1995

7. Basic Psychometric Issues In Licensure Testing

Howard W. Stoker
University of Tennessee

James C. Impara
University of Nebraska-Lincoln, jimpara@unl.edu

7

BASIC PSYCHOMETRIC ISSUES IN LICENSURE TESTING

Howard W. Stoker

University of Tennessee

James C. Impara

University of Nebraska-Lincoln

INTRODUCTION

The number of people in the United States who carry some responsibility for the writing of examination questions and the construction of tests is unknown. In the *Preface* to *The Construction and Use of Achievement Examinations,* published by the American Council on Education in 1936, the authors indicated that the number probably exceeded a million. That number has certainly grown in the past 60 years. Questions are posed to students by teachers at all levels of education; the Armed Forces have people whose job it is to construct tests which are used in the promotion of personnel; over 1,000 occupations are regulated by the states and many, ranging from the professions to the trades, require licensure or certification (Brinegar, 1990). Many licensure and certification decisions are based on test performance.

Throughout the years, the types of test questions being used have changed, emphasis has changed from performance testing to multiple-choice testing and back to performance assessment. Apprenticeship programs in the trades—a kind of continuous assessment of performance—have been supplemented, or even replaced, by written examinations, or by a combination of written and performance tests. More recently, the use of technology in testing has begun to come into the

picture. For example, computer administration of questions, interactive video, and CD-ROM are beginning to be used.

Regardless of the type of test, whether it was written 50 years ago or last week, there are some important concerns. Fundamental among these concerns are the reliability and validity of the measures. The purpose of this chapter is to focus on the psychometric issues of reliability and validity of measures as they pertain to licensure examinations. In addition, the chapter focuses on the relationship of the measures to various guidelines—those of the Equal Employment Opportunity Commission (EEOC, 1975) and *The Standards for Educational and Psychological Testing*, produced by a joint committee of the American Educational Research Association (AERA), American Psychological Association (APA), and the National Council on Measurement in Education (NCME) and published by the APA (1985). (We will refer to the EEOC document as the *EEOC Guidelines* and the AERA, APA, and NCME document as the *Standards.*)

Frequent references are made to the reliability and validity of examinations when, in reality, it is the scores and the decisions made on the basis of the scores that are, or are not, reliable and valid. In the context of licensure, scores are used to make decisions. Statistical analysis may show that the scores possess properties indicative of reliability. Studies may be conducted to show that the measures have some type of validity. However, reliable and valid scores may be used inconsistently or incorrectly, and when this happens, the *decisions* made on the basis of the scores may not be reliable or valid decisions.

The discussion of reliability and validity in this chapter focuses on the traditional concepts of reliability and validity rather than on a more contemporary approach broadly called generalizability theory. Our reasons for the focus on the more traditional conceptsare simply that most licensure and certification programs with which we are familiar have not yet made the transition to generalizability theory as their basic approach to reporting the psychometric characteristics of their tests.

Reliability

Reliability has both a mathematical and a conceptual definition. The conceptual definition relates to the extent that a particular observed score (the score an examinee makes on a test) is a close approximation of the examinee's "true" score on that test. This concept is operationalized by thinking about testing some hypothetical examinee an infinite number of times and calculating the examinee's average score over all these occasions. That average is the examinee's "true" score. We assume, of course, that each testing occasion is independent of every other occasion. In a perfect world, we might find that this hypothetical examinee obtained the same score on every occasion. Under those conditions, the test would be perfectly reliable! In the real world, however, that would not likely be the case. Virtually all tests are unreliable to some degree.

No matter how hard we try, every licensure examination will produce scores that are less than perfect representations of a candidate's "true" score. Various factors contribute to the random errors that influence a candidate's actual score and make it different from the "true" score. Such factors are related to: the test (e.g.,

ambiguous items or directions); testing conditions (e.g., lighting, temperature, or other environmental factors that may be more or less similar to conditions on the job); and, the physical attributes of the candidates (e.g., high motivation or illness). All such factors contribute to the generation of random errors in scores that lead to the unreliability of the scores. The larger the number of these random errors, the smaller will be the likelihood that a candidate's score has sufficient levels of reliability.

Our concerns with reliability are twofold. First, reliability is somewhat a technical concern. There are actions that can be taken to enhance score reliability. Second, reliability is a precondition for validity. Scores that are unreliable cannot be valid! Although this can be demonstrated mathematically, it is also logical. If you stood on a scale that showed a weight of 170 pounds, stepped off and back on and the weight shown was 150 pounds, which weight is trustworthy? Neither! If a different scale showed similar weights (e.g., 170, 169), then you may have confidence that the second scale is measuring your weight appropriately and in a consistent manner. Any inference you might want to make about your weight would be made more confidently with the consistent scale than it would with the inconsistent scale. If you wanted to make a decision about the effectiveness of your weight reduction program, using the first scale would be difficult, whereas the data from the second would provide more confidence in the decision.

If one cannot rely on the test scores as accurate representations of the behavior being measured (reliability), then no amount of statistical manipulation of the numbers will lead to good decisions (validity). Not too many years ago, a "good mechanic" listened to the noises a car was making and made decisions about what was wrong with the car. Now, the car is hooked up to a diagnostic machine, operated by a technician (possibly a mechanic), that identifies which "chip" is malfunctioning. It can be hoped that more reliable measures are being obtained from the machines than were obtained from the "good mechanic." More importantly, we hope the decisions made about what is wrong with the car are more valid—they certainly are more expensive!

A fact we must face is that we have not developed any diagnostic machine for constructing licensure examinations and making licensure decisions. A few programs may be using more sophisticated test administration procedures (e.g., computerized testing, interactive videos), but these procedures do not assure more reliable scores nor more valid decisions. Various guidelines and standards have identified the areas of concern, relative to reliability, validity, and safeguarding the public, but have produced no machine or magic formulas for us.

Adequate control of random errors can be maintained through careful construction of the licensure examination. Such control will do much to insure that qualified candidates will be granted licenses and the unqualified ones will be screened out. The guidelines and standards insist on such control for the purpose of protecting the public from unqualified practitioners. Such control will also help insure that candidates are treated equitably and that decisions are not capricious.

There are many sets of guidelines for constructing examinations, whether they are licensure examinations or examinations to be used for other purposes.

This book offers suggestions for developing a variety of types of items. Textbooks, mainly in the field of educational measurement, contain lists designed to guide one in the development of examinations, their administration and scoring, and the setting of cut scores (the scores used to make decisions). The construction of a licensure examination that will yield reliable scores and lead to valid decisions is a long and arduous process—not one to be taken lightly. The processes by which items may be developed are described in chapters 5 and 6 of this book and elsewhere. Any licensure board involved in test development should consider whether the examination should be constructed under the direction of testing professionals employed by the board or by consultants who are testing professionals.

Once the initial development of an examination is complete (i.e., decisions about individual items have been made), a tryout is generally scheduled. The purpose of the tryout is to obtain data to estimate score reliability and, perhaps, make preliminary decisions related to cut scores. The tryout data should be collected from a group that resembles the candidates for licensure as closely as can be managed.

In licensure examinations where the number of candidates is very small (e.g., polygraph examiner, embalmer), tryouts may be difficult, if not impossible, to arrange, due to the small number of candidates involved. Hence, it may be necessary to wait until the first administration of the examination to obtain such data. If no pretest is feasible, then careful test development plays an even more important role. The implications this situation has for decision making are discussed later.

Professional Guidelines

The *EEOC Guidelines*, (EEOC, 1975) and *Principles for the Validation and Use of Personnel Selection Procedures* (Society for Industrial and Organizational Psychology [SIOP], 1987) focus on the validity of measures and decisions for employment tests. Both documents represent the basics of good practice, but both are directed toward tests for employment rather than licensure tests. The relationship between these two different purposes is discussed in chapter 2. The *EEOC Guidelines* reference extensively the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1985).

The *Standards* make direct reference to tests used for both licensure and certification, along with other types of test uses. Explicit in the *Standards* is guidance pertaining to the reliability of tests and the use of the standard error of measurement in the interpretation of individual scores.

> Fundamental to the proper evaluation of a test are the identification of major sources of measurement error, the size of the errors resulting from these sources, the indication of the degree of reliability to be expected between pairs of scores under particular circumstances, and the generalizability of results across items, forms, raters, administrations, and other measurement facets.
>
> Typically, test developers and publishers have primary responsibility for obtaining and reporting evidence concerning reliability and errors of measurement adequate

> for the intended uses. The typical user generally will not conduct separate reliability studies. Users do have a responsibility, however, to determine that the available information regarding reliability and errors of measurement is relevant to their intended uses and interpretations and, in the absence of such information, to provide the necessary evidence.
>
> Reliability coefficient is a generic term. Different reliability coefficients and estimates of components of measurement error can be based on various types of evidence; each type of evidence suggests a different meaning. (AERA, APA, NCME, 1985, p. 19)

It is the responsibility of the licensure board to direct the test developer to obtain the types of reliability estimates most appropriate for the licensure examination. If internal consistency estimates are desired, then a single administration may be all that is necessary, but if either reliability estimates that reflect equivalence (of alternate forms) or stability are appropriate, then two separate test administrations will be needed. These different types of reliability estimates are described in more detail below. Moreover, because of the nature of the decision made on the basis of the test, decision-consistency reliability may be the paramount reliability concern.

Reliability Indices for Test Scores

Internal Consistency, sometimes referred to as *homogeneity*, is the easiest method one can use to estimate reliability. This coefficient estimates the degree to which items are contributing to a common underlying construct. It requires only a single administration of one set of items to a group of candidates. Several methods exist to estimate reliability from a single administration of an examination. Coefficient alpha (Cronbach, 1951), or the less general KR-20, are the most common methods. If one is using a "packaged" scoring program for multiple-choice tests, there is a high probability that one, or both, of these values will be generated as by-products of the scoring process. (Some programs may be using a method called split-half. We do not recommend this method. For most purposes it is obsolete and the result is potentially biased depending on how the decision is made to determine the two halves of the test.)

Coefficient alpha can be used to estimate reliability, no matter what type of items are on the test. When only dichotomous items are included (items scored right or wrong) KR-20 and coefficient alpha are the same. Formulas for calculating these coefficients of reliability can be found in almost any basic measurement text.

Stability is estimated by administering a single set of items to the same group of candidates at different times. The correlation between the two sets of scores is the reliability estimate. Most measurement texts refer to this method as test-retest. The lapse of time between the two administrations will, of course, have an impact on the obtained correlation. Hence, when reporting the reliability estimate, it is necessary to describe the group used to obtain it, and the time interval between the testings. A different coefficient for every time interval is expected. Generally, the interval should be kept short, probably less than a week if possible, to minimize any

differential learning or forgetting that might occur during the interval, and long enough to allow the candidates to "forget" how they answered an item the first time.

Equivalence, usually called *alternate forms* reliability, calls for two tests, designed and constructed to be essentially equivalent in their psychometric characteristics and to measure the same skills. As with the test-retest method, the reliability coefficient is the correlation coefficient computed between the scores of one group of examinees on the two tests. A counterbalanced administration is recommended. This means that one-half of the candidates take Form 1 first and the other one-half of the candidates take Form 2 first. The order of testing is reversed in the second administration. The time interval between the first and second administrations should be as short as possible. If several days pass between administrations, the obtained correlation between the test scores could be used as both an estimate of equivalence and of stability.

The *Standards* call for full reporting of data from the administrations of both tests—means and standard deviations, along with errors of measurement and the estimate of alternate forms reliability. In addition, the rationale for selecting the particular time interval should be reported.

How to choose? Whichever method is selected to estimate reliability and calculate the standard error of measurement will depend on several factors. As noted above, *internal consistency* calls for one test and one administration of the test. Hence, that method will produce the quickest results. Because of its ease of computation and because the information provided is useful, some internal consistency measure should be computed each time the test is administered. If the one test form could be administered to the same group at two different times, a coefficient of *stability* could be calculated, in addition to the internal consistency estimates for each administration. This would be preferable to a single administration of the test, but this is difficult to undertake in licensure testing.

We recommend the development of two equivalent forms of the licensure examination. The second form will be needed, eventually, for matters of security and to prevent candidates who repeat the test from "learning items" instead of learning the subject matter. We also recommend that item development be a continuing process. New items can be embedded in test forms and "banked" for later use. Most commercial test publishers use this process for test development.

The number of computer-based programs for storing items and constructing tests is large. A few years ago, one needed a large capacity computer to build tests using computer technology. Now, adequate programs can be purchased for virtually any desktop computer. In chapter 8, a full discussion of item-banking is provided.

As noted above, for almost every examination, an internal consistency estimate of reliability (either coefficient alpha or KR-20) should be calculated. The notable exception is any examination that has a speed factor. In the typical speeded test, the candidate's score is largely dependent on the number of items attempted, rather than on the candidate's range of knowledge. (This is generally not the case in licensure examinations, but forewarned is forearmed.) For speeded tests, alternate forms or test-retest are the only appropriate alternatives for estimating score reliability.

Regardless of the method, a coefficient of reliability is essential. This number will reflect (for the group tested, under stated conditions, etc.) a measure of the random error associated with the scores. A symbol used to represent the reliability estimate is "r_{tt}." Because of the different methods of estimating reliability and because reliability estimates vary across different samples, the reliability estimate alone is not sufficient as a way to characterize or interpret measurement error.

Standard Error of Measurement is another way to represent measurement error. It is computed by using the reliability estimate, r_{tt}, and the standard deviation of the test scores (S_Y):

$$SE_M = S_Y\sqrt{1 - r_{tt}}$$

The standard error of measurement, SE_M, as calculated by this formula, is the average error associated with individual test scores across the range of scores in the distribution. This value is most useful when interpreting individual scores. Because licensure examinations focus on individual scores, careful attention must be given to the standard error of measurement.

Two characteristics of the standard error of measurement are important. First, although reliability estimates will vary with the samples used to estimate them, the standard error tends not to fluctuate as widely. For example, suppose a licensure test was administered to a large sample that had a wide range of scores. The reliability estimate might be high (r_{tt} = .94) and the standard deviation might be 8 score points. The error of measurement, $SE_M = S_Y\sqrt{1 - r_{tt}}$, would be:

$$SE_M = 8\sqrt{1-.94} = 1.96$$

If another sample was more homogeneous, the reliability estimate for that group might be reduced to .85 and the standard deviation would be lower (e.g., 5), resulting in a standard error of measurement of:

$$SE_M = 5\sqrt{1-.85} = 1.94$$

This illustration makes two important points: As group homogeneity increases, the reliability estimate will tend to be reduced. This does not mean the test is less reliable for the second group, it is simply a function of the way reliability estimates are calculated; and, even though the reliability estimates differ across groups, the standard errors of measurement are nearly the same.

Second, although the standard error of measurement is interpreted as though it is constant throughout the score distribution, this interpretation has been shown to be false. The standard error is usually largest for high and low scores and at a minimum near the mean of the score distribution. It is extremely important in licensure examinations to know the standard error of measurement at the cut score, the score used to decide if a candidate is to be licensed. Setting the cut score at, or near, the mean of the scores (setting cut scores is discussed in chapter 10) will reduce the number of incorrect decisions that are due solely to measurement error. Note that setting the cut score near the mean does not imply that only half of the candidates will be licensed. It is likely that the score distribution will be skewed and, hence, more or fewer than 50% of the candidates will be licensed.

A formula for estimating the error of measurement at a particular score (e.g., the cut score), is:

$$E_{cs} = SE_M \sqrt{1 + \frac{1}{N} + \frac{(T' - \overline{T})^2}{\sum t'^2}}$$

Where: E_{cs} is the standard error of the score of interest;
SE_M is the standard error of measurement;
N is the number of examinees tested;
T′ is the estimated true score associated with the desired observed score. T′ is estimated by:

$$T' = r_{tt}(X - \overline{X}') + \overline{X}';$$

$\overline{T}' = \overline{X}'$ is the estimated true score mean. The estimated true score mean is equal to the observed score mean; and
$\Sigma t'^2$ is the sum of the deviation scores of the distribution of estimated true scores (i.e., all T′- $\overline{T}'$ scores).

Decision-Consistency Reliability

Decision reliability is related to the consistency of a decision for licensure; the decision to withhold or grant a license when there is a specified decision rule (e. g., pass candidates with scores greater than some cut score). This is conceptually similar to the reliability of scores, but in the case of licensure, the decision "score" can be thought of as either zero (withhold license) or one (grant a license). Estimating decision reliability takes place following test development, test administration, cut-score determination, and estimation of score reliability. As in estimating score reliability, estimating decision reliability may occur after a single administration of the test, after repeated administrations of the same test form, or after administering alternate forms of the test to the same group of examinees.

Feldt and Brennan (1989) summarized techniques for estimating reliability for criterion-referenced interpretations as in licensure or certification. They describe two squared-error loss methods (those proposed by Livingston and by Brennon and Kane) and four threshold loss methods (those proposed by Cohen, Huyhn, Subkoviak, and Raju). The squared-error loss methods consider "error" to be the distance between an individual's observed score and the cut score. The formulas take into account both measurement error and classification error. The squared-error loss methods require only a single administration of the test. Livingston's coefficient results in decision-consistency estimates that can be interpreted in the same way as coefficient alpha and KR-20 Brennan and Kane's index of generalizability can be interpreted like KR-21 (an estimate of KR-20). Depending on the location of the cut score relative to the mean of the test, coefficient alpha (and KR-20) and KR-21 will be lower limits for the the respective estimates of decision-consistency reliability. (If the cut score is equal to the mean, then the computations will result in the same values as would be obtained with coefficient alpha/KR-20 or KR-21, respectively.)

For the squared-error loss method we recommend using Livingston's (1972) formula, represented as follows:

$$k^2 = \frac{\frac{I}{I-1}[V_X - \Sigma V_i] + (\overline{X} - C)^2}{V_X + (\overline{X} - C)^2}$$

Where: I is the number of items;
$\overline{X}$ is the mean score for all individuals.
C is the cut score;
V_x is the total score variance; and
V_i is the variance of an item.

Feldt and Brennan (1989) indicate that the threshold loss methods take into account only classification errors and assume that any misclassification is equally serious. They also note that there are methods other than those they discussed and that some of these other methods permit differential weighting of misclassification errors. These other methods are, computationally, quite complex. Early strategies for the threshold loss methods required two administrations of the test. The two dominant methods are a simple coefficient of agreement (the proportion of individuals classified the same way after two administrations of the test) and coefficient kappa (Cohen, 1960).

Because of the opportunity to compute both squared-error and threshold loss coefficients, we believe the optimal determination of decision reliability occurs when scores are available from two administrations (or two forms) of the examination. However, as noted, test-retest and alternate forms administrations are often difficult to arrange in licensure examinations.

For a test-retest or alternate forms situation, we recommend the kappa threshold loss method for estimating decision consistency. For all practical purposes, kappa represents an index of the proportion of agreement of assignment to the license and fail-to-license categories, beyond that expected by chance.

For example, Table 1 illustrates the results which might arise from two administrations of a licensure examination to a single group of candidates.

		First Administration		
		License	Fail to License	
Second Administration	License	75	5	80
	Fail to License	10	10	20
		85	15	100

Table 1. Classifications Resulting from Two Administrations of a Licensing Examination

In this illustration, 75% of the applicants would be licensed based on the scores earned on both tests; 10% would not be licensed by both tests. Hence the proportion of agreement is

$$P_o = .75 + .10 = .85$$

To calculate the proportion of agreement to be expected by chance, marginal totals are used

$$P_c = (.85 \times .80) + (.15 \times .20)$$
$$= .68 + .03 = .71$$

Kappa, then, is an index of the proportion of agreement over and above what might be expected by chance.

$$kappa = \frac{P_o - P_c}{1 - P_c}$$

In this example:

$$kappa = \frac{85 - .71}{1.00 - .71} = .48$$

In general, kappa ranges from zero to one, with the higher values indicating higher agreement. Negative values are possible, indicating "less than chance" agreement, but are probably not interpretable (Huynh, 1976). As the cut score deviates from the mean score, measurement error tends to increase, which would lead to a decrease in kappa. According to Linn (1979), "kappa tends to be lower for criterion scores near the extremes, to increase with test length, and to increase with test variability (p. 100)."

Kappa has some clear limitations that condition its use, especially when the cut score deviates from the mean and when the distribution of passing and failing candidates is highly skewed. Although the theoretical range of kappa is zero to one, the maximum value of kappa depends on the specific marginal values associated with any particular set of data. If the scores represent the most extreme values (all candidates pass or all fail), then although the proportion of agreement is 1.00, kappa cannot be computed (it is undefined because the formula results in dividing zero by zero). In essense, kappa is interpreted as an index that represents the proportion of consistent decisons beyond that expected to occur under conditions of chance (Subkoviak, 1980).

One advantage of using kappa is that it can be calculated in situations where there are more than two decision categories. For example, licensure may be a multiple stage testing situation (i.e., obtaining a "passing score" or a "borderline" score on one test, prior to taking a second test). The passing score for the first test might be the passing score, plus one standard error of measurement, calculated at the cut score. This criterion would set up a three-level condition: clear fail (e.g., scores more than the cut score minus one SE_M), borderline (e.g., scores between plus and minus one SE_M around the cut score), and pass (e.g., scores greater than the cut score plus one SE_M). Such a strategy would reduce the number of

candidates incorrectly classified as failing the first examination and give them a second chance at becoming licensed. The second test would also have the same three score categories (of course, this would require some policy decision for dealing with candidates who were borderline on one or both of the tests). For such a situation, coefficient kappa would be quite appropriate for estimating decision consistency of either test or for the combined effects of both tests.

Methods for use when only one score is available (i.e., when the test has been administered only one time) are somewhat complex, computationally. One method, attributed to Subkoviak (1976), is easier to use than the Huynh method mentioned below, but it is still computationally complex. An individual's true score is estimated using one of two methods, and then the probability of that score being above/below the cut score is calculated for the actual test and for a hypothetical parallel test. The resulting coefficient would, of course, depend on the selection of the cut score. A disadvantage of this method is that it tends to be biased for short tests (in that case, it underestimates the level of agreement when cut scores are near the center of the distribution and overestimates the level of agreement when cut scores are near the extremes [Subkoviak, 1980]).

The Huynh (1976) model is based on kappa, and is much more computationally complex. If an examination has more than 10 items, as one would expect in a licensure examination, simpler methods can be used to approximate the calculations (Huynh, 1976). The calculations yield a number between zero and one, representing decision agreement based on the test administered and a hypothetical parallel test. The magnitude of the index depends on test length, the variability of the test scores, and the cut score. This method also tends to produce biased estimates of the level of agreement, but unlike Subkoviak's method, Huynh's method tends to underestimate the level of agreement throughout the distribution when the test is short (Subkoviak, 1980). This conservative approach may be justified in licensure testing.

In some licensure contexts multiple tests are used (either collectively as a total score or sequentially as in multi-stage testing). In these situations, the estimation of the reliability of the decision is not a straight-forward procedure (Raju, 1982).

Additional Reliability Issues

There are two additional reliability issues to be discussed. The first issue is related to the condition when two or more tests (or subtests) are used to make the licensure decision. The second issue is when the licensure decision is based wholly, or in part, on ratings other than (or in addition to) test scores.

Two or More (Sub)tests

The above discussion has assumed that licensure decisions rest solely on the score from a single test. Although this is true for many areas, some procedures include more than one test. The medical, dental, and legal professions have multiple examination procedures, as do CPAs and Certified Professional Secretaries, among others. In such situations, the licensure decision could be made by finding a total score across all (sub)tests—called a compensatory model; attaining

a minimum score on each test—called a conjunctive model, or some combination of those options—a disjunctive model. The disjunctive model, we think, has few applications in making licensure decisions, but it may have utility for certification decisions. Thus, when there are multiple tests or subtests used in the licensure testing situation, there are serious implications for the way in which the cut score(s) are set.

Estimating score reliabilities when there are multiple (sub)tests is difficult, because the unidimensionality assumption in the calculation of coefficient alpha and KR-20 is typically violated. Test-retest or alternate forms would be the preferred methods in these cases. A procedure for estimating the reliability of the total score from a single administration is a stratified coefficient alpha in which the total score consists of the sum of the subtest scores. The reliability of such a composite can be estimated by:

$$r_{tt} = 1 - \frac{\Sigma V_i (1 - r_{kk})}{V_t}$$

Where: r_{tt} = reliability of the composite;
r_{kk} = reliability of a subtest;
V_i = variance of subtest i; and
V_t = variance of the total score.

Reliability of Ratings

In many licensure situations, there is a performance or clinical component that is scored by judges' ratings. Measures that rely on human judgment for scoring usually have lower score reliability. The licensing agency must assume responsibility for establishing procedures that maximize the reliability of the judgment scores. Some discussion of the methods for examining reliability and for enhancing reliability are discussed in chapter 6 of this book. A summary of that discussion follows.

To enhance the reliability of ratings, the most critical factor is the training of the observers, scorers, and/or judges. Check lists, rating scales, etc., can help ensure that all the raters are looking for the same thing and, hence, increase interrater reliability. Another factor in enhancing the reliability of ratings is the use of multiple judges, with either a requirement that judges/raters agree on pass/fail decisions, or, if that is deemed too rigid, an averaging of ratings may be used. The need for multiple judges to increase the reliability of ratings is exemplified by the judging of athletic competitions, such as diving, synchronized swimming, gymnastics, etc. At a local meet, two or three judges may be used. As the competition moves to district, state, and national levels the number of judges increases and, in Olympic competition, up to eight judges may be present.

Intense training of judges and the use of multiple judges correspond to the two dimensions of reliability discussed by Ebel (1951). In this landmark discussion, Ebel provides rationale and statistical formulas for estimating the reliability of individual ratings or of average ratings. He suggests that if the decision is made on the average score across a number of judges, then the reliability of the average rating is needed. If, however, the judgment is made by judges working individually,

across a number of examinees, then the reliability of individual ratings is appropriate. He argues strongly for the computation of an intraclass correlation to estimate reliability and he also provides formulas for the computation of a coefficient when there are missing data. Many of the formulas Ebel demonstrates are consistent with newer applications of generalizability theory being advocated in estimating the reliability of ratings. Additional discussion of the problem and methods of estimating reliability of raters may be found in Feldt and Brennan (1989).

Validity

About 40 years ago, validity was well defined and understood. There was *content* validity—earlier called face validity—which was necessary to show that the tasks in a test were representative of some domain. *Predictive* validity was needed to show the relationship between performance on the test and some later performance. *Concurrent* validity called for a correlation between the test scores and criterion performance obtained at about the same time. In some measurement texts, predictive and concurrent validity were subsumed under statistical validity. Finally, there was *construct* validity, which called for a conceptual framework, frequently implying some underlying trait and usually considered to be the responsibility of researchers. Licensure examinations relied heavily, if not entirely, on content validity.

About 30 years ago, predictive and concurrent validity merged into *criterion-related* validity. The criterion could exist along some time continuum, but the idea was that there be a relationship between the test scores and some criterion. About 25 years ago, two other types of validity were introduced, largely as a result of court challenges to the use of test scores in making pass or fail decisions about high school students. These two types of validity were called *instructional* validity and *curricular* validity (McClung, 1978). Instructional validity "is the actual measure of whether the schools are providing students with instruction in the knowledge and skills measured by the test" (McClung, 1978, p. 397). Curricular validity is "an actual measure of how well test items measure the objectives of the curriculum" (McClung, 1978, p. 397).

The *Standards* (AERA, APA, & NCME 1985) state, "Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. A variety of inferences may be made from scores produced by a given test and there are many ways of accumulating evidence to support any particular inference... The *inferences* [italics added] regarding specific uses of a test are validated, not the test itself" (p.9).

The *Standards* add, "Traditionally, the various means of accumulating validity evidence have been grouped into categories called *content-related, criterion-related, and construct-related evidence of validity.* These categories are convenient...but the use of the category labels does not imply that there are distinct types of validity... Evidence identified usually with the criterion-related...categories, for example, is relevant also to the construct-related category."

The consensus today seems to be that validity is a *unitary* concept, and that all evidence to be collected is a part of construct validation. For those who may be interested in the changes in emphasis in test validation, Messick's chapter in *Educational Measurement* (1989), Geisinger's article in *Educational Psychology* (1992), and Shepard's chapter in *The Review of Research in Education* (1993) are highly recommended.

Shepard's proposal (Shepard, 1993) "is that validity evaluations be organized in response to the question, 'What does the testing practice claim to do?'" (p.429). Applying this question to licensure examinations, the primary claims to be considered are: Is the test designed and developed to identify candidates who possess the entry-level knowledge and skills sufficient for licensure? And, does passing the test insure that the public will be protected from incompetent candidates?

The first claim, test design, falls into the areas commonly referred to as content validity. The licensing agency would start with a job, or practice, analysis from which is derived statements of purpose and, perhaps, a listing of objectives, knowledge, or skills that candidates are expected to attain, or display. Following this would be the establishment of what is usually referred to as a test blueprint. The test blueprint will include the domain of knowledge and skills to be sampled and the types of responses candidates will be asked to make (responses to multiple-choice items, constructed responses, performance, etc.). This process is described in some detail in chapters 5 and 6.

The job analysis may indicate the need for some general knowledge and skills all candidates should possess. If a carpenter is to read the plans for a house and estimate the cost of materials, certain reading and mathematics skills will be required (although, with today's emphasis on "precut" homes, the level of these skills may be lower than before). In any case, the list of general knowledge and skills will probably be a long one. Even though the list is long, it is unlikely that the test will provide an estimate of proficiency on such general skills. Instead, those skills specific to the occupation or profession will be tested and scored.

The knowledge and skills specific to the profession and critical to the protection of the public should be identified. In developing a job analysis for electricians, a domain might be the use of tools (observing an electrician at work would reveal a large array of tools in a hip pack). One such tool is probably a "Klein off-set screwdriver." Non-electricians would not be expected to know the use for this particular tool, but an electrician should (and "handy" home owners would be well advised to learn). Job analyses can be accomplished by observing professionals at work, or by surveying them using mail, telephone, and/or personal interviews or some combination of these methods. Any method may be acceptable and, again, will produce a long list of knowledge and skills from which the knowledge and skills needed at the entry level for the protection of the public needs to be identified.

Subsequently, the list of general and specific critical knowledge and skills must be examined and prioritized. The measurement of general knowledge and skills tends to be easier than the measurement of job-specific skills. Care must be exercised in the selection of tasks to be included in the test so that the actual job performance is represented in the test. Even though reading may be required for

successful job performance, a reading comprehension test may not be appropriate for licensure.

The level of specificity associated with the identification of critical skills and abilities for an occupation or profession varies greatly. In some licensure settings, it is virtually impossible to obtain a listing of all the "critical" knowledge, skills, and abilities. For example, it might be argued there are domains of knowledge, skills, and abilities needed by a lawyer or physician that are critical, but within these domains it is virtually impossible to identify the specific knowledge, skills, and abilities that are critical. Specifically, a specialist in problems with the feet may not be expected to have much knowledge about throat infections. A specialist is licensed as a physician and at some later time may choose to seek certification in his or her specialty. Because of situations like this, some licensure tests may be undifferentiated in terms of critical knowledge, skills, and abilities (i.e., individual items may be difficult to classify as measuring "critical" things, but the domain from which items are drawn may be considered critical). In such undifferentiated professions it is assumed there is a broad-based, but nonspecific set of critical knowledge, skills, and abilities to be measured on the licensure examination. Law, medicine, elementary school teaching, and real estate sales are but a few examples of such professions.

Job analysis, prioritizing elements, and developing a test blueprint are critical steps in developing content validity evidence. The *Principles* (SIOP, 1987) list several aspects of content validity evidence that should be provided. These principles, as modified to focus on licensure testing, are:

A. *The job content to be sampled should be defined.* The job domain need not be exhaustive, but the definition of the domain should include the most important parts of the job. General knowledge and skills can be thought of as one end of a continuum and job-specific skills as the other end. Between them, one would expect to find blends of general and job-specific skills that the candidates for licensure would be expected to have.

B. *Special circumstances should be considered in defining job content domains.* If there are specific skills that are part of the job description, these should be included in the content domain description. Similarly, if there are parts of the job that would be difficult to test, a substitute method of measurement may be needed. For example, the task may require a piece of equipment that is too heavy or too costly to provide to the candidate in the testing situation. In order to deal with the use of this equipment, the test would have to deal with subordinate skills, related to the operation of the machine. Alternatively a simulation may be substituted (as in using a flight simulator prior to taking an actual flight). When testing subordinate skills or simulations are not feasible, then some other means to determine that skills and knowledge exist may be used. One such substitute is the requirement that the candidate graduate from a program and that graduation can not occur without demonstration of the knowledge or skill in quesiton. What adaptations

are made may well depend on the licensure situation and the specific circumstances.

C. *Job content domains should be defined on the basis of accurate and thorough information about the job.* The definition of a job content domain can be derived through an analysis of tasks, activities and/or responsibilities of the job incumbents. Worker specifications may include knowledge, abilities, job skills or even personal characteristics judged to be prerequisites to effective behavior on the job. For example, if licensure in a particular occupation implies that the licensee will need to establish rapport with clients, as might be the case for polygraph operators, the licensure board may decide that evidence of prior experience in maintaining such relationships be part of the licensure test.

D. *Job content domains should be defined in terms of what an employee needs to do or know without training or experience on the job.* It is important, when developing the test blueprint, to separate those skills that the licensing board would *like* the candidate to have from those that are *necessary* prior to licensure (entry level skills critical for the protection of the public).

E. *A job content domain may be restricted to critical or frequent activities or to prerequisite knowledge, skills or abilities.* The definition of the domain should include the *major* aspects of the job, and not seldom performed activities (unless such seldom performed activities are deemed critical for the protection of the public). There may be things that a licensed person *should* be able to do, but if these are not really job requirements, they should not be tested. It would be nice to assume that all candidates for licensure in pharmacy have good interpersonal skills. However, that is not part of licensure, even though the absence of these skills may doom the person to failure as a pharmacist.

F. *Sampling of a job content domain should ensure that the measure includes the major elements of the defined domain.* The test will not be long enough to include all of the skills included in the content domain. The actual test items will be a sample from the domain of possible items. A careful balance must be maintained such that the items selected are an appropriate representation of the domain.

G. *A test developed on the basis of content sampling should have appropriate measurement properties.* Wherever possible, the entire licensing procedure should be pretested. The usual statistics related to items and the test should be developed and examined.

H. *Persons used in any aspect of the development or choice of procedures to be validated on the basis of content sampling should clearly be qualified.* As note above, a responsibility of the agency is to see that all judges, and others involved in the licensure procedure are well trained.

There is no statistical index which attests to content validity. Some indices lend support to such evidence. In Principle G, above, for example, pretesting is

recommended, along with the derivation of means, variances, measures of internal consistency, item statistics, etc. These are all parts of the collection of content validity evidence.

Additional content validity evidence may be collected by using expert judges to examine and rate items in terms of how the items relate to the content specifications or objectives. Hambleton (1980) describes several ways that such judgments may be obtained and he provides illustrations of forms that may be used for this purpose. He advocates asking expert judges to match items with objectives (when objectives are the basis for the test specifications), but this method could be modified easily to fit a program that uses more traditional test blueprints. He also advocates asking different judges to rate the extent to which an item reflects the objective or domain specification. This method can also be modified to fit the more traditional test blueprint format. Do not be misled by the title of Hambleton's work: "Test Score Validity And Standard-Setting Methods." These content validity rating methods relate to score validity and the illustrations of formats are found in appendices. (There is also a useful rating scale for making judgments about individual multiple-choice test items, thatmay enhance the validity and reliability of any multiple-choice test.)

Smith and Hambleton (1990) also discuss other issues related to content validity. Such issues include the extent that local conditions (within a particular state) need to be taken into account in examining content validity in professions in which a national examination is used for licensure. This issue is also discussed by Nelson (1994). Smith and Hambleton suggest additional types of evidence that might be useful in examining content validity. They also discuss some interesting methods of using criterion-related evidence in a licensure setting.

Criterion-Related Validity

It has been argued that the collection of criterion-related validity evidence is a critical part of identifying competent candidates and protecting the public from incompetent ones (Hecht, 1979). Such a task is easy to describe. One simply correlates scores from the test with some criterion measure. However, the definition of the criterion is not an easy task. In the current literature in the licensure field, there seems to be a relatively high consensus that boards should not be putting much effort in gathering evidence of criterion-related validity.

The primary issue, of course, is what constitutes a reasonable external criterion measure. The criterion measure, in this aspect of validity, typically occurs after the administration of the licensure examination. Given that the purposes of licensure testing, as noted above, are identifying candidates with requisite critical knowledge and skills and protecting the public from incompetent candidates, what would constitute a valid criterion? At the time of testing, one either has or does not have sufficient knowledge and skills needed to be at the entry level, thus the criterion is actually determined by the content of the test. It would be tautological to say that the criterion is the score on the test (it is not reasonable to make the test its own criterion).

Similarly, if the criterion is some measure of "errors" that put the public in danger (the most reasonable criterion measure for licensure), then an effective

licensure test (one that has few false positives—licenses few people who should not be licensed) would successfully screen out those who might endanger the public and the criterion measure would not exist. Virtually all those who are licensed would score "zero" on the criterion measure (they would not make errors). If the licensure test did a poor job of screening, then the board would know about it quickly enough to take appropriate action without having to undergo special statistical studies of the test. In most cases, licensure boards have ways to suspend licenses for individuals who are a threat to the public.

At present we will concur with most of our colleagues that licensure boards should not be concerned with criterion-related validity. But the suggestions made by Smith and Hambleton (1990) on this topic may be of interest to some boards that feel a need for more than content validity evidence.

Instructional and Curricular Validity

Instructional and curricular validity may, or may not, be part of the validity evaluation process in licensure examinations. If the agency requires or provides training that precedes the test, evidence should exist showing that the knowledge and skills being tested appear in the curriculum. Instructional validity would be important should a challenge be lodged that candidates had no opportunity to learn what is being tested (i.e., the test was not instructionally valid). In other words, job-specific skills which can be learned only *after* licensure should not appear on the test.

The evidence from content validity evaluations should provide satisfactory evidence, within the construct validity concept, that the primary claims for licensure examinations have been met. Collecting the evidence is sometimes difficult and time-consuming, but will lead to better practices. Again, an agency may be well advised to seek professional assistance in either the design or the conduct of the evaluation study, or both.

SUMMARY

In this chapter, we have attempted to identify some of the basic, psychometric issues associated with licensure testing. In particular, we have looked at *reliability* as a general concept and the requirements pertaining to reliability that appear in various professional guidelines. Specifically, we discussed measures of *internal consistency, stability and equivalence, or equivalent forms.* We recommend the development of equivalent forms, wherever possible. Also discussed was *decision-reliability* and the methods that can be used to estimate it.

Validity was the other psychometric issued treated in this chapter. For many, if not most, licensure examinations, *content validity* is of primary concern. The *construct issues* deal with whether candidates possess sufficient knowledge and skills to qualify for licensure and whether passing the test will protect the public from incompetent candidates. We close the chapter by recommending that agencies spend the time and effort necessary to collect evidence with respect to these vital construct issues.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

American Council on Education. (1936). *The construction and use of achievement examinations.* Washington, DC: Author.

Brinegar, P. (Ed.) (1990). *Occupational and professional regulation in the states: A comprehensive compilation.* Lexington, KY: National Clearinghouse on Licensure, Enforcement and Regulation.

Cohen, J. A. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Pyschometrika, 16,* 297-334.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16,* 407-424. (Reprinted in Mehrens, W. A. & and Ebel, R. L. [1967]. *Principles of Educational and Psychological Measurement A book of Selected Readings,* pp. 116-131. Chicago: Rand McNally.)

Equal Employment Opportunity Commission and others. (1978). Adoption by four agencies of uniform guidelines on employment selection procedures. *Federal Register, 43,* 38290-38315.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement,* (3rd ed.; pp. 105-146). Washington, DC: American Council on Education.

Geisinger, K. F. (1992). The metamorphosis of test validation. *Educational Psychologist, 27,* 197-222.

Hambleton, R. K. (1980). Test score validity and standard setting methods. In R. A. Berk, (Ed.), *Criterion-referenced measurement: The state of the art* (pp. 80-123). Baltimore, MD: The Johns Hopkins Press.

Hecht, K. A. (1979). Current status and methodological problems of validating professional licensing and certification exams. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based education* (pp. 16-27). Washington, DC: National Council on Measurement in Education.

Huynh, H. (1976). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13*(4),253-264.

Linn, R. L. (1979). Issues of reliability in measurement for competency-based programs. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based education* (pp. 90-107). Washington, DC: National Council on Measurement in Education.

Livingston, S. A. (1972). Criterion-referenced applications of classical test theory. *Journal of Educational Measurement, 9*(2), 13-25.

McClung, M. S. (1978). Are competency testing programs fair? Legal? *Phi Delta Kappan, 59,* 397-400.

Messick, A. (1989) Validity. In R. L. Linn (Ed.), *Educational measurement,* (3rd ed.; pp. 13-103). Washington, DC: American Council on Education.

Nelson, D. S. (1994). Job analysis for licensure and certification exams: Science or politics. *Educational Measurement: Issues and Practice*, *13*(3), 29-35.

Raju, N. S. (1982). The reliability of a criterion-referenced composite with the parts of the composite having different cutting scores. *Educational and Psychological Measurement, 42*, 113-129.

Shepard, L. (1994). Evaluating test validity. In L. Darling-Hammond (Ed.) *Review of research in education 19* (pp. 405-450). Washington, DC: American Educational Research Association.

Smith, I. L., & Hambleton, R. K. (1990). Content validity studies of licensing examinations. *Educational Measurement: Issues and Practice*, *9*(4), 7-10.

Subkoviak, M. (1976). Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement, 13*(4), 264-276.

Subkoviak, M. (1980). Decision consistency approaches. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art* (pp. 129-185). Baltimore, MD: The Johns Hopkins Press.

Society for Industrial and Organizational Psychology, Inc. (1987). *Principles for the Validation and Use of Personnel Selection Procedures (3rd Ed.)*. College Park, MD: Author.