1995

6. Developing And Using Clinical Examinations

Jimmie C. Fortune
Virginia Tech

Theodore R. Cromack
Consultant

DEVELOPING AND USING CLINICAL EXAMINATIONS

Jimmie C. Fortune

Virginia Tech

Theodore R. Cromack

Consultant

INTRODUCTION

Generally, clinical examinations for licensing (sometimes called performance tests) involve the candidate completing one or more tasks (in licensing this is thought of as "services for a client") that have been selected from the supervised practice (job analysis) of an occupation or profession. The clinical examination may exist in contexts (occupations) that do not require client interactions. Such contexts include building trades, automobile repair, accounting, etc. The tasks may range from fixing brakes, to preparing a body for burial, to wiring a house, or auditing a set of business interactions.

Other contexts require the candidate to perform services or tasks while interacting with a client. Such tasks include filling a tooth, counseling, fitting contact lenses, hair removal, and similar services. These tasks would then be graded as part of the licensure examination. Supervision or scoring of the tasks in the context where interaction is not required is easier than those requiring the presence of the client. Interaction with clients makes the second type of tasks harder to supervise and to grade. In recent practice, some boards have moved from using live clients to using simulations (Yaple, Metzler, & Wallace, 1992). Oral interviews may be required prior to issuance of a license in some contexts, but such

entry orals or group interviews are not considered here as clinical examinations as they seldom are tasks germane to the job analysis of these occupations or professions.

Tests are used as a proxy to judge the ability of an individual relative to actual performance of a task. It might be useful to consider a continuum of faithfulness to the task ranging from a paper-and-pencil (multiple-choice) test to actual performance of the task. This continuum describes the concept of fidelity, or the degree to which the test requires the same behaviors as those required by the task. Unfortunately, this faithfulness to the task is only half of the equation, the other half is accuracy of the inference made about the candidate's ability to complete the task. This dimension speaks to the measurement concept of validity. Both fidelity and validity are complex, thus making a judgment about ability a complex activity.

Human judgment is complex and so intertwined with previous experience that total objectivity is virtually impossible to achieve. So called "objective tests," such as multiple-choice tests, generally moderate judgment by being constructed using multiple judges to determine content and to set cut scores, and by being scored in such a manner that individuals who perform the same task in the same way will attain the same score (often scoring is possible by machine or template). Moving along the continuum toward actual performance (i.e., from multiple-choice tests through essay tests, oral tests, and simulations, to actual clinical performance), the potential gain in fidelity can be offset by loss in objectivity.

Clinical tests appear on the side of the continuum closest to the actual performance of the task. The discussion which follows offers suggestions for enhancing the objectivity in the development and use of the clinical tests. Making scoring judgments explicit, reducing compounding of judgments, utilizing multiple judges in scoring individual performance, and providing statistical evidence of reliability, validity, and fidelity are among the topics discussed.

Performance on a clinical examination generally requires the candidate to use a combination of knowledge gained in training, skills acquired in the education or training program, physical attributes demanded in practice, interpersonal interaction skills, and attitudes. The clinical examination is believed to require candidates to demonstrate their ability to master and apply these different elements in concert. In a discussion with the Advisory Board to Southern Regional Testing Agency (SRTA) at Fort Walton Beach, Virginia, August, 1991, one dentistry board member explained that the clinical examination requires diagnosis, treatment, and patient education cast in the context of dealing with a fearful and uncomfortable patient. It requires practice of the profession along with human management. In dentistry, it was suggested that clinical examination required the candidate to work in the hard-to-reach areas of the mouth without compounding the patient's problems by injury of the areas that make access difficult. Common dental clinical examinations include: one or more types of restoration, prosthetics, and endrodentics.

Historically, these examinations were graded without psychometric analysis. In clinical examinations such as these, the complexity resulting from the joint application of several measurable actions, each of which could alone or in combination cause the service to be unsatisfactory, led the scorers to make single,

global judgments of the degree of satisfaction of the candidate's performance of the service. When civil rights became an issue in the early 1960s, efforts to guarantee fair treatment of the candidates brought about increasingly sophisticated psychometric treatment of clinical examinations (Weiss, 1987). Among efforts to improve practice are methods to: (a) make scoring judgments explicit; (b) display the criteria associated with these judgments; (c) control the objectivity and uniformity of these judgments; and (d) measure agreement in these judgments (Schroeder, 1993). In other words, a methodology of clinical testing is now under development.

"Getting items" for clinical examinations differs from standard (i.e., multiple-choice) test development procedures in both conceptual and practical ways even though both start from the same body of information: the job analysis. In building multiple-choice tests, standard practice is to build from an inductive perspective using a "table of specifications" (or test blueprint) framework and the notion of domain sampling from the critical dimensions of the job. In clinical examinations, scorable dimensions, or items, must be extracted from a task within the clinical process. Hence, the test is developed in a more deductive manner. First comes the task, then critical elements of the task are identified and defined as criteria to be scored, hence items.

Many times clinical tasks result in an end product, a dental plate or a properly fitted pair of eye glasses. In such cases, evaluation of the product may be the most appropriate assessment of adequacy of performance by a candidate. Conversely, some clinical tasks are more aptly referred to as process tasks and do not yield a readily assessable product. Among such tasks might be ability to reduce a dental patient's anxiety concerning use of the drill. Such process tasks require observation and evaluation of the process as opposed to assessment of a product.

DEVELOPMENT OF SCORING PROCEDURES

When scoring a clinical examination, there is a tendency to avoid a systematic set of procedures and to make an overall global judgment of "pass/fail" for the whole task. The global score is unsatisfactory because it fails to: distinguish between degrees of successful completion of the task, provide the candidate with adequate feedback, make explicit the judgment process, and permit an opportunity to look at the degree of agreement among the judges. The scoring process becomes mystical without a systematic means of arriving at an estimate of the accuracy and quality of completion of the task. Of course, if the judgment is based on a single step or performance, then a single judgment is appropriate.

Beyond the overall judgment or single "pass/fail," scoring procedures are usually designed to describe the adequacy with which the task is completed or the quality of performance of each step in the process of performing the task. Three strategies appear in use, the first strategy is a 100-point system, which we do not recommend. This system involves subtracting a fixed value for each error, usually 1 or 2 points from the customary 100 points assigned to the candidate at the beginning of the examination. Our objections are based on the arbitrary handling of points and on the lack of identification of the number of potential errors. The second strategy is the dichotomous scoring of each process step or criterion point

and the summation of the scores for correct steps. The third strategy is similar with regard to scoring the process steps, but weights are assigned to each step in accordance to some rule such as importance or criticality.

Making Scoring Judgments Explicit

Two strategies have been used to make grading judgments explicit. The first strategy breaks the performance of the clinical examination into explicit steps to be performed or skills to be displayed. This method appears most appropriate for process-based clinical examinations, especially those requiring the candidate to interact with a client. The second strategy is to define scoring criteria for the rating of accuracy or quality of the result. This method is most appropriate for product-producing clinical examinations. The first strategy involves a process similar to that of job analysis and the second strategy involves a process of deductive valuing, such as is done in consumer rating of different makes of automobiles.

The National Board performance test in optometry is an excellent example of the explicit step strategy (Gross, 1993). In this examination, 18 clinical skills are performed and evaluated using real patients. The performance of the skills occurs at five examination stations and each skill is scored independently by two judges. Two to six skills are performed at each station. At four of the stations the candidate is faced with a different patient and set of tasks to perform. At the fifth station an examiner portrays a patient from whom the candidate takes a case history.

For scoring, each skill is subdivided into its component items and each item is scored as pass or fail. Although one may observe that this process still requires subjective judgment, this judgment is made on a much narrower, well-specified area of performance. This specific performance can be addressed in the rater calibration process and is more directly linked to the final result. The performance test requires approximately 3 hours, each skill has 9 to 42 items and across the 18 skills there are 279 items scored independently by each judge. Criticality weights have been determined for each item. These weights reflect the consensus judgment of the nine-person examination committee of the relative importance of each item. It is hoped this criticality is related to the job analysis.

To obtain a candidate's score, each judge, for each item, multiplies the dichotomous item score by its criticality weight to get the weighted score for that item. Weighted item scores are then summed for each judge to get a skill subscore for each of the 18 skills. From the sum of these subscores, a pass-fail decision is made for the skill subtest. The process used to determine pass-fail for the skill subtest is based on the amount of error that can be "tolerated" for the skill. Gross (1993) reports that the tolerance level for a skill subtest is one point less than the highest weighted (most critical) item in the set of items associated with the skill. "Therefore, the pass-fail index for a clinical skill is designed to identify all candidates who perform all items correctly except for the most critical item" (Gross, 1993, p. 20). Gross points out that the pass-fail score for the whole examination is the sum of the pass-fail scores for the subtests. In cases where the skill subtests are not "Go No-Go" decisions, the candidate can make up poor subtest scores with high scores on other skill subtests. The dichotomous scoring of the

items that go into a skill score is designed to promote interrater agreement. Mock examinations are used to estimate interrater agreement. Gross reported no attempt at estimating intrarater agreement.

The six clinical examinations of the licensing tests for dentists administered by the Southern Regional Testing Agency, Inc. (SRTA) provide excellent examples of the product-producing clinical examination and of the method of making judgments explicit by identifying criteria to assess success. The SRTA clinical examinations include: a non-metallic restoration, a metallic restoration, a gold restoration, two prosthetics (casting and fitting), and endrodentics (Minnich, 1992).

Scoring for each examination was devised by working backwards and deciding what would prevent a work sample from being acceptable. A panel of seven experts met and agreed on criteria to score each clinical examination. The criteria were very specific with descriptions to help pinpoint critical degrees of correctness. For instance, the criteria for the nonmetallic restoration were categorized for scoring as to cavity preparation and then as to finishing. Included in the cavity preparation were five decisions:

1. Was the cavity cleaned of decay?
2. Were the cavity walls prepared so as to facilitate the restoration staying in the cavity?
3. Was the cavity prepared with an anatomy that would permit a solid restoration?
4. Was the depth of the cavity handled appropriately (cement or treatment used if depth is too severe)?
5. Was the preparation properly cleaned and connecting teeth and tissue protected?

Prior to each test administration, slides are used to calibrate the judges on these criteria. Similar criteria are specified for the finishing of the restoration (Minnich, 1992).

Establishing Rater Agreement and Estimating Reliability

Two concepts of agreement are important in a scoring system. One concept is agreement across raters for an examinee or set of examinees for each item, referred to as interrater agreement. This concept could be thought of as one examinee and multiple raters: stability over raters. A second concept is that of internal reliability, or the agreement within a clinical examination for one rater. This is often called intrarater agreement and refers to stability of judgments. To maximize the extent of inter- and intrarater agreement, judges receive training, sometimes referred to as calibration. Calibration represents the degree to which several judges identify the same level of correctness for a given clinical performance or the degree to which a single judge identifies the same level of correctness for the same clinical performance over several examinees.

In 1951 Ebel proposed an analysis of variance format to estimate the reliability of ratings. Medley and Mitzel (1964) expanded the process to study multiple types of agreement using a single analysis of variance (ANOVA) framework. This model was then translated into the Winer (1971) repeated measures concept of reliability

analysis. The extent that the judges are equivalent and the extent that intrarater scoring processes are uniform across test administrations determine the overall comparability of scores across test administrations and sites. Feldt and Brennan (1989) demonstrate the thoroughness with which generalizability theory addresses the multiple agreement and reliability needs of clinical examinations.

Feldt and Brennan (1989, p. 115) present a reference page of methods to address internal consistency using different types of data and different theoretical models. A coefficient and formula can be found to fit most clinical examination cases. In addition to the ANOVA to establish interrater agreement of two or more judges across several performances of the same clinical examination, Pearson product-moment correlation coefficients may be used for interval data or phi correlation coefficients for dichotomous data.

Reliability of the scoring process should be studied during pretesting of the clinical examinations and during the actual test administrations. Kenyon and Stansfield (1991) recommend and demonstrate the utility of pretesting in refining tasks for performance assessments. Their work is generalizable to clinical examinations in licensing. The double-blind process discussed in protection of candidates, if used, permits the estimation of interrater agreement on the scores used for licensure. Butzin, Finberg, Brownlee, and Guerin (1982) provide a model for the study of reliability of grades from oral examinations that can be used for other forms of clinical examinations. In many cases slides or simulations are used to establish agreement needed for calibration (Minnich, 1992). Friedman and Ho (1990) report a study of interjudge consensus and intrajudge consistency in standard setting. Their paper indicates the tradeoff between the two concepts.

Methods of Combining Scores from Standard Tests and Clinical Tests to Determine Eligibility for Licensure

Although there are formulae in measurement theory that allow one to combine the scores from several tests into a combined score taking into account the mean and variances of each test (Hopkins & Antes, 1990), we find that these formulae are seldom used in licensure testing. Instead, three methods of rendering the "pass-fail" decision appear to be the most commonly used in licensure settings. All three assume each test or examination represents an independent score on an essential criterion for practice. In the first method it is not assumed that the candidate must pass all of the examination parts, but the candidate must do well enough on all parts to accumulate enough total points to exceed the preset cut score (based on the combined results). In this first method, points are given to each test, written and clinical, and the points scored on the combined tests are summed to a total score which is compared to a preset cut score. This method permits the candidate to do poorly on one test or examination and to make it up by doing well on the rest. This method is frequently called an unweighted compensatory method.

In the other two decision processes, the candidate takes each examination, often one or more written and one or more clinical examinations, as separate, independent events. Both decision processes require the candidate to pass each of the examinations before eligibility for licensure is established (i.e., a conjunctive

model). In the second of these three processes, the candidate must pass all examinations in a single testing, one failure results in the requirement to retake the entire examination. Costs of the examinations have tended to reduce the use of the pass all-in-one-sitting requirement.

In the third method, partial credit is permitted. The candidate must pass all examinations, but credit is given for the passing of one or more parts or examinations and the candidate can return in a future examination period to retake the examinations or subtests not passed earlier. This permits the candidate to accumulate passed examinations and is called the part-credit model. Millman (1989) argues for the setting of higher cut scores if the latter method is used. He feels that the probability of passing is modified and a higher cut score is needed to maintain discrimination or to identify the absence of competence.

ISSUES WITH CLINICAL EXAMINATIONS

Several issues are frequently raised by licensure board members or persons interested in licensure testing. Among these issues are: Why should a clinical examination be given? How close to the task must the clinical measure be? How can testing conditions be made uniform and fair? Does the clinical portion have to be standardized? What procedures are needed to insure standardization of the clinical portion? How do these procedures relate to the scoring procedures? What test statistics are needed for clinical items? Can test statistics be computed in the same way as for paper-and-pencil tests or other kinds of performance tests? What special procedures are needed to set a cut score for the clinical portion and how do these relate to continued testing using part credit? And is there some indication to show that a clinical measure is obsolete?

Absolute and comprehensive answers to many of these questions do not exist. In the following pages we discuss considerations required to develop answers for these questions.

Why Should a Clinical Examination be Given?

Interviews with board members in dentistry, nursing, and several licensed commercial occupations suggest that the clinical examination came about from three conditions: a mistrust of the paper-and-pencil or multiple-choice test, a need to see the candidate work with people, and a need to see the candidate perform in a work setting integrating the physical and cognitive skill areas. Often the services selected for the clinical examination are services commonly performed in practice, in many cases they are among the most frequently performed services, but certainly they are services or tasks perceived as "critical" in the job analysis.

Schroeder (1993) presents a series of questions concerning the use of oral, practical (which we have elected to call clinical examinations), or essay examinations, which may be helpful in making the decision whether or not to use a clinical examination. These questions are:

1. Is the behavior being measured something that could not be evaluated by the use of a multiple choice or objectively scored examination?
2. Are the evaluators thoroughly trained prior to the examination and administration?

3. Are the evaluators free of conflicts of interests concerning the candidates?
4. Are there detailed criteria for evaluating and scoring?
5. Does each evaluator make an independent rating?
6. Are at least two independent evaluations made for each candidate?
7. Is the evaluation free of potentially biasing information about the candidate which is not related to examination performance?
8. Has the examination session been documented (proctored, audio or video taped)? (Schroeder, 1993, p. 19)

The publication then suggests what the answer should be to elect to use a clinical examination.

How Close to the Task Must the Clinical Measure Be?

The actual requirement of clinical examinations (we prefer to address each required task as a single examination) may differ by profession or occupational area. Yet, all clinical examinations should evolve from a job analysis directed toward the identification of potential practices that may threaten to harm the health or safety of the public. Generally, clinical examinations are chosen from job analyses because (a) they define a frequently performed and important activity in the occupation (i.e., primary job activities), or (b) they require a complex coordination of cognitive and physical skills for successful practice, or (c) the professional practice demands complex interpersonal interactions with the "clients", or a combination of these reasons.

The first two reasons, primary job activities and complex multiability tasks, may, though do not specifically have to, result in a product-producing examination. Such a product, resulting from the examination, can be subjected to review or even tried out to determine its adequacy. This was illustrated above in the discussionof the SRTA dentistry examinations (Minnich, 1992). The third reason requiring "client" interactions is likely to lead to a process-performing examination. As was illustrated above in the discussion of the optometry examination (Gross, 1993), "Interaction with a client" is a process, as opposed to "preparation of a partial dental bridge" (p. 20) which is a product. Processes are more subjectively evaluated making it more difficult to establish uniform conditions across candidates and to grade the adequacy of the process.

The reason for using a clinical examination should be embedded in the examination. If the clinical examination is selected because of its importance in defining primary activities, it should contain all of the basic elements of performance required in practice (i.e., diagnosis, treatment, client education, etc.). In optometry such an activity might be an eye examination; the examination in this case may be identical with the task.

However, if the clinical examination is selected because it requires a complex coordination of cognitive and physical skills for success, opportunity to perform in a real or near real situation is necessary. In dentistry, such an activity may be the restoration of a molar using nonmetallic filling. This task must be performed in the patient's mouth so as to see if the candidate can handle the physical

challenge of working in an awkward position, the challenge of bleeding, patient reaction, etc.

Clinical examinations selected because of required client interaction skills should deal with actual clients who hold real attitudes and perhaps limited tolerance for pain. In optometry, the fitting of contact lenses and the education of the client may be tasks where the candidate and client patience are taxed and the client's threshold of pain exceeded.

Clinical examinations may be selected because of two or more of the characteristics mentioned: (a) frequently performed or important activities, (b) complex coordination of cognitive and physical skills, and (c) complex interpersonal interactions. All three of these reasons appear operative in the case of the dental clinical examination in endrodentics. It is a common practice in dentistry to have to relieve pressure in the root of a tooth. The process of drilling to relieve pressure requires the coordination of cognitive and physical skills and the "client" needing the service or task performed is certainly in pain.

How Can Testing Conditions be Made Uniform and Fair?

For examinations that do not require the use of patients or "real" tasks, fairness and uniformity concerns focus on the candidate. Addressing these concerns for the candidate requires four steps: (a) Assure that the candidate knows what is to be done; (b) be certain that the candidate receives the correct reaction when the appropriate response is made; (c) make certain that the task required is relevant to the job analysis and is not just an exercise; and (d) equate the differences in tasks with regard to difficulty by avoiding the selection of either overly simple or highly complex tasks. Failing to follow these four steps precludes a fair examination as illustrated by the following case involving licensing of polygraph operators. A candidate for licensing as a polygraph operator who was being observed was subjected to an oral examination for which no script was written and in which the examiners "just winged it." The absence of a script and the spontaneous and potentially arbitrary behavior of the examiners made it impossible for the candidate to know what was to be done or what behavior was expected. Because the questions were ad-libbed and not shared with the candidate or even with other examiners prior to the oral interview, it is unlikely feedback on the candidate's responses was appropriately given. It is difficult to see the relevance of an impromptu set of questions to the administering or scoring of a polygraph. Hence, the oral interview was likely just an exercise and was not based on the job analysis. This became more evident when transcripts of other oral examinations were reviewed. These transcripts revealed lack of uniformity in questions asked and of relevance to the operation of a polygraph (Maust, Callahan, Fortune, & Cromack, 1988).

An important consideration in mounting a clinical examination requiring the participation of actual patients or live clients is making certain that the clinical examination is fair to the candidates and is safe for the participating patients. The fairness issue arises from the fact that amount or complexity of services required by patients varies and may offer more or less challenging cases to the candidates.

Certainly, this variance in severity of the clients' problems does not constitute all of the criteria involved in assessing of the fairness of a test, but it is a major consideration in the use of live clients. Concern for the patients is based on the very threats that give rise to the need to regulate. The Council on Licensure, Enforcement and Regulation (CLEAR) has recently published a monograph entitled, *Principles of Fairness: An Examining Guide for Credentialing Boards* (Gross & Showers, 1993), to assist board members in the examination process.

If the examination is to include live clients, explicit instructions must be provided for choosing a cooperating patient. These instructions will describe the task to be performed by each candidate so that patients will be selected having similar needs to be addressed by each candidate. Tasks performed by each candidate should not only be similar, but should be of similar difficulty. Given that these instructions can create uniform levels of difficulty of tasks to be performed, the next step is to assure that no bias occurs in candidate grading. This is usually taken care of through the use of a double-blind procedure for grading. In the double-blind procedure the clients are disassociated from the candidates and are seen by the judges who score the candidate's work independently. The candidate is not seen by the judges. The client does not know the judges' ratings and the candidate does not know who scored his/her work. There are several ways in which the blinds can be constructed, either by moving clients or by moving judges. Logistics can present a problem, but usually the assignment of a candidate number to a patient or moving the patient to the judge can allow the double-blind procedure to work (Gross, 1993; Minnich, 1992). Most methods to assure fairness either use blinds or multiple judges to average out biases. Regardless, the principles are approximately the same.

The double-blind grading procedure works as a protection for the candidate against several types of discrimination, such as race, gender, age, etc. Yet, this protection is somewhat costly in that opportunities to assess candidates' interpersonal skills and attitudes toward patients are lost. The skills and attitudes appear critical in all but a few incidents where clinical examinations are used.

Does the Clinical Portion Have to be Standardized?

Schroeder (1993) suggests that all clinical examinations be standardized in order to insure that each candidate took approximately the same examination. This standardization also aids in helping the judges look at approximately the same criteria to score the performances. "While there are many differences, oral practical and essay examinations also have much in common with objectively scored examinations. Both forms of examination should be standardized so that all candidates have the same opportunity to demonstrate competence" (Schroeder, 1993, p. 18). Standardization may occur in many ways, among them is the use of standardized patients, or patient simulation, where well-rehearsed "actors" are used to insure that each candidate is provided the same opportunity to perform such tasks as collecting a history. This methodology is reviewed by Vu & Barrows (1994). Standardization is desired in some areas to increase mobility through extended reciprocity (Allen, 1992).

What Procedures are Needed to Insure Standardization of the Clinical Portion?

Standardization involves creating the conditions that assure uniformity of the tests with regard to administration, difficulty, clarity in scoring, and establishing psychometric evidence of the quality of the test. One of the conditions demanded is making the scoring criteria explicit. Explicitness means that the number of judgments are listed and clear scoring instructions are written, thus permitting the judges to be calibrated. By calibrated, we mean that each judge's score has the same meaning as every other judge's score. A second condition demands that more than one task or client be required within a given clinical examination to preclude post-test discussions from giving future candidates an unfair advantage in the examination. Lastly, all of the tasks need to be prestudied in order to assure near equality with regard to difficulty (a fairness concern) and fidelity to the job analysis. Two ways of making criteria explicit are discussed earlier in this chapter.

How Do These Standardization Procedures Relate to the Scoring Procedures?

Most clinical examinations can be scored in a variety of ways. Scoring procedures can include several options for the assignment of numerical values to a performance. Such options range from the global judgment of adequacy to intricate tallying of correctness for every step in a process.

The most important factor to include in scoring procedures to insure standardization is difficulty of tasks (or steps). In many clinical examinations some candidate errors are more important than others. In fact, an error such as severe damage to a tooth adjacent to the one on which a dental procedure is being performed can be deemed by the examiners to be so critical that the candidate is failed immediately. Errors that require immediate failure are referred to as "go no-go" items. Other errors appear as very important, but not so important as to demand immediate failure. In the case of differences in step or task importance, weights may be assigned to assure that passing or failing an important step is more significantly reflected in the score than passing or failing a minor step.

Of the two most common methods for scoring, "points correct" and "points off," the second poses the most potential problems. Scoring by summing values representing the adequacy of performance for each criterion is the "points correct" system. Deducting values assigned to each error from a constant score is the "points off" system. When using the "points off" system, errors may be chained, that is, some errors cause other errors to occur later in the scoring process. There must be a provision to handle these chaining errors. Although chaining of errors may also occur in the "points correct" system, this procedure is more adaptable to assuring independent item scoring. Chaining of errors can occur only when items are not independent.

Standardization mandates careful and uniform administration and scoring of the examination. Hence, administrator instructions must be carefully reviewed, making a well-edited examination guide and explicit scoring criteria a necessity. Making the scoring criteria explicit aids both the candidates and the judges. The

candidates are aided through the articulation of examination expectations. The judges are enabled to render more uniform judgments, due in part to calibration or judges' training and in part to simplifying the judgment.

What Test Statistics are Needed for Clinical Items?

Schroeder (1993) treats clinical examinations similarly to objective examinations with regard to psychometric evidence of quality. "Both types of examinations, clinical and objective must have a minimum passing standard, and the validity and reliability of the examination program is crucial for both types of examination" (Schroeder, 1993, p. 18). In objective examinations several statistics attesting to reliability appear interchangeable (e.g., Coefficient alpha and Hoyt's method). We suspect that the same situation is emerging for clinical examinations.

The statistics needed to support the utilization of scores from a clinical examination are those that substantiate the fulfillment of the requirements for standardization such as uniformity of the examination content over candidates. Uniformity is necessary for reliability and for making valid interpretation of the examination results because unless the candidates all receive essentially the same test (i.e., they are tested uniformly), one cannot claim that they meet the minimum qualifications to be licensed.

Statistics are needed to show that the clinical tasks have similar performance profiles across successful candidates, there is interrater agreement among the judges, the examination scores are reliable and yield valid interpretations, there is intrarater agreement, there are no systematic exceptions especially in terms of difficult areas in the examinations, and there are similarities between examinee classes or groups to which the examinations are administered to substantiate interpretation of the statistics. By systematic exceptions is meant that candidates performing a task such as drilling a tooth are all given approximately equal or uniform tasks, there is no evidence of systematically assigned difficult tasks.

Clinical examinations are graded or scored using a fixed set of criteria, which indicate the successful completion of the steps required to complete the task and that are designated a priori to examination administration. Multiple administrations of a given clinical examination should produce similar percentages of correct responses across steps. Similarly, comparable percentages should pass the clinical examinations across testing sites and across different test administrations, or a careful review should be made to assure that administrations were fully standardized. Similarity of percentages can attest to the uniformity as equivalents to the difficulty statistics used with objective tests. Points or steps where the candidates have the most and least difficulty are of interest to the examination analyst as indicators of potentially too much difficulty or too little discrimination (Maust et al., 1988).

Statistics are needed to attest to reliability of the examination results. Several types of reliability are of interest and if the judgments are reduced to dichotomies, there are several options in the choice of reliability methods and statistics. Reliability was discussed earlier as it relates to the design of the scoring procedures as well as in the chapter by Stoker and Impara in this book (Chapter 7).

Statistics are also needed to attest to the making of a valid interpretation of the examination results. Statistics that are helpful here are those: which demonstrate the relationship of the examinations to the job analysis, which show that the clinical tasks have similar performance profiles across successful candidates, and which investigate the similarity between examinee classes or groups to which the examinations are administered.

Statistics are needed to monitor examination performance. Records should be kept on exceptions to prescribed process, frequent examination difficulties, and examination performance across time. Such records of examination performance are useful in identifying trends, signalling out-of-date material that should be replaced, indicating potential bias in tasks or scoring, and other indicators of need for examination review and maintenance.

Can Test Statistics be Computed in the Same Way as for Paper-and-Pencil, Multiple-Choice Tests, or Other Kinds of Performance Tests?

Most test statistics used with clinical tests involve dichotomous analogs to statistics used with objective tests or statistics that can be completed using differential item weights. Interrater and intrarater agreement become statistics needed to assure the scoring process and the work of the judges. Coefficients of agreement such as reliability can be calculated several ways. These were discussed above under "Establishing Rater Agreement and Estimating Reliability."

Clinical examination requirements focus on uniformity of the examination procedures and tasks designed to be equivalent. The most useful statistics in looking at uniformity appear to be frequencies of examination exceptions in administration and the effects of these on examination averages. Difficulty levels of the items making up the examination and of the total examination should be analyzed across tasks within a clinical examination, across examination administrations, and across examination administration exceptions.

Estimating item difficulty levels can be done on the judgments in much the same way as it is done in objective testing (Crehan, 1974), specifically, by calculating the proportion of all examinees who answer the item correctly (or are given positive credit for their performance). The same is true for discrimination indices (Millman & Greene, 1989) (e.g., by calculating the correlation between the item score and the total score). Test analysis can be conducted with simple statistics as described by Schroeder (1993): "mean score ... Changes in the mean score from administration to administration may signal either changes in candidate capability or examination difficulty. Often, large changes in overall score means are associated with scoring errors so it is important that score means are reviewed, and the reasons for the change in the score mean investigated" (p. 30); "*standard deviation* ... When the number of candidates is large, the standard deviation will usually be very stable from administration to administration. Large changes in the standard deviation may signal changes in the nature of the candidate group or errors in scoring" (p. 30); "*standard error of measurement*... A relatively small standard error of measurement means that one can be confident that the test scores have a high degree of accuracy. If the standard error of measurement is high, the

associated test scores may have a lower degree of accuracy" (p. 31); and "*score frequency distribution* ... By comparing frequency distributions from two or more administrations, changes in the nature of the candidate group can be identified. Large changes in frequency distributions may be indicative of scoring errors or of changes in the nature of the candidate group" (p. 31). For large licensure testing programs, application of item-response theory may also be applicable.

What Special Procedures are Needed to Set a Cut Score for the Clinical Portion and How Do These Relate to Continued Testing Using Part Credit?

(See Chapter 10 by Mills for a complete discussion of setting cut scores.) Although other methods exist, it has been our experience that the most frequently used methods for setting cut scores for licensure examinations are Angoff, modified Angoff, and Ebel methods with the Angoff method being used much more frequently than the other two methods. All three methods are test-centered continuum models using judges and rating of items (Jaeger, 1989). Angoff's method leads the judges to set a score that is expected of a minimally qualified population of candidates. The methods use panels to identify item weights for each item. The Angoff method develops weights on the probability of minimally qualified candidates getting the item correct. The modified Angoff methods get the judges to assign item weights and the Ebel method develops item weights using relevance and difficulty classifications.

In working with performance tests the criterion points (or steps) can be treated as items. Because the task was chosen from the job analysis, and because the steps were determined as essential to the completion of the task, dichotomous scoring greatly simplifies the work of the panel of judges as it transforms the judgments to an analog of a right/wrong item. Complications occur when the steps can be partially correct and still the effort results in a successfully completed task.

When clinical examinations are more complex, such as those involving assessment of a candidates portfolio, standard setting is also much more complex. Several articles appear in special issue of *Applied Measurement in Education*, Volume 8(1) that examine issues related to setting standards in such a situation.

Is There Some Indication to Show That a Clinical Measure is Obsolete?

Usually, the clinical measure becomes obsolete when the task is no longer practiced due to a change in the profession. The harbingers of this need for replacement are usually research reports and workshops designed to have incumbents learn new practices in the occupation. Hence, members of the board who are practitioners would know of the changes and anticipate when a new job analysis should be made to see if the clinical examination should be revised. For instance, in optometry the diagnostic examination would continue, but changes in prescribing glasses for the near sighted may end if the emerging surgical procedures to reshape the cornea become widespread, making eye glass correction for myopia virtually obsolete. (Note: The operation is generally successful and the new laser procedure has proven very successful in Canada.) In dentistry, the molding of gold

restorations is no longer an important practice because almost all gold restorations are molded in the laboratory. The latter case was verified through a four-state survey of dental practices conducted for the Virginia Board of Dentistry (Fortune, 1991). With regard to psychotherapy, performance or clinical testing is currently under challenge in several states and in Canada (Trebilcock & Shaul, 1983). Clinical examinations in this area suffer from the lack of clients who can participate in testing without adverse effect. In part, clinical examinations are not used in psychotherapy due to the difficulty in making the tasks standard and in the lack of belief in the oral examination process.

IN SUMMARY

We have provided an overview of the rationale and procedures associated with developing, scoring, and using clinical examinations. Moreover, we have tried to provide answers to the following questions that have been raised by licensure board members:

- Why should a clinical examination be given? Is there some indication to show that a clinical measure is obsolete?

If a clinical examination has been indicated through the job analysis, documentation of its disappearance from practice must be made before it should be removed from use. Board members are often the first to question the continued use of a specific clinical examination.

- How close to the task must the clinical measure be?

A clinical measure should be as nearly identical as possible to the condition that gave rise to its existence. If the examination is given because of human interactions, those human interactions must appear in the clinical examination. If the clinical examination has been developed because of the required joint application of complex psychomotor and cognitive skills, then the candidate should have to exhibit those complex skills. Jointly the choosing of the task to fit the dictates of the job analysis answers a validity question and choosing the tasks to be performed very close to tasks in practice addresses the fidelity issues.

- How can testing conditions be made uniform and fair? Does the clinical portion have to be standardized? What procedures are needed to insure standardization of the clinical portion? How do these procedures relate to the scoring procedures?

Standardization is needed for clinical tests to assure fair and uniform treatment of each candidate. Double-blind grading is recommended as the preferred scoring procedure to assure uniform and fair testing.

- What test statistics are needed for clinical items? Can test statistics be computed in the same way as for paper-and-pencil tests or other kinds of performance tests? What special procedures are needed to set a cut score for the clinical portion and how do these relate to continued testing using part credit?

Reliability as indicated through inter- and intrarater agreement, explicit criteria used in the determination of satisfactory test performance, and cut scores are the

primary statistics needed in clinical testing. Logistics may prevent their being calculated in the same manner as paper-and-pencil testing, yet pretesting is encouraged.

REFERENCES

Allen, D. L. (1992). Standardized national dental clinical examinations. *Journal of Dental Education, 56*(4), 258-261.

Butzin, D. W., Finberg, L., Brownlee, R. C., & Guerin, R. O. (1982). A study of the reliability of the grading process used in the American Board of Pediatrics oral examination. *Journal of Medical Education, 57*(12), 944-946.

Crehan, K. D. (1974). Item analysis for teacher-made mastery tests. *Journal of Educational Measurement, 2*(4), 255-262.

Ebel, R. L. (1951). Estimation of the reliability of ratings. *Psychometrika, 16,* 407-424.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.; pp. 105-146);. Macmillan: New York.

Fortune, J. C. (1991, July). *Report on the analysis of examination performance across a six state licensing area.* Unpublished report commissioned by the Southern Regional Testing Agency, Virginia Beach, for the Virginia Board of Dentistry.

Friedman, C. B., & Ho, K. T. (1990, April). *Interjudge consensus and intrajudge consistency: Is it possible to have both in standard setting?* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Boston, MA. (ERIC Document Reproduction Service No. ED 322 164)

Gross, L. J. (1993, Winter). Assessing clinical skills in optometry: A national standardized performance test. *CLEAR Exam Review,* pp. 18-23.

Gross, L. J. & Showers, B. (1993). *Principles of fairness: An examining guide for credentialing boards.* Lexington, KY: Council of State Governments.

Hopkins, C. D., & Antes, R. L. (1990). *Classroom measurement and evaluation.* Itasca, IL: Peacock.

Impara, J. C., & Plake, B. S. (Eds). (1995). Standard setting for complex performance tasks [Special issue]. *Applied Measurement in Education, 8*(1).

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed; pp. 485-514). New York: Macmillan.

Kenyon, D., & Stansfield, C. W. (1991, April). *A method for improving tasks on performance assessments through field testing.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL. (ERIC Document Reproduction Service No. ED 334 226)

Maust, A. P., Callahan, D., Fortune, J. C., & Cromack, T. R. (1988). *An evaluation of the licensing examination function, Virginia Department of Commerce.* Alexandria, VA: Research Dimensions, Inc.

Medley, D. M., & Mitzel, H. E. (1964). Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), *Handbook of research on teaching* (pp. 247-328). Chicago: Rand McNally.

Millman, J. (1989). If at first you don't succeed: Setting passing scores when more than one attempt is permitted. *Educational Researcher, 18*(6), 5-9.

Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed; pp. 335-366) Macmillan: New York, .

Minnich, R., (1992). *Examiner's manual and orientation.* Virginia Beach, VA: Southern Regional Testing Agency, Inc.

Schroeder, L. L. (1993). *Development, administration, scoring and reporting of credentialing examinations: Recommendations for board members.* Lexington, KY: Council of State Governments.

Trebilcock, M. J., & Shaul, J. (1983). Regulating the quality of psychotherapeutic services. *Law and Human Behavior, 7*, 165-278.

Vu, N. V., & Barrows, H. S. (1994). Use of standardized patients in clinical assessments: Recent developments and measurement findings. *Educational Researcher, 23*(3), 23-30.

Weiss, J. (1987). The Golden Rule bias reduction principle: A practical reform. *Educational Measurement: Issues and Practice, 6*(2), 23-25.

Winer, B. J. (1971). *Statistical principles and experimental design.* New York: McGraw Hill.

Yaple, N., Metzler, J., & Wallace, W. (1992). Results of the Ohio non-patient dental board examinations for 1990 and 1992. *Journal of Dental Education, 56*(4), 248-250.