21 Clinical Performance Assessments

EMIL R. PETRUSA
Duke University School of Medicine

SUMMARY

Evaluation of clinical performance for physicians in training is central to assuring qualified practitioners. The time-honored method of oral examination after a single patient suffers from several measurement shortcomings. Too little sampling, low reliability, partial validity and potential for evaluator bias undermine the oral examination. Since 1975, standardized clinical examinations have developed to provide broader sampling, more objective evaluation criteria and more efficient administration. Research supports reliability of portrayal and data capture by standardized patients as well as the predictability of future trainee performance. Methods for setting pass marks for cases and the whole test have evolved from those for written examinations. Pass marks from all methods continue to fail an unacceptably high number of learners without additional adjustments. Studies show a positive impact of these examinations on learner study behaviors and on the number of direct observations of learners' patient encounters. Standardized clinical performance examinations are sensitive and specific for benefits of a structured clinical curriculum. Improvements must include better alignment of a test's purpose, measurement framework and scoring. Data capture methods for clinical performance at advanced levels need development. Checklists completed by standardized patients do not capture the organization or approach a learner takes in the encounter. Global ratings completed by faculty hold promise but more work is needed. Future studies should investigate the validity of case and test-wise pass marks. Finally research on the development of expertise should guide the next generation of assessment tasks, encounters and scoring in standardized clinical examinations.

International Handbook of Research in Medical Education, 673–709.
G.R. Norman, C.P.M. Van der Vleuten, D.I. Newble (eds.)

INTRODUCTION

The assessment of clinical competence is a key step in assuring that health-care providers have sufficient knowledge and skills to carry out their professional responsibilities in a competent and safe way. Techniques for assessing clinical competence have evolved dramatically over the last 15 years. In particular, techniques have been developed to measure the dynamic doctor-patient relationship while at the same time assessing the learner's clinical performance. First called pseudo-patients (Barrows & Bennett, 1972), then simulated patients and now referred to as standardized patients (Anderson, Stillman, & Wang, 1994), these people, typically laypersons, are able to accurately portray problems of real patients such that advanced practitioners are unable to tell the difference between standardized and real patients (Rethans, Drop, Sturman, & Van der Vleuten, 1991; Gallagher, Lo, Chesney, & Christianson, 1997). A relevant sample of clinical challenges, portrayed by standardized patients and organized in an efficient circuit, may be wedded with modern measurement characteristics to produce a credible and high-quality assessment of clinical performance.

While SP-based examinations have increased in use, the time-honored method of oral examination remains the preferred assessment approach in many countries. With variations, the typical oral examination includes an experienced faculty examiner who observes a learner workup one real patient. When a history and physical examination are complete, the learner presents his or her findings, a differential diagnosis and the rationale for this list. The examiner, who may or may not have observed the history and physical examination, then asks a series of questions to probe the learner's understanding of signs and symptoms and pathophysiology as well as clinical reasoning and conclusions. Variations on this general scheme are called *viva voce* or oral examinations. From a measurement perspective and from the literature on clinical performance, there are major limitations to the traditional oral examination. Very different patients for different examinees, limited number of patient workups for each examinee and usually only one examiner undermine the measurement quality of the oral examination. In the last twenty years an approach has been developed to overcome almost all of these limitations. This chapter will focus on the variations of this new approach and key research findings regarding its measurement and educational characteristics. A number of excellent and more thorough reviews of standardized patients and SP-based examinations have been published (Van der Vleuten & Swanson, 1990; Colliver & Williams, 1993; Vu & Barrows, 1994; Van der Vleuten, 1996). The reader is directed to these reviews for additional understanding of the research supporting these methods.

In general, the new approach typically consists of several patients who have been trained to portray their case in the same manner for each examinee. This training is intended to have each portrayal be as standardized as possible for each examinee. Patients may actually have the disease or clinical problem they are portraying or

they may simulate them. Patients, real or simulators, who have been through training to produce consistent portrayals, are called standardized patients (SPs). Most variations of this approach have patients learn checklists of observable behaviors that are expected from examinees that the SPs or observers complete after the examinees finish the tasks of each station. Examples of the kinds of clinical challenges that may occur in a station are to interpret an X-ray, to attach leads for an electrocardiogram to a person, to describe a medical instrument and its purpose, to intubate a manikin and to instruct an SP mother about breastfeeding. These clinical challenges are arranged in a circuit for efficient administration (Harden, Stevenson, Downie, & Wilson, 1975; Harden & Gleeson, 1979). Examinees are assigned a starting point on the circuit and then all move clockwise or counterclockwise for their next station. Time allowed for each station is usually the same for everyone in the circuit. Examinees are signaled when to begin the next station. As an example, with a circuit of 12 stations each up to 15 minutes in length, 12 examinees' clinical performance would be evaluated on 12 different clinical challenges in approximately three hours. Oral examinations typically allow 60 to 90 minutes for the examinee to do the history and physical examination and then present to the examiner. An examinee's clinical performance could be evaluated with two patients in this approach in three hours. Developers of this new approach with all the variations must provide evidence that the new approach is better than established methods that purport to evaluate the same thing. This chapter reviews key evidence about the measurement and educational features of this multi-station approach for evaluating clinical performance.

Before presenting this evidence, a brief overview of measurement characteristics will be described. Primary among these measurement concepts are reliability and validity. Following this section, methods for setting pass/fail marks for cases and for a whole examination will be reviewed. Throughout these sections the consistency (or more accurately the lack of consistency) between models of clinical performance, materials in the examination, scoring and selection of measurement characteristics will be highlighted. Recommendations for important new research are addressed within each section.

GENERAL ISSUES IN ASSESSMENT AND MEASUREMENT

The two primary criteria on which tests, measurements or assessments are evaluated are reliability and validity. Reliability refers to the consistency of results. This consistency may be between two observers marking the same interaction (inter rater reliability), between two ratings of an interaction by the same observer at two different times (intra rater reliability), or between observed performance and an estimate of future performance (test reliability). There are several accepted approaches to estimating test reliability (Guilford, 1965; Linn, 1989). Challenges in an evaluation may be divided into two groups (e.g., first half-second half or odd-even). The correlation between scores on the two halves provides an estimate of test

reliability. Another approach is to estimate test reliability from the proportion of cases or challenges answered correctly. More recently generalizability theory is the preferred framework for estimating test consistency (Brennan, 1983). The major advantage of generalizability theory is that components of the assessment exercise are identified and their separate contributions to the total variance in scores are quantified. This separation of components allows for more precise refinement or better interpretations of the data.

A body of evidence, not a single calculation, should establish the validity of an assessment procedure, the degree to which an examination measures what it claims to measure. Evidence may bear on the quality of predictions of future performance, whether that future performance is on the same assessment or on different tasks (predictive validity). If relationships between the assessment outcomes and other features of the concept being assessed (e.g., "if clinical performance improves with more education, then learners earlier in the educational process would be expected to score lower than more advanced learners") are empirically supported, then the assessment is said to have "construct validity". A third type of validity, concurrent validity, is claimed if strong relationships can be demonstrated between an assessment and other assessments administered at the same time. Educational validity may be claimed if there is alignment of learning opportunities with tasks and challenges in the assessment. For example if clinical performance is conceptualized as *a set of skills,* e.g., relationship, communication, history taking, physical examination, ordering diagnostic studies and an assessment and follow-up plan, then these *skills* ought to be the focus of the test. However, if clinical performance is conceptualized as "adequately addressing the patient's problem(s)", the focus ought to be the *complete encounter* as the unit of measurement. For the skills concept of clinical performance, validity of the examination rests with whether these skills are the focus of challenges and the units of measurement. For the second concept, however, validity would not rest with separate skill, but rather with the quality of the full encounter. Certainly "skills" are part of this full encounter, but are considered *interwoven* with clinical knowledge, clinical judgment and the ability to communicate with the patient.

Can both concepts of clinical performance be "correct"? Yes, but not for the same level of expertise. Excellent work by Elstein and colleagues clearly demonstrated that experienced clinicians are continuously considering convergent and divergent historical and physical examination information that bears on the diagnoses they are entertaining at a given point in the workup (Elstein, Shulman, & Sprafka, 1972). Further, measurement research with patient management problems rejected a "skills" orientation for evaluating performance due to the non-independence of items within skills and skills within problems (Yen, 1993). Both of these lines of research used advanced clinical tasks, addressing a "patient's" problem. It is conceivable that, at much lower levels of expertise, all that is expected is the skill of structuring history questions, for example. Does the learner ask the seven key questions about the chief complaint? Does she use open-ended

questions early in the encounter? Are his questions unitary or does he string together three or four options as he asks the question? These evaluation questions focus on the *process* of taking a history and far less on how a learner directs the content of her questions, based on the patient's responses. If these evaluation questions are used with more advanced learners, but more advanced learners would be expected to invoke clinical knowledge and reasoning to guide the content and organization of their history questions while also using appropriate "skill" in conducting the medical interview, then the construct validity of the assessment is compromised. Thus, the validity of a clinical performance examination must be judged in the context of its purpose, the nature of performance expected from those being evaluated and the alignment of an examination's materials and scoring with the purpose and nature of expected performance. The quality of this alignment is one of the shortcomings of published work on clinical performance examinations. We will return to this issue of what ought to be evaluated at increasing levels of expertise later in this chapter. Now we will consider the frame of reference or the types of comparisons that may be made with examination results and the implications these frames of reference have for selection of various kinds of psychometric characteristics for the examinations.

Two types of comparisons might be desired from a clinical performance examination. One is the comparison of test takers to one another. A common use for this comparison is to determine the highest score so that an award may be given. This is called a norm-referenced framework. Another comparison is each test taker's performance with an expected performance. An example is where the expected performance is 70% of actions with a given patient. Every test taker's performance is compared with this number. All could accomplish more actions; all could accomplish less. This is called a criterion-referenced framework.

The index for reporting the reliability of an examination is different for these two frames of reference. The psychometric index for the consistency of a norm-referenced test is a generalizability coefficient. As mentioned earlier, whether actually measured or estimated, the reliability of a test will be higher when test taker scores are spread widely and where some test takers perform relatively well on most patient cases while others perform relatively poorly. For criterion-referenced comparisons, the consistency of classification is of interest. How consistently (reliably) are test takers classified as scoring above or below an expected level? This consistency may be indexed by a dependability coefficient, a phi coefficient or kappa. With the focus on classification, spreading test taker scores is no longer desired. An extreme example is where every test taker accomplishes 100% of expected actions. A good outcome occurs, there is no spread of scores, test reliability is zero and classification is perfect. In evaluating the measurement quality of a standardized clinical performance examination, first one must understand the intended frame of reference to determine whether the appropriate index of reliability or consistency is reported and then evaluate the size of the index.

How high does a generalizability or classification index have to be?

http://www.springer.com/978-1-4020-0466-7

International Handbook of Research in Medical Education

Norman, G.R.; van der Vleuten, C.P.M.; Newble, D.I. (Eds.)

2002, XIII, 1106 p. In 2 volumes, not available separately., Hardcover
ISBN: 978-1-4020-0466-7