Purposeful Assessment Techniques (PAT) 
Applied to an OSCE-Based Measurement of Competencies in a Pediatric Dentistry Curriculum


Abstract: Careful measurements of knowledge, attitude, and psychomotor and communication skills are necessary components of testing in a competency-based approach to education in dentistry. In an effort to address these requirements, Baylor College of Dentistry (BCD), Dallas, Texas, has applied Purposeful Assessment Techniques (PAT) to the Objective Structured Clinical Examination (OSCE) currently in use. PAT are those techniques that allow one to work toward development of linear measurement scales that are “person-free” and “item-free.” Person-free measurement means that useful data are produced regardless of the group being measured, and item-free measurement means that it does not matter which mix of items is completed over the course of an assessment. The Rasch probabilistic model and a guiding definition of Objective Measurement were used in an effort to implement PAT for the BCD OSCE. A Rasch analysis of a BCD-administered OSCE produced an item map that demonstrated the range of difficulty of items by student performance. This item map can be used to determine which items can be repeated on subsequent tests to allow for linear measurement of students’ progression through the curriculum. The movement toward PAT described in this paper demonstrates how careful and evolving measurement in dental education can be of great benefit to faculty, staff, students, and the public.

Dr. Boone is Associate Professor of Science and Environmental Education and Coordinator of the Program in Science and Environmental Education, Indiana University-Bloomington; Dr. McWhorter is Associate Professor and Dr. Seale is Regents Professor and Chairman, Department of Pediatric Dentistry, Baylor College of Dentistry, Texas A&M University System Health Science Center. Direct correspondence to Dr. N. Sue Seale, P.O. Box 660677, Department of Pediatric Dentistry, Baylor College of Dentistry, Texas A&M University System Health Science Center, Dallas, TX 75266-0677.

Key words: assessment, competencies, OSCE, Rasch analysis

Submitted for publication 7/18/01; accepted 9/12/01

The movement towards competency-based education in dentistry challenges traditional testing techniques because careful measurements of knowledge, attitude, and psychomotor and communications skills are required. This paper describes Purposeful Assessment Techniques (PAT) as they are being applied to the Objective Structured Clinical Examination (OSCE) testing format currently being used to measure student performance in the Department of Pediatric Dentistry at Baylor College of Dentistry, Texas A&M University System Health Science Center (BCD), Dallas, Texas. The OSCE examination format uses multiple stations designed to be interactive and to test mastery of various skills. Examinees rotate through a circuit of stations at which they are asked to demonstrate a variety of skills.1

The assessment techniques used at BCD have resulted from the continued use and refinement of the OSCE testing format, the need to continually evaluate competency, and the necessity of objective student measurement. Although the assessment strategies outlined in this paper have grown out of using an OSCE, the techniques are of immediate value to all involved in the use of assessments—be they students, faculty, administration, policymakers, or the public.

What do we mean by PAT, and why is this issue of importance in dental education? We define PAT as those assessment techniques that help provide valid and reliable measures that are informative to students, faculty, administrators, and the public. Furthermore, PAT are those techniques that allow one to work toward the development of linear measurement scales that are “person-free” and “item-free.”2,3 Linear scales are those that express performance (i.e., one set of students) on a metric that allows parametric statistical tests, such as the widely used t-test and ANOVA analysis of variance, to be used. Person-free measurement means that useful data results, regardless of the group being measured. Thus, it does not matter how competent or incompetent a group is in terms of performance when completing a test. Useful, valid, and reliable person-free measurement can be conducted regardless of the cadre of students. Item-free measurement means that it does not mat-
ter which mix of items is completed over the course of an assessment. Thus, within an OSCE framework, item-free measurement occurs when students can be compared regardless of the mix of stations completed by test takers. The following sections of this paper will concern: 1) the importance of PAT when considering the issue of dental competencies, and 2) selected strategies used by Baylor College of Dentistry to move toward PAT within an OSCE framework.

A competency-based approach to education implies that learning is a continuum; it involves a linear progression. Competency-based dental education involves the identification of those terminal competencies that a safe beginning dental practitioner must possess. Terminal competencies are composed of subcompetencies and foundational knowledge that the dental student must master in the process of movement toward readiness for unsupervised, entry-level practice. Methods must be in place to measure the degree to which students can demonstrate progress toward the terminal competencies. As a result, measurement tools must take into consideration a range of issues related to this process. First, a class of dental students is not homogenous. Within a class there will be a mix of students, with some more advanced and some slower in their mastery of curricular components. Second, because the educational process is a continuum over four years, multiple examinations are required. These examinations must provide a continuous measurement of student progress (mastery of increasingly complex subcompetencies) through the educational process. Well-developed measurement tools with good PAT are mandatory for competency-based education, as they will measure students with the same accuracy regardless of their learning (person-free) and be compact yet sensitive enough to measure the students’ progression through the curriculum (item-free).

In an effort to implement PAT for the OSCE, the probabilistic Rasch model⁴ was utilized, as well as a guiding definition of Objective Measurement.⁵ Objective Measurement can, in part, be viewed as the repetition of a unit amount that maintains its size, within an allowable range of error, no matter which instrument is used and no matter who or what relevant person or thing is measured.⁵ This definition reminds researchers of the importance of developing and monitoring measurement instruments. Members of the research community in industry, medicine, and education have determined that this model helps in the development of valid measurement instruments. Since the development of this measure-ment model by Rasch⁴ and its subsequent expansion by Wright and others,²,³ a wide range of investigators have utilized the model to create tests and calculate test-takers’ performance. In Australia, the model has been widely used to evaluate and communicate student performance to parents.⁶ More than twenty years ago in the United States, George Ingebo developed a Rasch-based test items bank for a school district in Portland, Oregon. In recent years the Rasch model has been used for many large-scale projects such as the Third International Math and Science Study and for the long-term evaluation of reform in public schools.⁷ Many dental/medical groups have used this model for board certification and/or recertification (American Dental Association, American Board of Pediatric Dentistry, American Society of Clinical Pathologists, American Board of Medical Examiners).

The Rasch Model

The Rasch model developed by the Danish mathematician George Rasch⁸ is a probabilistic model and differs from many other models that are commonly deterministic. The equation that describes this relationship between items and person is expressed by: $\log (Bn-Di) = P(Bn-Di)/1-P(Bn-Di)$ where $Bn$ is the ability of person “n” and $Di$ is the difficulty of item “i.”

Inspection of the model equation quickly reveals that it is based upon probabilities, that persons and items are expressed in the same units (called logits), and that the probability of a person correctly answering an item is solely related to his or her ability and the difficulty of the item being answered. The model is the one that meets Thurstone’s⁹ requirement for scale validity, and Norman Campbell’s requirement for fundamental measurement, which is that the combination of $A$ and $B$ must be more than $A$ or $B$ and that $A$ and $B$ are linked to produce $A + B$.¹⁰ Specific details about the model are provided by Wright and Stone,² Wright and Masters³ and Andrich.¹¹ The authors of this study view the Rasch model as a definition of measurement that is easily understood and exceedingly useful.

Beyond Rights/Wrongs

Assessments of students usually result from the calculation of a total raw score. Thus, on a 100-item
test, a particular student’s performance is often based upon the student’s correct raw score (e.g., 83 of 100). However, there are many deficiencies in the use of the raw score total. One particular problem is that, by emphasizing raw scores, it is easy to ignore that there are many issues greatly influencing a student’s raw score—for instance, the mix of items presented on a test in terms of difficulty. Figure 1 provides an item map constructed following a Rasch analysis of an OSCE administered by the Department of Pediatric Dentistry at BCD in the middle of the second-year preclinical course designed to prepare students to enter the clinic in all areas of pediatric dentistry (T1). The Xs to the left of the central vertical line represent the distribution of students taking the test. The letter/number combinations to the right of the vertical line correspond to items on the examination. The item map provides the ordering of items according to student performance, from more difficult to less difficult, as if reading a thermometer. For instance, item 6E was the most difficult item for this group of students, while items 13A through 9M at the bottom of the figure were the easiest items on the test. Item ordering tells how students performed on that particular examination. Faculty members generally have an idea of how difficult an individual item will be for the students, and it can be surprising and revealing when an item previously identified by faculty as easy appears on the item map as difficult. For instance, a question involving a visual and manual examination of a tooth that had a sealant placed on the occlusal surface required the student to determine the adequacy of the sealant placement. The faculty had assessed the item as easy prior to administration of the examination; however, this item appeared on the item map as the most difficult item on the examination out of a total of ninety-six items. In another case, an item required the performance of a manual skill in which the student used a handpiece to prepare a pulpotomy access on a typodont tooth. The faculty anticipated that this station would be difficult for the students in a time-restricted setting, but it appeared on the item order map as a moderately easy item. In the first situation, possible explanations for the item appearing more difficult could include inadequate teaching of the item/topic and/or student confusion over a poorly written question. In the second situation, student performance indicated a level of mastery beyond faculty expectations.

This visual presentation (Figure 1) provides an overview of the actual curriculum, and the construction of an item map has helped greatly in the movement toward PAT. Plausible explanations for items appearing easier or more difficult than anticipated have implications for curriculum design and the preparation of subsequent examinations. In the case of the sealant question, originally thought to be easy, identifying that students had not mastered the item meant that the associated subcompetency needed to be retaught and retested before moving on in that component of the curriculum. On the other hand, identifying that a concept had been more fully mastered than expected allows that portion of the curriculum to move forward and prevents unnecessary repetition. Feedback from the item maps allows curricular adjustments that accommodate the most recently measured student performance.

### Developing New Tests

As part of the BCD effort to implement PAT, the data presented in Figure 1 were also utilized to: 1) revise and develop subsequent tests administered to students, and 2) revise the initial test completed by students. The data facilitate the creation of subsequent tests by allowing faculty to select items for inclusion in later tests that can be used longitudinally to measure student performance. In the spring of 2001, the faculty reviewed Figure 1 and selected items of medium and high difficulty for inclusion in the OSCE (T2) which was to be taken by those same students six weeks following administration of the initial test (T1). By selecting items from the initial test, it is possible, through a technique called test linking,² to express the performance of students on the same metric regardless of the test that was completed. This facilitates a longitudinal evaluation of students. Selected items from T1 that represented skill sets of particular importance to the students’ preparedness to enter the clinic, such as ability to take radiographs or a medical history on a child patient, made ideal links between T1 and T2 in the BCD preclinical course.

By linking tests through a careful repetition of selected items, it is possible to monitor the students’ increased level of mastery of curricular content and measure their grasp of subcompetencies and foundational knowledge. Improved student performance on an item identified earlier as difficult is often in-
Figure 1. Person/item map from Rasch analysis of test data from an OSCE administered to second-year dental students at BCD

Students' distribution is shown to the left, and item distribution is shown to the right of the central vertical line. Items near the top are considered “harder” items for this group of students, and items on the lower end of the map are considered “easier.”
dictative of their progress through the curriculum as they move toward the terminal competencies of pediatric dentistry. Additionally, areas of confusion can be identified, retaught, and retested.

Careful Steps

In almost all cases of test development, decisions are made with good intention. However, it is critical to point out that, in many cases within traditional tests, the ramifications of key decisions may often be overlooked. In the OSCE format, judges for instance, are often used to evaluate students’ performance at stations. Due to the number of candidates completing the Baylor OSCE, the examination time constraints, and the space needed for stations, the Baylor OSCE often is administered to two groups of students simultaneously. This means that during a particular OSCE administration, two sets of identical stations are created. To envision this mechanism, one can conceive of a group of twelve stations in one room for ten students and an identical set of twelve stations in another room for ten more students. Many stations require the presence of a judge (proctored station); therefore, because there are two sets of the OSCE being administered, two judges for each proctored station must be trained. In the past, BCD has worked toward requiring that judges use the same criteria for the evaluation of students and behave in the same manner when judging student performance. However, administrations of the OSCE using different judges for the same station have provided evidence that judges do not always behave in the same manner. One judge may be more severe in his or her assessment of student performance. For instance, at an OSCE station where the judge is acting as the child’s parent while the student takes a medical history, one judge may inadvertently “lead” the student through the process with positive body language, while a more severe judge may exhibit less encouraging body language. While neither judge would be intentionally varying from the standardized criteria for student assessment, their differences could affect the students’ performance on the two “identical” stations. Another example of differences in judge interaction may result from the manner in which the judges interpret the students’ performances. In situations where the student doesn’t perform the task perfectly, a more lenient judge may give the student the benefit of the doubt while a stricter judge may give partial credit.

A new judge rotation has been developed to monitor and overcome the problems of discrepancy in the behavior of judges, thus moving beyond an initial PAT level. A third standardized judge was added who moved between the two original judges and made an additional evaluation at the same time that the original judge was evaluating the student. Instituting an overlap of judges injects a degree of increased quality control into the assessment process. By overlapping judges, one can immediately compare the way in which they are utilizing uniform criteria used to evaluate student performance. This process takes into consideration the same range of issues revealed when BCD worked toward purposeful selection of items for inclusion in subsequent OSCE tests. Just as T1 of the Baylor OSCE is linked with later tests through the inclusion of common items, it is important to work toward ensuring that there is a purposeful overlap of judges at stations.

Assigning Item Weight

Decisions concerning the weight given to test items have a substantial impact on the scoring of the OSCE. In the T1, a total of ninety-six items was presented to students and graded by judges. Visual display of item weights in tandem with the data displayed in Figure 1 demonstrated the importance of creating a mechanism to assign weights to particular test items within the OSCE. The faculty realized that the decision to weight particular items is as important as the decision to include or exclude items. Weighting items can have a profound effect upon the final measures computed for each test-taker. As a result, it is important to develop a protocol to determine and evaluate the weights assigned to items. The development of a weighting protocol starts with an understanding that competency-based education is a continuum, and there will be variation in the relative importance of the subcompetencies to be mastered along the way. For instance, when the students prepare to enter the clinic, those subcompetencies that facilitate their immediate success in the clinic are the most important. Therefore, they should be weighted the most heavily at that point along the continuum, even though at the same time they are learning portions of the curriculum that seem more sophisticated and will have far greater value as they assume more responsibility for patient care later in their educational experience.
Once the appropriate questions have been identified, additional aspects of item weighting may be related to the number of parts to the question—for example, a proctored station may have a number of items to be judged during performance versus a manual skills station or problem-solving station that might include only two to three parts. Regardless of the time allocated for the station, the relative value of the test items at each station must be determined by considering the number of concepts involved, the manual difficulty of the task to be performed, and the complexity of the skill sets required for the item within a hierarchy of subcompetencies. Some stations may have two distinct components to the answer, while others may have only one. Each question component must be evaluated individually for its value to the test and the importance of the subcompetency or foundational knowledge that it measures. Simply giving equal value to all stations because they are all “four-minute stations” may result in disproportionate weighting of an item. Automatically assuming that a question with many parts in the answer should be weighted more heavily than a question with only one or two can also lead to disproportionate and inappropriate weighting of items.

**Conclusion**

The terms “validity” and “reliability” are often overused and at the same time misunderstood. Many types of validity and a wide range of issues impact the reliability of an instrument. Too often, when instruments used in dental education are discussed, one type of validity might be considered (e.g., face validity) and an overused calculation of a Cronbach Alpha reliability is made. Although such steps do provide some information, the steps outlined above that help develop PAT provide much greater insight for improvements in the validity and reliability of an instrument.

In summary, the Department of Pediatric Dentistry at BCD has used Rasch methods and a philosophy of objective measurement to carefully select test items for an OSCE that will allow longitudinal measurement of the students’ attainment of competencies. Faculty assessment of the OSCE led to the development of a judge rotation system to enhance scoring calibration and recognition of the need to weight test items.

**REFERENCES**

8. Thurstone LL. Attitudes can be measured. Am J of Sociol 1928;33:529-54.