Neurology objective structured clinical examination reliability using generalizability theory

Angela D. Blood, MPH, MBA
Yoon Soo Park, PhD
Rimas V. Lukas, MD
James R. Brorson, MD

Correspondence to Angela D. Blood: angela_blood@rush.edu

ABSTRACT

Objectives: This study examines factors affecting reliability, or consistency of assessment scores, from an objective structured clinical examination (OSCE) in neurology through generalizability theory (G theory).

Methods: Data include assessments from a multistation OSCE taken by 194 medical students at the completion of a neurology clerkship. Facets evaluated in this study include cases, domains, and items. Domains refer to areas of skill (or constructs) that the OSCE measures. G theory is used to estimate variance components associated with each facet, derive reliability, and project the number of cases required to obtain a reliable (consistent, precise) score.

Results: Reliability using G theory is moderate (Φ coefficient = 0.61, G coefficient = 0.64). Performance is similar across cases but differs by the particular domain, such that the majority of variance is attributed to the domain. Projections in reliability estimates reveal that students need to participate in 3 OSCE cases in order to increase reliability beyond the 0.70 threshold.

Conclusions: This novel use of G theory in evaluating an OSCE in neurology provides meaningful measurement characteristics of the assessment. Differing from prior work in other medical specialties, the cases students were randomly assigned did not influence their OSCE score; rather, scores varied in expected fashion by domain assessed. *Neurology*® 2015;85:1623–1629

GLOSSARY

D-study = decision study; **G-study** = generalizability study; **G theory** = generalizability theory; **OSCE** = objective structured clinical examination; **SP** = standardized patient.

Objective structured clinical examinations (OSCEs) are a method of performance-based assessment[1,2] in which learners rotate through multiple cases, typically comprising of standardized patients (SPs), or persons who are trained to portray a patient presentation in a consistent and believable manner,[3–5] in order to measure a variety of clinical skills.[6] Estimating reliability, or the consistency of the assessment, is challenging because OSCEs are inherently multifaceted with variance in scores influenced by many sources. Prior studies in neurology have described the utility and predictive value of OSCEs using traditional quantitative methods.[7,8] However, traditional methods may bias score interpretation,[9,10] and do not provide guidance as to how to improve reliability. This study demonstrates the use of generalizability theory (G theory),[11] which identifies factors influencing the variability of scores, calculates a refined measure of reliability, and makes post hoc projections of reliability when altering the structure of the OSCE.

BACKGROUND

OSCEs. OSCEs were first described in 1979 as a means of assessing clinical skill.[6] In an OSCE, students rotate through multiple stations that typically include SPs or partial task trainers, "models meant to represent only a part of the real thing and often comprise a limb or body part."[12] OSCEs are currently utilized to assess clinical skills during the medical licensing examination.[13]

While a student's clinical ability could be measured through direct observation or written tests, OSCEs are often preferred as they allow for application of skills and knowledge that can be difficult or impossible to measure otherwise.[14–19] Direct observation of clinical skills can be challenging because of variability of clinical settings,[20] lack of validity evidence to support assessment instruments,[19] subjectivity of faculty evaluations,[8] and

Supplemental data at Neurology.org

From Rush University Medical College (A.D.B.); Department of Medical Education (Y.S.P.), The University of Illinois at Chicago; and Department of Neurology (R.V.L., J.R.B.), The University of Chicago, IL.

Go to Neurology.org for full disclosures. Funding information and disclosures deemed relevant by the authors, if any, are provided at the end of the article.

reliance on secondary evidence such as oral case presentations or patient notes to make inferences about general clinical skills.[20] Written tests, while useful for some applications, may not sufficiently measure clinical skills because they do not allow the student to demonstrate their clinical skills in a holistic and authentic manner.

OSCEs require students to perform an activity or construct a response, thus allowing for evaluation of both the *process* used in solving a task as well as the *product*.[21] OSCEs also typically yield multiple scores, thus reducing time and cost for item development and test administration.[22] An additional benefit to an OSCE is that it provides a unique opportunity for feedback on the learner's performance, often from multiple sources, and thus facilitates deliberate practice, a cornerstone of modern simulation-based education.[23,24]

OSCEs are intended to measure a variety of clinical skills.[6] For example, if students are asked to produce a differential diagnosis, the OSCE allows students to demonstrate not only whether they can recite a differential diagnosis, but how well they can arrive at the diagnosis through an appropriate history and physical examination. The specific skills measured through an OSCE can vary based on discipline, level of learner, and objectives of the curriculum. Examples of skills that can be measured include history-taking, physical examination, procedural skills, communication and interpersonal skills, professionalism, and documentation.

In the University of Chicago Neurology Clerkship OSCE, each case includes the same format. Students read a door chart with introductory information about the patient and a list of tasks to complete. They have 30 minutes for the patient encounter and 15 minutes to complete a post-encounter patient note on a computer. While each of the cases is unique, the clinical competencies of history-taking, physical examination, communication, and data interpretation are consistent. The particular cases developed by the co-clerkship directors (R.V.L. and J.R.B.) were chosen because they included both positive and negative pertinent findings that could be simulated by trained persons (SPs), and because they represent different content areas covered during the clerkship. The "overall aim in case selection is to avoid infrequent or atypical scenarios in favor of classical or canonical entities that students need to recognize."[32] The case diagnoses are not discussed here because these materials are actively in use for summative assessment purposes. However, all case materials and assessments are available and password-protected through MedEdPORTAL.[32]

G theory. In measurement theory, a learner's score is a function of their "true" ability and measurement error. In G theory, the learner's true score is "carefully defined in relation to all identified facets of the measurement process, and thus provides a more comprehensive assessment of reliability."[25] In an OSCE, examples of facets include students (p), cases (c), domains (d), items (i), and raters (r), among others. Using G theory, the variance associated with each facet is estimated and provides meaningful information on the measurement characteristics of the assessment.

Students (p) refer to the variability in scores caused by true differences between learners. In an assessment that discriminates learner performance, one would expect that if student A outperforms student B on an assessment, this is a reflection that student A has greater ability. It is desirable for the student variance to be greater than any other facet because it indicates that differences in scores are truly attributable to student ability rather than measurement error. Cases (c) refer to the variability in difficulty associated with the patient encounters assigned in an OSCE. In this study, students were randomly assigned to 2 of 3 cases. Domains (d) refer to variability in difficulty associated with constructs or areas of skill that an assessment measures. In this study, the 3 dimensions within the OSCE include: the students' ability to gather a patient's history, the students' ability to perform physical examination maneuvers, and the students' ability to document an evidence-based differential diagnosis and treatment plan. Items (i) refer to the variability in difficulty associated with checklist items within each case and across domains. Finally, raters (r) refer to the variability in severity associated with evaluators, which in OSCEs often include SPs and faculty. Interactions between facets also provide meaningful information; for example, person by case ($p \times c$) interaction would indicate differences in student performance between cases, rather than consistency in performance across cases. The degree of variance contribution from each facet provides useful information to identify whether the assessment properly discriminates between high and low performers, whether cases or items sufficiently vary in difficulty, or whether sources of error caused by raters threaten the reliability of the assessment.

In addition to the variance sourced from each facet, G theory estimates the overall reliability of the OSCE. Reliability refers to the consistency of assessment scores, such that a student's performance is trustworthy from application to application. This consistency gives the educator confidence that the assessment scores can be used to make decisions. For example, scores could be used to determine whether or not a student passes the neurology clerkship and moves on to the next training experience. While reliability is important for any measurement, it is especially critical in high-stakes testing when significant consequences are attached.

Analyses using G theory are generally divided into 2 parts: (1) generalizability study (G-study), and (2) decision study (D-study). In a G-study, variance components from the facets noted above (e.g., students, domains, items) are estimated and an overall reliability is calculated. Then, using estimates of variance components, a post hoc projection of reliability is examined through the D-study. The G- and D-studies are not separate studies, but rather different analyses from the same overall study.

At times educators may be told that an absolute minimum number of cases for an OSCE exists, but in reality the number of cases required to meet acceptable reliability is context dependent. These contexts motivate the use of a D-study. For a given application, the minimum number of cases may be 3, while for another application it may be 10. Any adjustments made to the OSCE have implications for time, effort, and funding. It is therefore valuable to conduct a D-study to inform decisions before adjustments are made.

Case specificity. Case specificity can be defined as the phenomenon of student assessment performance varying from case to case, as some students may have greater knowledge or past experience with some cases rather than others.[26] Previous research regarding case specificity has yielded conflicting results,[27–30] with some studies showing degrees of case specificity in both performance-based and written examinations. Previous concepts of case specificity held that case specificity was naturally a large component of variance in multicase examinations, and thus many cases would be needed to arrive at a reliable score. More recent research has demonstrated that the number of cases, while a factor, is not necessarily the driving force behind variance, and instead larger sources of variance may be attributed to the items[29,31] or some other factor.

Determining the degree of case specificity within a neurology OSCE has both measurement and practical consequences. For example, if a significant portion of the variance was sourced from cases, this might indicate that the students' performance on the OSCE was influenced by which case he or she was assigned, rather than by his or her true ability. In an assessment with high case specificity, achieving sufficient levels of reliability may require adding more cases.

METHODS

Data. Data used for this study are based on a high-stakes summative neurology OSCE. This study was approved by the institutional review board. The facets included in this study are as follows:

- Students (p): One hundred ninety-four third-year medical students participating in the neurology clerkship rotation.
- Cases (c): Three cases, of which each student is randomly assigned to 2. Each patient case is portrayed in a standardized manner, and includes history and physical examination positive and negative pertinent findings.
- Domains (d): Three domains include gathering patient history (h), performing a complete neurologic physical examination (px), and an assessment of the patient including differential diagnosis documented in a note (a).
- Items (i): The domain of patient history includes a range of 11–14 case-specific items. The domain of physical examination contains 55 items consistent for all 3 cases; there are also up to 4 case-specific additional physical examination steps. The domain of patient assessment contains 20 case-specific items.

Analysis. Reliability and variance composition of the neurology OSCE were examined using G theory. Within a G-study design, all facets must be listed as either crossed or nested. If a facet is crossed, all elements of the facets interacted with each other. For example, in this study, domains (d) are crossed with cases (c) because all 3 of the cases tested all 3 of the domains (history-taking, physical examination, and differential diagnosis with treatment plan). Facets that are nested are those for which one facet was contained in another. For example, in this study, items (i) are nested within cases (c) because some items within each case were unique.

In this study, the overall design is $p \times [i\colon (d \times c)]$, where students are crossed with items nested within domains, which are crossed with cases. Cases and items are assumed to be random, sampled from the population (universe) of potential cases and items. In other words, in a universe of potential cases and items, this OSCE includes a sample. Domains are assumed to be fixed at 3, as the finite set of components measured. Specifically, there are a limited number of domains that can be measured within the universe of clinical skill, and this OSCE fully measures the 3 domains within its structure: gathering a patient's history, performing physical examination maneuvers, and documenting an evidence-based differential diagnosis and treatment plan.

Using estimates of variance components, G theory allows for the calculation of reliability: (1) G coefficient, and (2) Φ coefficient. These coefficients represent the reliability of the score provided by the OSCE. When normative decisions are needed, the G coefficient can be used; when criterion-based decisions are needed, the Φ coefficient can be used. D-study was conducted to examine the projected reliability when increasing the number of cases. D-study is a post hoc analysis that uses estimates of variance components from the G-study to make projections in reliability. Figure e-1 on the *Neurology*® Web site at Neurology.org shows the variance components associated with this design (including the linear model equation). Estimation was conducted using urGenova[9] for the G-study.

RESULTS

Descriptive statistics indicate a comparable distribution of performance across cases (see table 1). Student performance on a given domain is not significantly different from case to case. Instead, much of the variability in scores is sourced from the domain tested (history-taking, physical examination, and documentation), thus supporting the decision to consider the domains as separate constructs measuring different attributes.

Table e-1 represents the variance components from each facet within the OSCE. The analysis of variance components from the G-study reveals that a majority of variance is attributed to the domain

Table 1 Descriptive statistics

Checklist	No.	Mean	SD	Min	Max
Case 1					
SP history questions	121	0.75	0.14	0.36	1.00
Physical examination	121	0.91	0.07	0.67	1.00
Assessment: Patient note	121	0.68	0.14	0.25	0.90
Case 2					
SP history questions	109	0.74	0.15	0.36	1.00
Physical examination	109	0.94	0.05	0.80	1.00
Assessment: Patient note	109	0.76	0.10	0.50	1.00
Case 3					
SP history questions	159	0.79	0.11	0.42	1.00
Physical examination	159	0.91	0.06	0.69	1.00
Assessment: Patient note	158	0.66	0.12	0.35	0.95
Overall					
SP history questions	389	0.76	0.13	0.36	1.00
Physical examination	389	0.92	0.06	0.67	1.00
Assessment: Patient note	388	0.69	0.13	0.25	1.00

Abbreviation: SP = standardized patient.
Students are assigned 2 cases as part of their examination. There were 194 unique examinees who participated in this test. Student performance within domains across cases did not significantly vary, while student performance across domains within cases varied.

(11.7%). This further supports the findings in the descriptive statistics that each domain is truly unique from the others, and that successful performance on one domain does not necessarily translate into successful performance on another. A lesser amount of variance was sourced from cases crossed with items nested in domains (11.0%), students crossed with items nested in domains (8.6%), and items nested in domains (5.2%). Case specificity, or the degree to which the case the student is assigned influences the variance, was *not* a concern (0.0%). Finally, the 4-way interaction of all the facets (students, cases, domains, and items), as well as residual error, accounted for 59.2% of the variance components.

Results from a D-study are shown in table 2. D-studies provide projections in reliability estimates when facets are adjusted (e.g., increase the number of cases, decrease the number of items). This is a useful and practical tool to improve assessments based on evidence and assign resources for test development (e.g., faculty time needed to develop cases, feasibility in students' testing time).

For the overall reliability of the OSCE, in a G-theory framework, there are 2 reliability indices: Φ coefficient is used for criterion-based assessments (students measured against a predetermined standard), and the G coefficient is used for normative assessments (students compared to each other). The reliability estimates were as follows: Φ coefficient = 0.61 and G coefficient = 0.64. Students are randomly assigned to 2 of 3 cases, and thus calculations were made for each possible combination of cases (cases 1 and 2, cases 2 and 3, cases 1 and 3), as well as a "combined" calculation that represents an aggregate of estimates across all case combinations. The combined data reveal that 3 cases are required to reach a G coefficient above 0.70, and 6 cases are required to reach a G coefficient above 0.80. Educators would need to double the number of cases per student (3–6) in order to increase the reliability by a tenth of a point, but only one additional case is needed to reach the desired threshold. Figure 1 illustrates this for visual purposes. Because 0.70 is a commonly accepted minimum level of acceptable reliability for an internal examination,[33] the judgment about whether or not to double the number of cases becomes not simply one of measurement but also of institutional policy. The D-study results *dispel* the belief that there exists an absolute number of cases required for an OSCE in order to reach acceptable reliability and provide guidance for educators about how to determine the appropriate number of cases.

DISCUSSION A previous study examining the reliability of the neurology clerkship OSCE determined that student OSCE scores were significantly correlated from one case to another, significantly correlated with National Board of Medical Examiners subject

Table 2 Decision study results: Increasing the number of cases

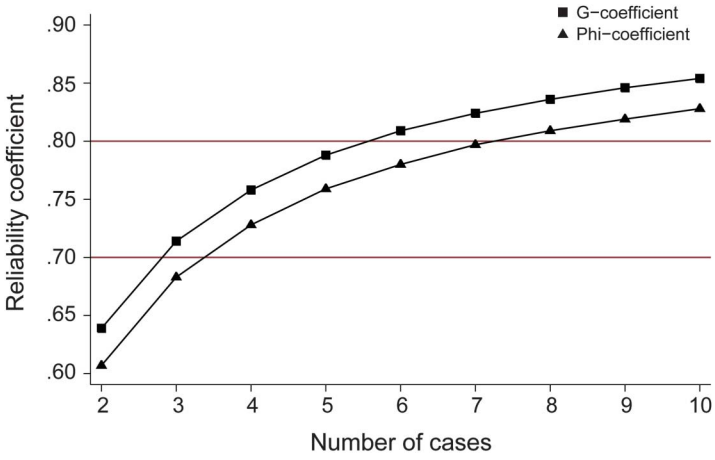
	Cases 1 and 2		Cases 2 and 3		Cases 1 and 3		Combined	
No. of cases	G	Φ	G	Φ	G	Φ	G	Φ
2	0.639	0.607	0.473	0.427	0.525	0.483	0.639	0.607
3	0.714	0.683	0.563	0.511	0.612	0.566	0.714	0.683
4	0.758	0.728	0.622	0.566	0.666	0.619	0.758	0.728
5	0.788	0.759	0.664	0.606	0.704	0.655	0.788	0.759
6	0.809	0.780	0.695	0.636	0.732	0.682	0.809	0.780
7	0.824	0.797	0.719	0.658	0.753	0.703	0.824	0.797
8	0.836	0.809	0.738	0.677	0.770	0.719	0.836	0.809
9	0.846	0.819	0.754	0.692	0.784	0.732	0.846	0.819
10	0.854	0.828	0.767	0.704	0.795	0.743	0.854	0.828

G and Φ coefficients calculated using variance components from table e-1, assuming domain to be fixed, while case and items are random. "Combined" represents an average across all case combinations. The combined data reveal that 3 cases are required to reach a coefficient above 0.70 and 6 cases are required to reach a G coefficient above 0.80.

examinations, and not correlated with faculty assessments of students' clinical performance.[8] These results are helpful in arguing to include an OSCE as a component of students' assessment portfolios within a neurology clerkship. However, in order to know which facets within the OSCE would be most influential if adjusted to increase the reliability of the assessment, new analyses using G theory are needed. The use of G theory builds on the prior study in that it provides data specific to each facet of the OSCE (students, cases, items, and domains) to guide decisions about adjusting facets to improve reliability.

This study offers an example of how to examine reliability using G theory when presented with an unbalanced dataset. Because each student is assigned to 2 of 3 possible cases, each student has missing data for the third case that he or she does not participate in. Many OSCEs have case banks or multiple cases that are exchanged from rotation to rotation, or year to year, in order to collect data on multiple case measurements and maintain test integrity. The value of breaking the data into 3 separate datasets (students who participated in cases 1 and 2, cases 2 and 3, cases 1 and 3) and then aggregating the estimates allows us to examine an OSCE with a rotating case bank.

Figure 1 Decision study results: Increasing the number of cases



Increasing the number of cases per student from 2 to 3 increases the G coefficient beyond 0.70, a typical benchmark for institutional examinations. Increasing the number of cases per student from 2 to 6 allows the G coefficient to reach above 0.80. As the number of cases increase, there is a diminishing return on reliability. G and Φ coefficients are from the averaged result across combination of cases.

According to a recent systematic review of the OSCE as a testing format, the average reliability of scores from high-stakes OSCEs was 0.65,[34] below the generally accepted guidelines of minimum reliability.[33] The level of reliability desired is a policy decision, and from a budgetary standpoint, having guidance regarding the number of cases required is more cost-saving than making an educated guess. This finding *contradicts* the argument that case specificity is a significant contributor to variance in all applications of OSCEs. Instead, educators should consider the context in which their OSCE is applied. It may be that summative OSCEs that sample curricular content across a medical program indeed have a greater degree of case specificity than OSCEs targeted within a single medical discipline such as a clerkship. This is an area in which future research should be focused.

Finally, within the field of neurology, reliability has been examined both in educational endeavors[35] and clinical applications.[36–38] While OSCEs are a common example of a multifaceted assessment, many if not all applications of assessment involve multiple facets. This demonstration of G theory presents educators and researchers with a method by which to

examine multiple facets, and therefore multiple sources of variances, within assessments.

Limitations. This study is limited in that it was conducted at one institution. It may be that applications of the same OSCE material would yield varying results depending on the student population to which it was applied. In addition, the case selection for clerkships in other settings may vary from the case selection analyzed in this study because curricula across institutions differ.

Conclusion. Simulation, specifically the use of SPs in OSCEs, is increasingly used in medicine as an assessment of individual ability. Simulation-based assessment is used to make decisions for advancement in institutional programs, for state licensure, and for some credentialing programs for current practitioners.[13,39] Despite significant consequences associated with simulation-based assessment, measurement principles are not consistently applied.

This study contributes to the body of knowledge in this field because it provides an example of applying measurement principle (G theory) to simulation-based assessment (OSCE). The results from this study reveal that student performance on a given domain does not significantly differ from case to case, but instead much of the variance in scores is sourced from the domain tested. It may provide an example for other educators utilizing OSCEs and call attention to issues that have yet to be fully explored, such as case specificity related to OSCE stations. In this study, case specificity did not contribute significantly to the variance in scores, dispelling the belief that all OSCEs must necessarily include large case banks to reach acceptable levels of reliability. This underscores the practical importance of conducting site-specific D-studies before adjusting facets within an assessment such as an OSCE.

Moreover, this research demonstrates the use of G theory to study a high-stakes local examination in medical education and in particular, neurology. Implications from the use of G theory that indicate differences in domain-level scores can be used to refine test development in future uses of this test design.

Comment: Generalizability theory and assessment in medical training

Even though it is more than 50 years old,[1] generalizability theory,[2] or G theory, is not well known to the medical community. It combines aspects of classic test theory and analysis of variance in order to estimate reliability ("G-study") of a measurement instrument as a function of design aspects ("facets" in G-theory literature), and to improve instrument design ("D-study"). G theory is unrelated to G estimation used in causal inference.

This study used G theory to analyze data from objective structured clinical examinations (OSCEs) in Neurology clerkships.[3] Specifically, the authors examined data from OSCEs of 194 students, each randomly assigned to 2 of 3 clinical cases, with 3 domains examined for each case: medical history, physical examination, and assessment. Each domain includes 11 to 55 specific assessment items. The authors' G-study shows that their OSCE has moderate reliability, and they identify 2 major sources of explainable variability: the 3 domains, and interaction between the 3 cases and the items nested within domains. The authors' D-study shows how reliability of their OSCE would be likely to improve if the number of cases was increased.

G theory has the potential to help create better instruments that can improve the evaluation process for Neurology clerkship and residency training. This article is a model for how this can be done: After a G-study analysis is used to describe an existing assessment procedure, D-study analyses show how modifications to the procedure can improve its reliability. Clerkship and residency directors should seriously consider this data-driven methodology to improve assessments.

1. Cronbach LJ, Rajaratnam N, Gleser GC. Theory of generalizability: a liberalization of reliability theory. Br J Stat Psychol 1963;16:137–163.
2. Brennan RL. Generalizability Theory. New York: Springer-Verlag; 2001.
3. Blood AD, Park YS, Lukas RV, Brorson JR. Neurology objective structured clinical examination reliability using generalizability theory. Neurology 2015;85:1623–1629.

Charles B. Hall, PhD

From the Department of Epidemiology and Population Health, and the Saul R. Korey Department of Neurology, Albert Einstein College of Medicine, Bronx, NY.

Study funding: No targeted funding reported.

Disclosure: C.B. Hall has received honoraria for serving as a reviewer for NIH study section review panels. Dr. Hall has received salary support from the National Institute of Occupational Safety and Health (U01 OH010412, and U01 OH010711, for which Dr. Hall is principal investigator [PI]; U01 OH010513 and U01 OH010728, PI Mayris P. Webber; U01 OH010711, PI Thomas K. Aldrich; and contracts 200-2011-39378 and 200-2011-39489, both with PI David J. Prezant), the National Institute on Aging (P01 AG003949, PI Richard B. Lipton; R01 AG034119, PI Chengjie Xiong; R01 AG022092, PI Leslie Wolfson), the National Cancer Institute (P30 CA013330, PI I. David Goldman), and the National Center for Advancing Translational Sciences (UL1 TR001073, PI Harry Shamoon). Go to Neurology.org for full disclosures.

AUTHOR CONTRIBUTIONS

Angela D. Blood and Yoon Soo Park contributed to the design and conceptualization of the study, analysis of data, and drafting the manuscript. Rimas V. Lukas and James R. Brorson contributed to the interpretation of data as well as revising the manuscript. Angela D. Blood and Yoon Soo Park performed the statistical analysis.

ACKNOWLEDGMENT

The authors thank Rush University Medical College, The University of Chicago Simulation Center, The University of Chicago Department of Neurology, The University of Chicago Pritzker School of Medicine, and The University of Illinois at Chicago Department of Medical Education for their support in conducting the Neurology Clerkship OSCE and for conducting this research.

STUDY FUNDING

No targeted funding reported.

DISCLOSURE

A.D. Blood and Y.S. Park report no disclosures relevant to the manuscript. R.V. Lukas serves as a contract worker to perform medical review of published content for EBSCO Publishing. He also acts as a consultant to the American Physician Institute, contributing to Board Review material for Continuing Medical Education. J.R. Brorson has received compensation as a consultant for CVS Caremark, Inc., for the National Peer Review Corporation, and for legal proceedings involving Octapharma, Inc. Go to Neurology.org for full disclosures.

Received February 5, 2015. Accepted in final form May 14, 2015.

REFERENCES

1. Fitzpatrick R, Morrison EJ. Performance and product evaluation. In Thorndike RL, editor. Educational Measurement, 2nd ed. Washington, DC: American Council on Education; 1971:237–270.
2. Wiggins G. A true test: toward more authentic and equitable assessment. Phi Delta Kappan 1989;70:703–713.

3. Barrows HS. An overview of the uses of standardized patients for teaching and evaluating clinical skills. Acad Med 1993;68:443–451.
4. Barrows HS, Abrahamson S. The programmed patient: a technique for appraising student performance in clinical neurology. J Med Educ 1964;39:802–805.
5. Levine HG, McGuire C. Role-playing as an evaluative technique. J Educ Meas 1968;5:1–8.
6. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979;13:39–54.
7. Gupta P, Dewan P, Singh T. Objective structured clinical examination (OSCE) revisited. Indian Pediatr 2010;47: 911–920.
8. Lukas RV, Adesoye T, Smith S, Blood A, Brorson JR. Student assessment by objective structured examination in a neurology clerkship. Neurology 2012;79:681–685.
9. Brennan RL. Generalizability Theory. New York: Springer; 2001.
10. Kreiter CD, Bergus GR. Case specificity: empirical phenomenon or measurement artifact? Teach Learn Med 2007;19:378–381.
11. Clauser BE, Harik P, Margolis MJ. A multivariate generalizability analysis of data from a performance assessment of physicians' clinical skills. J Educ Meas 2006;43:173–191.
12. Bradley P. The history of simulation in medical education and possible future directions. Med Educ 2006;40: 254–262.
13. National Board of Medical Examiners (NBME) and Federation of State Medical Boards (FSMB). United States Medical Licensing Examination (USMLE) Step 2 CS. 2014. Available at: http://www.usmle.org/step-2-cs/. Accessed December 19, 2014.
14. Schmahmann JD, Neal M, MacMore J. Evaluation of the assessment and grading of medical students on a neurology clerkship. Neurology 2008;70:706–712.
15. Heckmann JG, Knossalla F, Gollwitzer S, Lang C, Schwab S. OSCE in the neurology clerkship: experiences at the neurological department of the University Hospital Erlangen. Fortschr Neurol Psychiatr 2009;77:32–37.
16. Yoshii F. Advanced OSCE in neurology [in Japanese]. Rinsho Shinkeigaku 2007;47:897–899.
17. Prislin MD, Fitzpatrick CF, Lie D, Giglio M, Radecki S, Lewis E. Use of an objective structured clinical examination in evaluating student performance. Fam Med 1998; 30:338–344.
18. Gledhill RF, Capatos D. Factors affecting the reliability of an objective structured clinical examination (OSCE) testing neurology. S Afr Med J 1985;67:463–467.
19. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. JAMA 2009;302:1316–1326.
20. Williams RG, Klamen DL. Examining the diagnostic justification abilities of fourth-year medical students. Acad Med 2012;87:1008–1014.
21. Lane S, Stone CA. Performance assessment. In: Brennan RL, editor. Educational Measurement, 4th ed. Westport, CT: American Council on Education and Praeger Publishers; 2006:387–431.
22. Goldberg GL, Roswell BS. Are multiple measures meaningful? Appl Meas Educ 2001;14:125–150.
23. Ericsson KA. Deliberate practice and the acquisition and maintenance of expert performance in medicine and related domains. Acad Med 2004;79:S70–S81.
24. Ericsson KA, Krampe RT, Tesch-Romer C. The role of deliberate practice in the acquisition of expert performance. Psychol Rev 1993;100:363–406.
25. Kreiter CD. Generalizability theory. In Downing SM, Yudkowsky R, editors. Assessment in Health Professions Education. New York: Routledge; 2009:75–92.
26. Swanson DB, Norcini JJ. Factors influencing reproducibility of tests using standardized patients. Teach Learn Med 1989;1:158–166.
27. Colliver JA, Markwell SJ, Vu NV, Barrows HS. Case specificity of standardized-patient examinations: consistency of performance on components of clinical competence within and between cases. Eval Health Professions 1990;13:252–261.
28. Schuwirth LW, van der Vleuten CP. The use of clinical simulations in assessment. Med Educ 2003;37(suppl 1):65–71.
29. Norman G, Borage G, Page G, Keane D. How specific is case specificity? Med Educ 2006;40:618–623.
30. Wimmers PF, Fung CC. The impact of case specificity and generalizable skills on clinical performance: a correlated traits-correlated methods approach. Med Educ 2008; 42:580–588.
31. Gagnon R, Charlin B, Roy L, et al. The cognitive validity of the script concordance test: a processing time study. Teach Learn Med 2006;18:22–27.
32. Blood A, Brorson JR, Lukas RV. The neurology clerkship objective structured clinical examination (OSCE): a series of standardized patient cases. 2013. In: MedEdPORTAL. Available at: www.mededportal.org/publication/9368/. Accessed March 2013.
33. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ 2004;38:1006–1012.
34. Brannick MT, Erol-Korkmaz HT, Prewett M. A systematic review of the reliability of objective structured clinical examination scores. Med Educ 2011;45:1181–1189.
35. Schuh LA, London Z, Neel R, et al. Education research: bias and poor interrater reliability in evaluating the neurology clinical skills examination. Neurology 2009;73: 904–908.
36. Benbadis SR, LaFrance WC, Papandonatos GD, Korabathina K, Lin K, Kraemer HC. Interrater reliability of EEG-video monitoring. Neurology 2009;73: 843–846.
37. LeMarre AK, Rascovsky K, Bostrom A, et al. Interrater reliability of the new criteria for behavioral variant frontotemporal dementia. Neurology 2013;80:1973–1977.
38. Galvin JE, Roe CM, Xiong C, Morris JC. Validity and reliability of the AD8 informant interview in dementia. Neurology 2006;67:1942–1948.
39. American Society of Anesthesiologists (ASA). ASA workgroup on simulation education white paper: ASA approval of anesthesiology simulation programs. 2006. Available at: https://education.asahq.org/drupal/sites/default/files/asasimwhitepaper.pdf. Accessed September 22, 2015.