where scorers review the purpose of the examination, discuss detailed proficiency criteria, observe example performances at and below the required proficiency, and should be operationalized during the task with clear ratings or checklists. Scoring may be standardized on a particular sample of examinees for NRT interpretations (Shimberg, 1981). However, this is not appropriate for CRT decisions where each examinee must be compared to a stable required performance level (and not to any sample of candidates).

In performance assessments, the component parts of the clinical interaction are often scored using objective checklists, and scorers may themselves be calibrated using trained physicians (e.g. Ross et al., 1996). Increasing the number of judges may also minimize inter-judge differences, but may not be feasible as a routine practice. Some methods have also been developed using IRT models to adjust for differences in judge leniency (e.g. Rasch modeling). More commonly, judges receive periodic review and re-training to encourage similar evaluations. In medical certification, the items or tasks must be representative of those critical performances that occur in clinical practice. The challenge of licensing testing lies in the attempt to make the tasks as meaningful as possible while obtaining desirable empirical test properties. In the next section, issues concerning the appropriateness of interpretations made using the test scores will be discussed as validity evidence. It might be important to note that the test development process, including blueprinting, standard-setting and evidence of reliability are important components of a cogent validity argument.

Validity Theory

Educational measurement has developed a unitary theory of validity in the last two decades (APA, 1985; Messick, 1993). This theory orders the types of validity evidence (previously viewed as separate) under the common umbrella of construct validity. The test score interpretations, and decisions made based on these scores must be supported by the content and the empirical properties of the test. Validation analyses touch on all aspects of test design, administration, scoring and test use. Validity studies extend the requirements of reliability analyses by taking into account competing hypotheses for systematic variation in test scores. Common evidence for test validity includes criterion-based methods such as convergent or divergent evidence, and requires demonstrations that scores are unbiased. Investigations of the design process and face validity analyses are also important sources of evidence for test validity.

Validity of Medical Licensure Tests

LaDuca (1994) argues that for certification examinations content validity is of primary concern. Certification examinations must be defensible in terms of manifest content, this differs somewhat from achievement testing that may measure latent abilities or skills indirectly. The problem of finding a relevant criterion (discussed earlier in the context of reliability) makes it difficult to provide convergent evidence of validity. For this reason, generalizability, content validity, freedom from bias, and consequential validity have each received emphasis in studies of certification testing.

Construct Irrelevant Variance

Generalizability theory (G-theory) makes use of an empirical analysis method based on separating the variance introduced by different aspects of the testing situation. G-theory is a broader perspective than simple reliability analyses, but the general assumptions are similar. The construct that is being measured is assumed to be stable, so that components of variation can be attributed to various other sources (Engelhard, 1992). Many facets in the assessment situation are identified so that irrelevant variation may be linked to raters, test forms, cases, competencies, or examinee characteristics. This type of analysis is well suited to detect possible biasing factors in the testing situation (e.g. Campbell & Fiske, 1959). Philips (1996) suggests the main threat to validity of certification tests is under-representation of the universe of content (e.g. fewer tasks can be included in clinical skills assessments with simulated patients). Kane (1994) argues that the correct approach to validation analyses is refutation of alternative hypotheses. G-theory analysis is one method of exploring whether known extraneous factors may influence test scores.

Although often used as a simple tool to discuss reliability of scores in medical education, in fact G-theory analyses is more appropriately applied to performance assessments to determine if test scores are generalizable (e.g. Boulet, 1996). In the credentialing examination context, tasks

are intended to be representative of the universe of critical activities of the practitioner. Because it is not possible to include all possible activities, the sample of tasks or items included must be defended based on the criteria for proficiency and the test blueprint. This external aspect of validation is critical evidence that these tests are sampling the domain of actual proficiency. G-theory provides a mathematical model where variation associated with error in scores can be attributed to raters, cases, or content or skill differences in the observed ratings. Other forms of validation analyses simply take the possible elements contributing to scores singly (ignoring interaction effects), and evaluate the degree to which the evidence supports the intended inferences from test scores or suggests bias due to error.

Content Validity

Several authors have described the necessary steps required to link exam scores to necessary skills, judgment and knowledge for medical practitioners (Jaeger, 1995; Kane, 1994; Philips, 1996; Smith & Hambleton, 1990). The general method begins with role delineation through exhaustive job analysis or blueprinting process where each listed proficiency is rated for importance to the proficiency decision. Independent verification of the representativeness and appropriateness of items to the criteria of proficiency can demonstrate content validity, but multiple lines of evidence are preferred (e.g. validation analysis based on examinee performance data). Periodic item bank reviews, job analyses, and other studies are often combined with evidence based on test scores, keeping in mind the limitations due to the mastery decision discussed previously.

It may be important to weight items according to test specifications, or to increase the number of items that represent important domains of skills and knowledge in order to ensure the score represents the intended constructs. Scoring procedures are usually linked to the blueprint development process, item types and usually require piloting to demonstrate reasonableness for the interpretations made. CRT examinations depend on precision at the cutscore, and normalizing transformations based on a reference cohort (z or t scoring) are usually not supportable:

> "[In an NRT context]…the standard used in interpreting test performance is a relative one and the score given to the examinee is called a norm-referenced measure. The second type of information is the degree to which the student has attained the goals of instruction...If the items on the test are chosen randomly from all possible items, the proportion-correct score can be considered an estimate of the proportion of the facts the student knows and can be interpreted without knowing how other examinees have performed on the test. In this example the proportion-correct score is a criterion-referenced measure." (Crocker & Algina, 1986, p 192).

However, it is usually important to develop a stable and consistent score for exams that may include different items, or different numbers of observations. For this reason, it may be necessary

to transform raw scores to an easily understandable scale, with a common cutscore for score reporting.

It might useful to note there is a distinction between the reported passing score and the underlying proficiency required to pass the examination, usually on a percent-correct scale determined by the standard setting. The proficiency required to pass may be mapped onto any reported score scale, with a few considerations in mind.

The appropriate transformation to a reporting scale must retain the properties of the raw score scale in addition to providing a common passing score across test administrations. For example, when raw scores are based on categorical judgments (master, non-master), an ordinal scale results. It is not reasonable to transform the raw scale in such a way that interval or ratio interpretations might be encouraged. For example, a z-score transformation not only locates the center of the sample of scores on a mean from a sample, but also uses the standard deviation of that sample to force linear normality on the reported score distribution. The use of various samples to determine a score reporting scale is questionable in the context of a CRT where competency is judged for mastery.

Shimberg (1981) suggests that linear scoring is often not realistic for credentialing tests, and a minimum standard also may be required for subsections of the test. In this case multiple criteria are considered as a requirement for passing (e.g. passing language proficiency exams and knowledge exams may be required for certification). It is also necessary to periodically re-evaluate the validity of cutoff scores to ensure that they relate to current professional roles. As new tests are constructed, changes occur in the role of the professional, or when additional skills or knowledge included in a licensing exam, empirical study and re-validation of the standard and score scales are necessary to preserve the appropriateness of decisions and test score validity.

Test and Item Fairness

Test and item bias, or fairness, is often conceptualized through evidence that no group of people are given any unfair advantage because of the examination process. Exploration of test bias may involve judgments of item content (e.g. language must be clear and free from irrelevant culturally specific terms), or comparisons of specific groups on items who have equivalent levels of ability. IRT can be used to demonstrate whether or not item difficulties are similar for groups of candidates who have the same proficiency level, but differ in other ways (e.g. gender, country of training, ethinicity, age). In spite of the importance of this area of analysis, these bias studies are somewhat limited by the ability of some testing programs to gather demographic data on test takers (Smith & Hambleton, 1990). Case law involving licensure examinations places responsibility with the testing program to provide evidence the results of the exam are not biased

against subgroups, and this has resulted in guidelines that require bias studies be conducted (AERA, APA & NCME, 1999 and 2011 Standards).

It is just as important to demonstrate that alternative forms of the test, or scores from different judges, yield response data with equivalent properties. One method of controlling possible biasing factors is to mitigate unintended systematic bias by standardizing testing conditions (e.g. through administration conditions, eligibility requirements, objective scoring, trained raters, or standardized patients). As these sources of error are decreased, test validity will improve. When biasing factors are controlled, alternative explanations for the resulting classification decisions can be excluded and the internal validity of test scores can then be defended. In the next section, we turn to external validity evidence based on the social and other consequences of test use.

Consequential Validity

Theoretical models of practice are critical for evidence for the external validity of the certification testing process. The criteria for the tests should be based on social constructions of the professional role, and not solely on cognitive or knowledge-based models of practice. The impact of decisions made using test scores on various stakeholders form the consequential aspects of validity as described by Messick (1989). Classification decisions based on the test must reflect this purpose. Fairness to examinees is maintained only when the content of proficiency testing is limited to content "...unimpeachably relevant to the field and essential for effective practice" (LaDuca, 1994, p183).

It has been suggested that task analysis of professional activities may not be sufficient basis for development of content for defensible certification testing. Rather, the ecology of the profession should determine the content of testing. Shepard (1991) recommends a model that looks at problem solving abilities of physicians as critical skills. This author suggests that examinations be constructed with multi-faceted items that reflect the situations encountered in clinical practice. Examples of theoretical frameworks in medical education would include the CANMEDs roles, and role-defined competencies that are beyond the traditional knowledge-based examinations for physicians. Hence, the popularity of performance assessments is supported due to the need to include relational competencies based on the evolving role of physicians in society. When it is difficult or infeasible to include some nuanced competencies in a standardized national licensing exam, it is usually sufficient to embed those competencies in the experience and education requirements for credentialing. Examples might include workshops with ethical dilemmas for jurisdictional licensure, supervised practice and feedback programs during residency, or continuing education requirements for maintenance of licensure. Sometimes assessment of specific skills and knowledge required for proficiency are rightly viewed as the

responsibility of the institution that provides accreditation and training to candidates. Therefore, the responsibilities for proficiency examinations will be defined with respect to the compensating roles of the entire credentialing process (medical education programs, specialty training programs etc.) in the specific jurisdiction.

Summary

This discussion has summarized issues of test development, reliability and validity in credentialing examinations for medical practice. Critical aspects included defensibility in test design and standard setting which must reflect the proficiency required for certification decisions. The purpose of licensure tests is to engender public trust by preventing incompetent practitioners from access to employment. Also, the examinations must be fair to candidates, so that qualified examinees are able to reach the standard. The social construction of the role of the health practitioner and the relevance of the content to minimum proficiency are important factors. Standard setting is the heart of certification testing, and some techniques of ensuring that cutoff scores are fair to candidates, and methods of establishing cutoff scores to minimize misclassification have been presented.

Principles of reliability and validity are the common guidelines for fairness in testing. Reliability of certification exams depends on the accurate isolation of sources of error in test scores (e.g. differences between scorers) and in mitigation of errors in classification decisions. Some methods of improving reliability for licensure testing were suggested which can reduce the possible contribution of extraneous factors to classification decisions. Validity expands on the principle of reliability, and raises issues of the generalizability of performance assessment and appropriateness of interpretations made from test scores. Content validity and consequential aspects of classification decisions were described as critical for certification testing. These related educational measurement principles are critical for fairness in testing for certification of medical practitioners.

References

Allan, R. E. (Ed.) The Oxford Dictionary of Current English. Oxford University Press: New York.

AERA, APA & NCME, (1999). Standards for Educational and Psychological Testing. American Educational Research Association, American Psychological Association, and National Committee on Measurement in Education. AERA Pub., DC.

Angola, W.H. (1971). Scales, Norms, and Equivalent Scores, in R.L. Thorndike (Ed.) Educational Measurement (2nd ed.). American Council on Education, Washington. 508-600.

Berk, R.A. (1976). Determination of optimal cutting scores in criterion-referenced measurement, Applied Measurement in Education, 1 215-222.

Boulet, J.R. (1996). Generalizability of performance assessments for physician proficiency. Examination Committee for Foreign Medical Graduates, Philadelphia. P.A.

Boyle, M.H. & Torrance, G.W. (1984). Developing multiattribute health indexes., Medical Care, 22, (11), 1045-1057.

Brennan, R., Ed., (2006). Educational Measurement, American Council on Education Pub.

Campbell, D.T, & Fiske, D.W. (1959). Convergent and discriminant validity in the multitrait-multimethod matrix, Psychological Bulletin, 56, 81-105.

Cavanaugh, S. H. (1991). Response to a legal challenge: five steps to defensible credentialing examinations. Evaluation and the Health Professions, 14, 1, 13-40.

Crocker, L. & Algina, J. (1986) Introduction to Modern and Classical Test Theory. Holt Rinehart & Winston Inc., Fla.

Cronbach, L.J. (1988) Validity, in Wainer, H. & Braun, H.I. (Eds.) Test Validity. Lawrence Erlbaum Associates, Inc., NJ.

Cizek, G.J. (1996). Setting passing scores. Educational Measurement: Issues and Practice. Summer, 20-31.

Davis, L.J. & Morse, R. M. (1987). Patient-spouse agreement on the drinking behaviors of alcoholics. Mayo Clinic Proceedings, 62, August.

Engelhard, G Jr. (1992). Historical views of invariance: evidence from measurement theories of Thorndike, Thurstone, and Rasch, Educational and Psychological Measurement, 52 275-291.

Ebel, R.L. (1972). Essentials of Educational Measurement. Prentice-Hall. New Jersey.

Freidman, M. (1991). Rethinking critical issues in performance assessment. Academic Medicine, 66, 7, 390-395.

Feldman, A.B., Haley, S.M. & Coryell, J. (1990). Concurrent and construct validity of the pediatric evaluation of disability inventory. Physical Therapy, 70, 10, 602-610.

Gavin, D.R., Ross, H.E. & Skinner, H.A. (1989). Diagnostic validity of the drug abuse screening test in the assessment of DSM-III drug disorders. British Journal of Addiction, 84, 301-307.

Haladyna, T.M. (1987). Three components in the establishment of a certification testing program. Evaluation and the Health Professions, 10, 2, 139-172.

Hambleton, R.K. (1984). Criterion-referenced measurement, in Husen & Postlethwaite (Eds.) International Encyclopedia of Education. Pergamon Press, NY.

Hambleton, R.K. & Jones, R.W. (1993). Comparison of classical test theory and item response theory and their applications to test development, Educational Measurement: Issues and Practice, fall.

Hambleton, R. K. & Fennessy, L.M. (1993). Technical advances in credentialing examination development. Modern Theories of Measurement: Problems and Issues.

Jaeger, R.M. (1982). An iterative structured judgment process for establishing standards on proficiency tests: theory and application. Educational Evaluation and Policy Analysis, 4, 461-476.

Jaeger, R. M. (1990). Establishing standards for teacher certification tests. Educational Measurement, Issues and Practices, 9, 4, 15-20.

Jaeger, R. M., (1995). Setting standards for complex performances: an iterative, judgmental policy-capturing strategy. Educational Measurement: Issues and Practice. Winter.

Kane, M. (1994), Validating the performance standards associated with passing scores, Review of Educational Measurement, 64, (3) 425-461.

Kane, M. (1982). The validity of licensure examinations, American Psychologist, 37, 8, 811-918.

Katz, N., Itzkovich, M., Averbuch, S. & Elazar, B., (1989). Loewenstein occupational therapy cognitive assessment (LOTCA) battery for brain-injured patients: reliability and validity. The American Journal of Occupational Therapy,43, 3, 184-192.

LaDuca, A. (1994). Validation of professional licensure examinations. Evaluation and the Health Professions, 17, 2 , 178-197.

Livingston, S.A. & Lewis, C. (1995). Estimating the consistency and accuracy of classifications based on test scores. Journal of Educational Measurement, 32, 2, 179-197.

McDowell, I. & Newell, C. (1987), Measuring Health: a Guide to Rating Scales and Questionnaires. Oxford University Press, NY.

Meguerditchian, A., Bordage, G., Tamblyn, R (2011). A Blueprint for Addressing Preventable Adverse Events and Medical Errors during Training, on Credentialing Examinations and in Professional Enhcancement. Technical report, Medical Council of Canada.

Messick, S. (1989). Meaning and values in test validation: the science and ethics of assessment, Educational Researcher, 18 (2) 5-11.

Messick, S. (1993) Validity in Linn, R.L. (Ed.) Educational Measurement. American Council on Education, Macmillan Inc., NY. 13-103.

Mills, C.N. (1983). A comparison of three methods of establishing cutoff scores on criterion-referenced tests. Journal of Educational Measurement, 20 , 283-292.

Mischke, H.D. & Venneri, R.L. (1987). Reliability and validity of the MAST, Mortimer-Filkins Questionnaire and CAGE in DWI assessment. Journal of Studies on Alcohol, 48, 5 492-501.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: implications for performance assessment, Review of Educational Research, 62 (3) 229-258.

Nedelsky, L. (1954). Absolute grading standards for objective tests. Educational and Psychological Measurement, 14, 3-19.

Norcini, J., Brownell, A., et al. (2011). Ottawa Conference Reports, Medical Teacher, 33 (11), 206-233.

Phillips, G., W. (1996). Technical Issues in Large Scale Performance Assessment (U.S Department of Education, Office of Educational Research for Improvement, NCES 96-802). National Centre for Educational Statistics.

Raymond, M. & Neutsel, S. (2006). Determining the Content of Credentialling Exams, in Handbook of Test Development, Downing & Haladya (Eds.) Lawrence Erlbaum and Associates, 181 – 224.

Ross, L. P., Clauser, M.J., Margolis, N.A. Orr, N.A., & Klass, D.J. (1996). An expert-judgment approach to setting standards for a standardized-patient examination. Academic Medicine, 71, 10, October supplement 4-6.

Shepard, L. A. (1991). Psychometricians'' beliefs about learning. Educational Researcher, 20, 7, 2-6.

Shimberg, B. (1981). Testing for licensure and certification, American Psychologist, 36 10, 1138-1146.

Sireci, S. & Zenisky, L. (2006). Innovative Item Formats in Computer-Based Testing: In Pursuit of Improved Construct Representation, in Handbook of Test Development, Downing & Haladya (Eds.) Lawrence Erlbaum and Associates, 181 – 224.

Smith, I. L., & Hambleton, R. K. (1990). Content validity studies of licensing examinations. Educational Measurement: Issues and Practice, Winter.

Streiner, D.L. & Norman, G.R. (1995). Health Measurement Scales: a practical guide to their development and use (2nd ed.). Oxford University Press, UK.

Streiner, D.L. (1993). A checklist for evaluating the usefulness of rating scales, Canadian Journal of Psychiatry, 38 March, 140-147.

Swaminathan, H., Hambleton, R. K. & Algina, J. (1974) Reliability of criterion-referenced tests: a decision theoretic formulation. Journal of Educational Measurement, 11, 263-268.

U.S. Department of Justice (1978). Uniform Guidelines for Employee Selection Procedures. Federal Register, August 25.

van Alphen, A, Halfens, R., Hasman, A., & Imbos, T. (1994). Likert or Rasch? Nothing is more applicable than good theory, Journal of Advanced Nursing, 20 196-201.

Ware, J.E. (1995). The status of health assessment 1994. Annual Review of Public Health, 16 327-354.

Worthen, B.R., Borg, W.R. & White, K. R. (1993). Measurement and Evaluation in the Schools. Longman Pub., New York.

Wright, J. G. & Feinstein, A. R. (1992). A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. Journal of Clinical Epidemiology, 45 (11) 1201-1218.

Yersin, b. Trisconi, Y., Paccaud, P., Gutzwiller, F. & Magnenat, P. (1989). Accuracy of the Michigan Alcoholism Screening Test for screening of alcoholism in patients of a medical department. Archives of Internal Medicine, 149, 2071-2074.