Standardized Oral Examinations

Measurement Research Associates, Inc.
505 N. Lake Shore Dr. Suite 1317
Chicago, IL 60611
Phone: 312-822-9648
Fax: 312-822-9650

Standardized Oral Examinations

Oral examinations are included in the certification process because clinical application skills can be tested more effectively. A number of decisions about the structure, administration, and scoring of the examination must be made to preserve the flexible nature of the oral examination, while simultaneously providing the standardization that will facilitate fair comparison of candidates. The primary purpose of the oral examination is to identify *differences* in candidate ability. Standardization enhances the consistency and fairness of the scoring. Oral examinations are constructed to provide individualized examinations of the candidate's ability to apply their knowledge and skill in the clinical environment. With oral examinations there must be enough standardization to compare candidates, but also enough flexibility to allow examiners to tailor the examination to the each candidate. There are many options for structuring orals and their scoring, so that differences among candidates can be described in a standard and consistent way.

Standardization of oral examinations occurs through organization, scoring, and statistical analysis. There are many ways to organize oral examinations without compromising their validity and individualized nature. Cases, protocols, skills, or guided questions are often standardized, that is, written by subject matter experts, reviewed and perhaps even pretested. Examiner training may help to standardize the way examiners use the scoring system and the examination stimulus materials. Organization controls the type and amount of information collected about each candidate. However, scoring is the primary source of standardization. The scores earned define the differences and similarities among candidates.

The Oral Examination Process

There is no "typical" process for an oral examination, but there are often four or five examination facets that influence the outcomes of most oral examinations.

Candidates are the first facet. The purpose of the examination is to document variation in ability among candidates.

Examiners are the second facet, and are essential for assessing the quality of candidate performance. Each examiner is a unique person with unique perceptions and expectations for satisfactory clinical performance. Examiners must be trained in the administration and scoring of the oral examination. It is, however, highly likely that examiners will maintain standards for acceptable practice that have been established through their education and experience over a lifetime. For this reason, several examiners provide independent assessments, and a statistical correction to eliminate varying examiner severity and bias is desirable.

The third facet is the cases, problems , guided questions or other stimuli that serve to present the content and stimulate the candidate to demonstrate his/her knowledge and skill. Board constructed standardized cases present a specific patient's signs and symptoms, with x-rays, laboratory reports and other pertinent information available to the candidate on request. Another alternative is for candidates to present selected cases from their practice, or to work with actual patients.

A fourth facet is the tasks or skills necessary for safe and effective practice. Among the pertinent skills tested may be diagnosis, treatment, management, technical skills, interpretation, patient assessment, problem solving, and clinical judgement.

A fifth facet may define administrative differences, such as, days of the week, times, or locations at which the examination is administered, sessions, or any other factors.

The rating scale, the number of rating categories, and the meaning assigned to those categories are critical for scoring and interpretation. Candidates may earn pass/fail ratings , or they may be

rated on a scale of one to four or one to 100 points. The **number** of ratings given to each candidate by each examiner affects score precision. A single rating provides little documentation of the candidate's strengths and weaknesses, while several ratings provide more specific information about the candidate. The definition of the facets determines how the examination is organized, and the meaning associated with the outcomes. The type and number of ratings standardize the conditions of observation (skills, cases, questions) and determine the amount of information collected about each candidate.

Content Validity

Maintaining the content validity of the oral examination is essential. Certification boards establish the content specifications for constructing oral examination stimulus material to ensure content validity. If a blueprint for the content of the oral examination is not in place, one should be developed. The blueprint will likely reference the clinical content areas that are cross-referenced with the skills being tested. Protocols, cases, or problems, can then be developed and classified. Candidates can be instructed to submit cases in pertinent content domains.

Some clinical problems are more difficult than others, so the difficulty should be assessed both subjectively by the content experts, and statistically for verification. . The following may be considered when examination materials are developed and selected. 1) What criteria are used to insure proper content sampling? 2) What are the stimulus materials designed to measure, and how is their relative difficulty determined? 3) If different cases are used on different days or sessions, how are they matched for consistency and comparability?

Decision Reliability

In order to score candidates reliably, as much information as possible should be collected. Generally, more ratings provide more information about the candidate's ability so that a more precise estimate of the candidate's ability can be calculated. When only one holistic rating is recorded, the error of measurement is very large. As more ratings are recorded, the error of measurement decreases, so there is more reliability and more confidence in the outcomes. Therefore, a careful assessment of how many ratings are given to each candidate, and the rationale for giving those ratings is an essential part of standardization. The meaning of each category on the rating scale must be defined, because the interpretation of the final score relates directly to the definitions associated with each scoring category. The following issues should be considered with regard to scoring: 1) on what do examiners rate candidates; 2) how are scores combined or how is the overall score derived; 3) how do the scores relate to the intended purpose of the measurement; and 4) how are the scores used to support the reliability of the oral examination?

Standardization through Calibration of the Examination Facets

Calibration locates the relative ability, difficulty, or severity of the elements of each facet of the examination, so that their relationship with regard to difficulty, severity or ability can be observed. The Rasch-based multi-facet model (Rasch,1960/1980; Linacre, 1989; Wright and Stone, 1979) performs this statistical analysis. The propositions of the multi-facet model, predict that when the case and skill together are *more difficult* than the candidate is able, after the severity of the examiner is taken into account, the candidate has less than a 50% probability of earning a passing rating. If the candidate's proficiency *exceeds* the difficulty of the case and skill, after the severity of the examiner is taken into account, the model predicts that the candidate has a greater than 50% probability of earning a passing rating. It also predicts higher ratings on easier cases or skills, and higher ratings for more able candidates. These basic assumptions are commensurate with the primary oral examination goal of identifying differences among candidates.

Other methods of analysis, such as inter-judge correlation (see Lunz and Stahl, 1992 and Lunz, Stahl, and Wright, 1994) , do not account for the impact of the difficulty of the stimulus material, or the severity of the particular examiners encountered by the candidate. Methods such as generalizability theory predict the same sources of variability in the facets of the oral examination, but do not use that information in the calculation of candidate ability. Therefore, candidate results are dependent on the severity of the examiner who did the rating and the difficulty of the cases encountered (Lunz and Schumacher, 1997).

Calibration is accomplished by using **all** of the ratings given to **all** candidates by **all** examiners on **all** stimulus materials during an oral examination, then analyzing or calibrating each facet independently, but on a common scale. Examiner **severity** is the term used to encompass all of the characteristics of an examiner, and is calibrated using all of the ratings given by an examiner during the examination. Item, case, problem, question, or project **difficulty** is calibrated using all ratings given by all examiners to all candidates. The difficulty calibration can be validated through verification by content experts.

Figure 1 shows an example of the pattern of statistical facet element calibrations. The facets shown are candidates, examiners, protocols and skills. A * represents each element in the vertical histograms. The first column shows the distribution of candidates by ability. The second column shows the distribution of examiners by severity. The third column shows the distribution of protocols by difficulty, and the fourth column shows the distribution of skills by difficulty. Using this map, it is possible to track the difficulty of the oral examination taken by each candidate.

Statistical calibration methods are currently used on a number of nationally administered certification examinations for medical specialties, allied health, dentistry, and education. Raw scores do not serve as well, because the severity of the examiner(s) who gave the ratings, and the difficulty of the particular stimuli are *not* considered. When oral examination facet elements are calibrated, differences in the individual candidate oral examinations can be identified and taken into account in the candidate scoring process. .

The ratings given to a candidate must be independent among examiners and skills, even if some skills are conceptually related. Another factor is the difficulty of the stimulus. More difficult stimuli earn lower ratings, regardless of their conceptual relationship to the other materials. Examiners must give the candidate ratings independently, without collusion among examiners or skills. For example, a candidate may miss the diagnosis and earn an unsatisfactory rating. The examiner informs the candidate of the correct diagnosis. Subsequently, the treatment suggested by the candidate is acceptable and an appropriate. An independent rating for treatment is then given. Independent ratings are crucial when collecting information about the candidate, because it is from these ratings that candidate ability among skills and on the total test is determined.

Selecting Administration Parameters: **How Much Standardization is Enough**

The goal is to determine skills, tasks, topics, guided questions or other stimulus that provide the most information about the candidate. It is advisable to set up a data-collection plan that allows examiners to give a minimum of three ratings per candidate. Pertinent areas in which the candidate is rated must be defined explicitly and agreed upon by the examiners. Then all examiners must assess candidates on those three stimuli.

Ratings and Rating Scales

In order to achieve any consistency, the rating scale must be standardized. All examiners must use the same rating scale and understand the definitions for each rating scale category. Candidate ratings are the data for interpreting candidate performance, comparisons and outcomes. The rating scale attaches scores to the quality of the observed candidate performance. The rating scale provides the opportunity for a disciplined dialogue between the quality of the performance and the scoring. The score has meaning

based on its reference to the quality of candidate performance, as represented by the definitions established for each category on the rating scale.

All examiners must be familiar with the rating scale, the meaning of the categories, and have some training and experience using it. The number of rating scale categories determines the specificity of the distinctions in candidate ability that examiners record. A two category rating scale is a gross generalization of candidate performance, while a 100 point rating scale requires distinctions too specific to observe . The rating scale categories are the distinctions in performance quality. Candidate ability estimates are often based on the sum of ratings or scores. When a candidate earns a rating equivalent to 'satisfactory', one assumes that his/her performance on that case or skill was perceived to be 'satisfactory.'

Holistic and Analytic Ratings

The precision and accuracy of the candidate results determines the level of confidence in the decisions. Holistic scoring requires the examiner to process all aspects of the candidate's performance simultaneously, and make one gross or global rating. Analytic scoring requires examiners to make a series of explicit ratings for a representative sample of cases or skills.

Analytic ratings break down the scoring process into distinct assessments, and thus produce independent ratings from a representative sample of skills. More ratings provide more information about the candidate, lower measurement error, and more confidence in the accuracy of candidate scores and subsequent outcomes (pass or fail). Several ratings are more likely to be closer to "true ability" than one single assessment.

Analytic scoring requires the examiner to make detailed, but independent assessments. Analytic ratings are usually associated with the specific tasks, skills or topics, etc. A candidate may earn a high rating on diagnosis because the diagnosis is correctly determined, and a lower rating on treatment, because the treatment selected by the candidate is not currently considered the best alternative.

Holistic scoring forces examiners to balance the importance of all cases, skills, guided questions, etc., and then incorporate this information into one all encompassing rating. Holistic judgments are *very subjective*, making it impossible to separate the score from the impressions of the particular examiner. Another examiner may rate the same candidate quite differently and award a substantially different holistic score. The rating is irrevocably connected to the perception of the individual examiner. There is a lot of measurement error, so confidence in the accuracy of the pass or fail decisions is often low. This is why oral examinations are generally considered unreliable.

Content-Related Materials

Content-related materials may take many forms, such as, projects, cases, clinical scenarios, practice based cases, work sample projects, case lists, protocols, guided questions, or numerous others. Content distribution is most closely controlled when content experts prepare materials according to specific guidelines. This also limits the content that is covered. The same standardized materials are used to examine all candidates. Examination forms are made comparable to insure coverage of the content specifications. Generally, standardized protocols or cases describe a specific patient and the candidate diagnoses and treats that patient. Standardized questions may be asked of all candidates. The examiners may inform the candidates of errors to insure that candidates continue through the standardized case as established. A potential downside is that the oral may dissolve into an oral "recall" examination, instead of a clinical application test. The number of content areas covered can be representative, but may be limited.

Somewhat less control occurs when candidates present selected cases from their practice, or when examiners select the case materials from their practice to cover specific content areas, or students create portfolios of their work using specified guidelines. As control of the examination materials decreases, the basis for rating candidate performance must be more explicit to insure consistent measurement among candidates.

Skills or Other Ratings

The skills or tasks, on which the candidate is rated, standardize the scoring regardless of the content. The specifics of the content tested may vary among candidates; however, the skills are common to all candidates. Skills or tasks may serve as the criteria for rating candidates. Clinical skills such as diagnosis, management, problem solving, and clinical judgment are pertinent, regardless of the case-specific problem. Analytic skill ratings provide consistent and organized information about the candidate.

Examiners

Standardizing examiners through extensive training and re-training is virtually impossible, but inter-judge reliability coefficients have often been considered as an indication of oral examination reliability (LeMahiew, Gitomer, and Eresh, 1995). The literature suggests that examiner training is not completely successful. In fact, even if examiners do correlate perfectly, there is still no guarantee that they will rate candidate performance comparably (see Lunz, Stahl, and Wright, 1994 and Lunz, 1992). Examiner training usually includes an introduction to the examination materials, information on how to conduct the examination, and a description of the rating scale points and their meaning. Asking examiners to provide documentation for their scores may provide some insight into the thinking of the examiner at that time.

In summary, the standardization provides organization of the material presented and the scoring procedures. In order to compare candidate performances, it is necessary to rate all candidates using the same criteria and the same rating scale. The statistical calibration of the stimulus material, and examiner severity enhance the standardization of the oral examination by reducing the bias caused by the individual characteristics of each oral examination.

Standard Setting for Oral Examinations

Establishing a criterion or norm referenced standard for passing and failing candidates, standardizes the process of making pass and fail decisions. The information about candidates is supplied by the examiners, but the Board sets the standard. Historically, a criterion or norm-referenced passing standard was not been established for oral examinations. Rather, the wisdom and experience of individual examiners determined whether candidates passed or failed the examination. Different examiners made different decisions about the same candidate due to their expectations from personal experience and/or pressure from their peers. This lack of standardization in decision making is a major reason why oral examinations have been considered unreliable.

However, recent research (Lunz, in press) has identified several potential methods of standard setting when analytic ratings are collected, and oral examinations are analyzed using the multi-facet model (Linacre, 1990). The standards established are independent of the particular examiners, and/or the specific stimulus material presented. It is possible to set criterion or norm-referenced standards only when a sufficient amount of information about the candidate has been collected, so the error of measurement is small enough to have confidence in the accuracy of the pass or fail decisions.

A norm-referenced standard can be established by identifying a pre-score or reference group, calculating their mean, standard deviation and standard error of measurement for their total test scores. The standard can be applied to all candidates. Candidates who meet the standard pass, while the others fail.

Criterion referenced standards are difficult to establish for oral examinations. Each candidate experiences a different form of the oral examination, because different stimulus materials are presented and rated by different examiners. Oral examinations are temporal, making it unlikely that the exact conditions of scoring can be reconstructed. The only concrete record of the examination is the ratings given by the examiners. When analytic ratings are given, it is easier to understand the performance of the candidate and the perception of the examiner during the examination. It is also possible to establish criterion-referenced standards.

Criterion-referenced standards for oral exams have been established in several ways. The first alternative is the creation and scoring of a "synthetic" candidate. This hypothetical candidate is rated by the committee of experts. The collective expectations of how a candidate, who will just pass the examination, determine the ratings. These ratings must take into account the difficulty of the stimuli presented, and an examiner of moderate severity. The "synthetic" candidate ratings are then scored along with the candidates. The estimated ability of the synthetic candidate should be in the region of the criterion-referenced standard. This method is most useful for portfolios or other exams for which the candidate produces examples of their work, which can be assessed by a standard setting committee.

A second alternative is the "fair average," when the multi-facet analysis is used. After accounting for the difficulty of the candidate's examination form, the estimated candidate ability is translated back to an average score by the Facets program (Linacre, 1990). It is reasonable to expect a passing candidate to earn a mean score that represents a rating of satisfactory when the differences in the difficulty of examination forms are accounted for. If raw scores are used, average scores are confounded by the severity of the examiners and/or the difficulty of the stimuli presented to the candidate. The examiners must agree, a priori, on the characteristics that represent satisfactory performance. Thus, a fair average that represents the point between satisfactory, and less than satisfactory, designates the region of the pass point.

A third alternative for setting criterion standards involves assessment of expected performance on standardized cases, protocols, or questions. This is similar to a modified 'Angoff' method for multiple choice tests. The mean of the expected performance of the minimally competent candidate, across all protocols, can be used to identify the region of the criterion standard. Of course, adjustments to any standard can be made based on the error of measurement to avoid Type I or II error. With any of the above standard setting procedures, the certification Board, rather than individual examiners, sets the standard and applies it consistently to all candidates.

Standardized Reporting of Oral Examination Results

Because candidates take different oral examinations, boards must decide what type of information to report to candidates. The following are suggestions. 1) Verbal notification of pass or fail results only. If holistic ratings are used, *only* pass or fail decisions can be reported. 2) Scaled scores for the total examination and/or content or skill areas can be reported to failing candidates only, or to all candidates. 3) National percentile ranks and/or other comparative distributions can be calculated. It is not advisable to report raw scores for oral examinations, because they do not account for the unique characteristics of the individual examination forms taken by candidates. Candidate reports require a sufficient number of ratings to calculate scores, and reduce the error of measurement. This provides confidence in the accuracy of the scores reported.

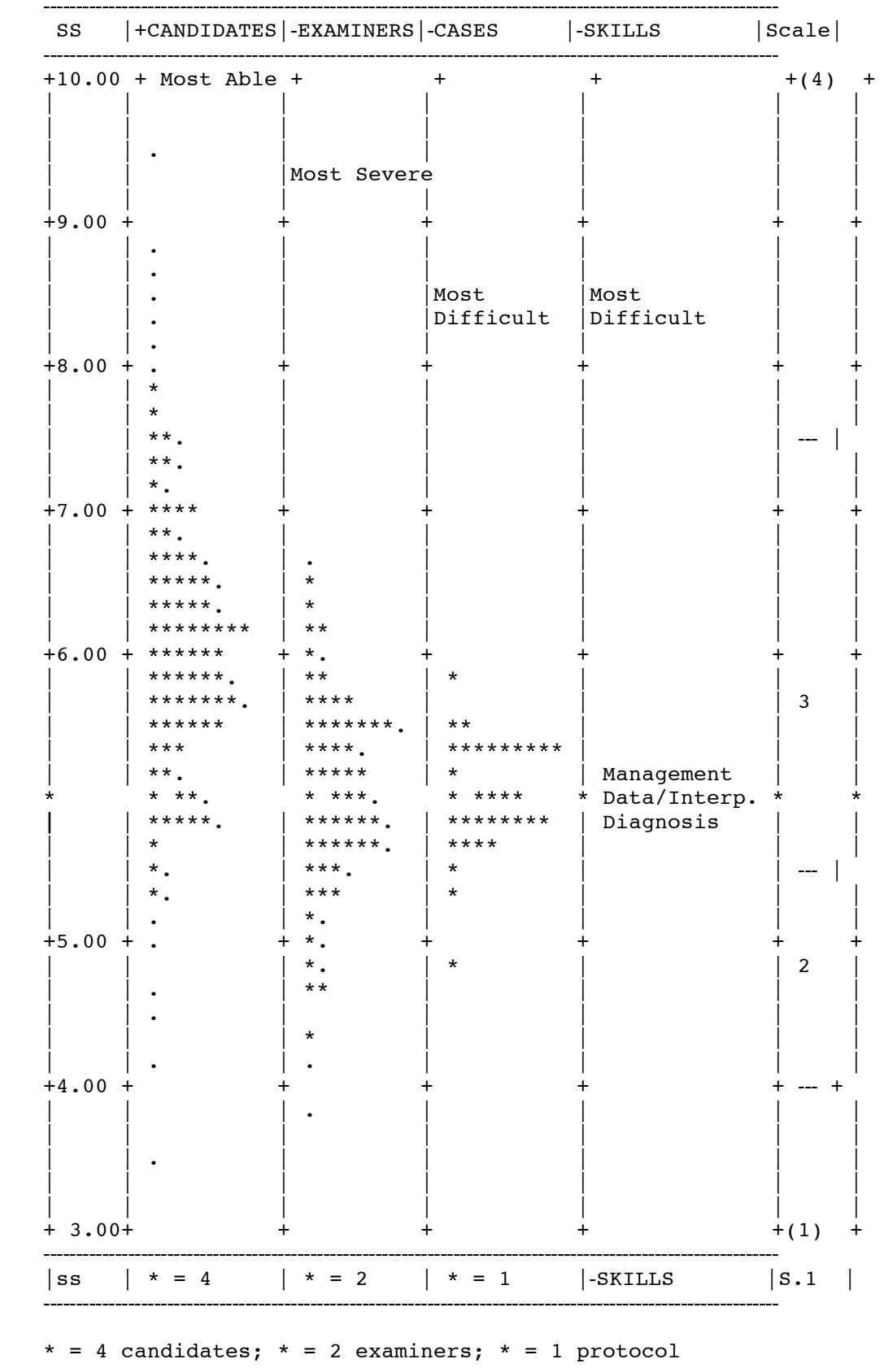
Conclusion

The standardization of oral examinations requires a great deal of planning, from conception to reporting. Standardization of content materials, skills, ratings scales, and scoring, combined with examiner training, improves the validity and reliability of oral examinations. Statistical calibration, using the multi-facets analysis, provides information about the nature and form of the examination. The combination of facets that create each individualized examination form influence the ratings earned by candidates. The ratings are the *documentation* of candidate performance and support the validity and reliability of the oral examinations. Establishing a criterion-referenced standard for passing or failing, gives all candidates a comparable opportunity to pass, if they have the requisite knowledge and skill. Standardized reporting to

candidates completes the process.

Overall, standardization improves the oral examination process, without impinging upon the clinical flexibility of the oral examination. The measurement accuracy of the examination is enhanced, as well as, the consistency and the fairness of the process of evaluating candidates.

Figure 1
Map of Oral Examination Facets
Show relative difficulty of cases and skills compared to examiner severity and candidate ability.



* = 4 candidates; * = 2 examiners; * = 1 protocol

References

Le Mahieu, P., Gitomer, D. & Eresh, J. (1995). Portfolios in large-scale assessments: Difficult but not impossible. Educational Measurement: Issues and Practice, Vol. 14, 3, 11-28.

Linacre, J. M. (1989). Many-faceted Rasch measurement...Chicago, IL: MESA Press.

Linacre, J.M. (1990). FACETS: A computer program for multi-facet ... Chicago, IL: MESA Press.

Lunz, M.E. (In press.). In *Objective Measurement: Theory into Practice* (eds. Engelhard. G. and Wilson, M.).

Lunz, M. & Stahl, J. (1992). New ways of thinking about reliability. Professions Education Researcher Quarterly, 13, 4, 16-18.

Lunz, M., Stahl, J. & Wright, B. (1994). Interjudge reliability and decision reproducibility. Educational and Psychological Measurement, 54, 4, 913-925.

Lunz, M. & Schumacker, R. (1997). Scoring and analysis of performance examinations: A comparison of methods and interpretations. Journal of Outcome Measurement, Vol. 1, 3, 210-238.

Rasch (1960/1980). *Probabilistic Models for Some Intelligence and Attainment Tests.* Chicago, IL: University of Chicago Press.

Wright, B. & Stone, M. (1979). Best Test Design. MESA Press.