Reliability analysis of the objective structured clinical examination using generalizability theory

Juan Andrés Trejo-Mejía¹*, Melchor Sánchez-Mendiola¹, Ignacio Méndez-Ramírez² and Adrián Martínez-González¹

¹Secretariat of Medical Education, UNAM Faculty of Medicine, México City, México; ²Institute of Applied Mathematics and Systems, UNAM, México City, México

Background: The objective structured clinical examination (OSCE) is a widely used method for assessing clinical competence in health sciences education. Studies using this method have shown evidence of validity and reliability. There are no published studies of OSCE reliability measurement with generalizability theory (G-theory) in Latin America. The aims of this study were to assess the reliability of an OSCE in medical students using G-theory and explore its usefulness for quality improvement.

Methods: An observational cross-sectional study was conducted at National Autonomous University of Mexico (UNAM) Faculty of Medicine in Mexico City. A total of 278 fifth-year medical students were assessed with an 18-station OSCE in a summative end-of-career final examination. There were four exam versions. G-theory with a crossover random effects design was used to identify the main sources of variance. Examiners, standardized patients, and cases were considered as a single facet of analysis.

Results: The exam was applied to 278 medical students. The OSCE had a generalizability coefficient of 0.93. The major components of variance were stations, students, and residual error. The sites and the versions of the tests had minimum variance.

Conclusions: Our study achieved a G coefficient similar to that found in other reports, which is acceptable for summative tests. G-theory allows the estimation of the magnitude of multiple sources of error and helps decision makers to determine the number of stations, test versions, and examiners needed to obtain reliable measurements.

Keywords: OSCE; reliability; generalizability theory; clinical competence; México

Responsible Editor: Brian Mavis, Michigan State University, United States.

*Correspondence to: Juan Andrés Trejo-Mejía, Secretaría de Educación Médica, Facultad de Medicina UNAM, Edif. B, 3er Piso, Ave. Universidad 3000, C.U. México, D.F. 04510, México, Email: andretreme@gmail.com

Received: 17 March 2016; Revised: 25 July 2016; Accepted: 25 July 2016; Published: 18 August 2016
It can be calculated using multiple approaches, such as Cronbach’s alpha and generalizability theory (G-theory).

Cronbach’s alpha estimates only one facet as a source of measurement error, that is, the scores on the stations; this imposes limitations on the estimation of the reliability of the OSCE. A facet represents a dimension or source of variation over which generalization is desired (9, 10). Since classical test theory has some limitations, Cronbach developed generalizability theory, which enables the calculation of multiple sources of error with a single measurement (11). It was initially used in behavioral sciences, education, and psychology, and later in medical education for assessment methods with many sources of variation, like the OSCE (7, 11). In the OSCE, a measurement error can be attributed to the following: 1) stations, 2) standardized patients, 3) examiners, 4) scenarios (if the OSCE is conducted at multiple sites), and 5) occasion effects (if the OSCE is conducted at different times).

Generalizability theory can be applied in formative and summative examinations, and its use is recommended to investigate the sources of error and the number of observations required for a given level of reliability (12, 13). Analysis of the sources of error in summative exams is useful as a quality control procedure to ensure reliable inferences from the results. G-theory assumes that the observed score of a person (the object of measurement) consists of a universal score (analogous to the true score in the classical test theory) and one or more sources of variation or facets (14).

G-theory begins by defining a finite universe consisting of observations on all possible levels of the factors that are of interest to the researcher. For example, if one is interested in estimating the contribution of examiners, versions, and stations for the measurement of communication skills, our universe is defined in terms of the number of levels of examiners, versions, and stations. Thus, the universal score is the average score of individuals across all levels of all factors in this specific finite universe (15).

G-theory uses the techniques of analysis of variance (ANOVA) for measurement in the behavioral sciences. It allows the estimation of the importance of factors such as examiners, students, and stations on the reliability of the OSCE and how different numbers of examiners, students, and stations change reliability (15). Two types of coefficients are used in G-theory: the G coefficient, which indicates the reliability level, and the dependability index (D), which is generally smaller than G. The D coefficient is used to design D studies to calculate the standard absolute error and its confidence intervals (14, 15).

The strengths of G-theory, which complement Cronbach’s alpha, lie in its ability to identify which facets of the OSCE (stations, test versions, sites, or students) are the greatest source of measurement error. It also allows the decision maker to determine the number of examination occasions, the test formats, and the examiners needed to obtain reliable scores (14, 15). The scores obtained from an OSCE are vulnerable to potential errors because of seasonal variations, different versions, sites, and students. The monitoring these sources of variation functions as a quality control mechanism in order to ensure a valid interpretation of psychometric data and informed decision-making.

There are published reports of summative OSCE studies using G-theory, and the coefficients were between 0.51 and 0.78 (16–22). A large spread in G coefficients has been reported for formative OSCEs, ranging from 0.12 to 0.86 (7, 23, 24). All these studies were performed in developed countries, and there are no published replication studies from developing countries and Latin America. The goal of this study was to measure the reliability of a summative OSCE in a developing country medical school by using G-theory and to explore the usefulness of this approach for quality improvement.

Method

Subjects and setting

The available study population was 708 internship medical students from the National Autonomous University of Mexico (UNAM) Faculty of Medicine program. Most of the medical schools in Mexico have a 6-year program, of which the first 2 years include basic science courses, the next two and a half years comprise clinical courses, and the sixth year is an immersive full-time clinical experience in health care institutions called ‘medical internship’ (25). The internship rotation sites in our medical school are spread all over the country; therefore, we decided to use the available interns in the Mexico City area as sample for the study.

A 50% stratified random sample was selected to represent all internship hospitals and healthcare institution sites in Mexico City, which included 354 students. A total of 278 students (39.3% of the total population) took the summative end-of-career exam, which included the OSCE test used in our study. Their average age was 23.1 years: 72.7% were women and 27.3% were men. All students were familiar with the testing modality, as they had gone through a formative OSCE exam prior to the internship year.

Design

This is a cross-sectional study with a summative OSCE assessment in the end-of-career graduating exam. The exam had four versions, each with 18 equivalent stations from six areas (pediatrics, obstetrics and gynecology, surgery, internal medicine, emergency medicine, and family medicine), for a total of 72 stations. The equivalent stations were designed from the same blueprint and explored similar clinical skills, although they were different.
cases. Efforts were made to explore similar problems and topics in the different versions of the test. The exam was applied in six clinical sites simultaneously during a 2-day period. The students were randomly distributed among the six sites.

We considered four facets or sources of variation: the students, the equivalent stations, the test versions, and the sites. The stations, the patients, and the examiners were considered a single facet of analysis because there was only one examiner in each station, and the patients did not evaluate the performance of the students. It was not possible to separate the effects of the station, the examiner, and the patient.

A generalizability study was carried out using a random effects design, in which the items used to determine reliability were the total scores of each student at each station (5). The dependent variable was the score obtained by students in the OSCE; the independent variables were the students, the stations, the sites, and the test versions.

Case development
The components of clinical competence that were evaluated were 10 history-taking stations, three physical exam stations, one diagnosis and clinical management station, two radiographic interpretation stations, one laboratory studies interpretation, and one critical appraisal of a research paper station.

The assessment tools were developed and validated by a committee of experts on the six knowledge areas of the Medical Internship Program previously mentioned. The OSCE exam lasted for 2 h, with 6-min stations, and there were two rest stations. The raters evaluated the students with a station-specific checklist and marked the percentage of correct items for each station.

Raters’ training
A total of 108 clinical teachers from the UNAM Faculty of Medicine participated in the examination process, 18 in each of the six clinical sites, that is, one examiner for each station. The students and the examiners did not know each other. All examiners were clinical faculty, with training and experience in the OSCE methodology.

The examiners reviewed the stations, the checklists, and the global rating scale of communication skills before the test. In 13 stations with patients, the examiner, apart from evaluating the checklist, assigned a score to the students’ communication skills using a global rating scale of 1 to 9 (1 to 3 = unsatisfactory, 4 to 6 = satisfactory, and 7 to 9 = superior performance).

Each examiner used an op-scan sheet to score each student with the checklist and the global rating scale. All examiners attended the six sessions (three per day) and scored the four versions of the test.

Standardized patients
A total of 124 patients participated in the examination; they were trained with a workshop and were evaluated with a formative OSCE 8 months before the exam. All the patients had participated in a minimum of five previous OSCE tests and were evaluated by the clinical raters. All standardized patients with an acceptable performance were included in the examination. Two standardized patients were excluded due to poor performance in the previous exams.

Statistical analysis
Reliability tests with G-theory were carried out using the results of the ANOVA from the ‘model fit’ routine in JMP (SAS Software). The object of measurement, also called the facet of differentiation, was labeled ‘p’, for ‘person’ (Table 1). A random effects model was used to identify the main sources of variation. (Fig. 1) With this design, we obtained the estimates of the components of variance for each of the following facets: stations, sites, test versions, and all their interactions. We included the interaction ‘station × students’ nested within test versions and sites. Furthermore, we estimated the residual component of variance, which is the variability not explained by any of the facets.

In our OSCE study, the ‘p:sev × s’ design means that students (p) are nested within the site (s) and the test version (v) and that each student is crossed with each station (s). The test versions are crossed with the sites. The stations, in turn, are crossed with the sites and with the test versions (14, 15). We used JMP software (version 11) for the analysis of linear statistical models with both fixed and random effects. The formulas used for computing coefficients G and D are given in the Appendix.

Ethical aspects and funding
The protocol for this study was approved by the Research and Ethics Committee of UNAM Faculty of Medicine and received support from the Program for Support of Projects for Innovation and Improvement of Teaching (PAPIEME), UNAM, with project number PE207410. Participation in the study was voluntary. We preserved the confidentiality and anonymity of the students and present results in aggregate form.

Table 1. Notation system with G-theory

| p:sev × s |
|-----------|
| p = person (student or object of measurement) |
| s = site |
| v = version |
| s = station |
| . means ‘nested within’ |
| × means ‘crossed with’ |
Results

The OSCE overall mean score was 63.2 ± 5.7 (mean ± SD). The distribution of the students among the exam versions was 97 in version 1, 53 in version 2, 98 in version 3, and 30 in version 4.

The generalizability study revealed that the greatest source of variance (Table 2) was the residual error (51%). The stations were another source of variance (11.4%). No significant source of variance was found in the sites or in the test versions (0 and 0.2%, respectively).

To determine the reliability of the test scores, we used the components of variance to calculate the G and D coefficients.

This design means that students (P) are nested within the site (S) and the test version (V) and that each student is crossed with each station (E). The test versions are crossed with the sites. The stations, in turn, are crossed with the sites and with the test versions.

The G (generalizability) and D (Dependability) coefficients obtained from the analysis of the OSCE are shown in Table 3, with standard errors of measurement (SEM).

Discussion

This paper presents the results of an OSCE reliability study using G-theory. The OSCE can be improved by using G-theory because it provides information about the main indices of quality and validity evidence in the results of an assessment. The total variance of the OSCE was low; the reason might be that at the end of the internship year, and due to the similar educational process to which the students are subjected, they become more homogeneous (8, 14).

Regarding the estimates of the variance components in our OSCE, the students’ variance was 4.8%, which shows how much they differed in their performance.

Table 2. Variance component estimates obtained by analyzing the OSCE with G-theory (n = 278)

| Components | Variance component estimates | %  |
|------------|------------------------------|----|
| Station    | 42.157                       | 14.3|
| Version    | 0.737                        | 0.2 |
| Version x station | 38.968            | 10.6|
| Site       | 0                            | 0.0 |
| Site x station | 46.631              | 12.7|
| Site x version | 0.867                | 0.2 |
| Site x station x version | 34.692            | 9.4 |
| Student [version, site] | 17.652            | 4.8 |
| Station x student [version, site] | 187.374       | 51  |
| Total      | 367.024                      | 100 |

Estimates based on equating mean squares to expected value. $[] =$ Nested, $\times =$ interaction. $^a$Residual error.

Fig. 1. Model of statistical analysis to obtain variance component estimates to calculate the G coefficient.

The stations’ component of variance was 11.4%, which reflects the variance of the constant errors associated with levels of difficulty in the universe of the stations; the relative position of the students differed from one station to another.

The variation in the scores of the stations could be due to the OSCE content: history, physical examination, diagnosis and management, interpretation of radiographic and laboratory studies, and critical appraisal of a research paper (26).

The variance in the effect of the interaction between students and stations indicates that there are differences in the management of cases by the students. For example, a student might find it easy to manage some stations and may have difficulty to manage others. These results indicate that the difference in difficulty between stations differs from one student to another.

The interaction between versions and stations (10.6%) indicates that there are differences in the scores of stations according to the version. The interaction between the sites, versions, and stations (9.4%) is explained by the differences in the scores of stations according to the sites and versions.

Regarding the calculation of the G and D coefficients, an approach based on G-theory allows examining the implications of increasing or decreasing the size of the number of stations to assess its effect on the G coefficient. The D study, which is based on the findings of the G study, predicts what would happen if the number of stations is increased (15). In our study, an increase to 22 stations, with a duration of 2 h 24 min, would increase the G coefficient from 0.93 to 0.94. Versions and sites have a component of variance close to 0%, therefore, the increase in reliability would be marginal if they get changed.

Table 3. G and D coefficients

|                | OSCE          |
|----------------|---------------|
| G coefficient  | 0.93          |
| SEM relative   | 1.16          |
| D coefficient  | 0.83          |
| SEM absolute   | 3.42          |

SEM = standard errors of measurement, OSCE = objective structured clinical examination.
All variances are considered to calculate the D coefficient (dependability), except those involving the student; the D coefficient was 0.83 (9, 15).

During the internship year, the students developed clinical competence; the lower variability of their components was reflected in the low values of the absolute and relative SEM of the OSCE (Table 3).

The G coefficient measures the proportion of the total variation produced by the variation in knowledge and skills of the students (18). A higher value of G implies that the other sources of variation are less important compared to the variation among students. Furthermore, the G coefficient value is considered to represent an acceptable reliability for a summative examination lasting 2 h (19).

A meta-analysis found that the unweighted average of the generalizability coefficient was 0.49, as would be expected (27).

Lawson, Auewarakul, and Hatala conducted OSCEs and obtained a G coefficient from 0.62 to 0.68 (17, 19, 20). These values were lower than ours, which could strengthen our validity arguments.

Baig, Vallevand, Boulet, and Donnon also published OSCE studies and obtained a G coefficient from 0.51 to 0.78 (16, 18, 21, 22). These studies lasted longer but their G coefficient was equal to or lower than ours.

This study reports the experience of only one medical school in Mexico. We had only one examiner per station, which did not allow us to determine interexaminer reliability. It is recognized that G-theory is complex and requires expertise in its use (12).

There are several software programs that can be used to carry out generalizability studies, such as urGENOVA (15), SAS, G_String (13), and SPSS (14). G-theory involves a linear random effects model that is processed by JMP routines, which has the necessary tools to calculate estimates of the components of variance of studies with crossover and nested designs. In JMP, all you need is to put the scores (dependent variables) and all the individual facets, with interactions and nestings (independent variables), in the appropriate boxes, mark all of them as random, and run the program.

The reliability study using G-theory allowed us to identify several sources of variation that are involved in the OSCE, including the students' variance and the residual used to calculate the G coefficient; it also allowed us to predict that increasing the number of stations would increase the reliability of the OSCE. The analysis of the stations will allow us to improve the quality of the OSCE.

The use of G-theory can solve some problems inherent to interexaminer reliability, such as overestimation of reliability, as pointed out by Fan (28). A large number of raters participated in our OSCE, and we made the reasonable assumption that any error due to differences between examiners was randomly distributed. We assume that the error of variance due to differences between examiners is small because they were all trained clinicians with standard performance in the OSCE.

Student characteristics such as gender and skill level were not considered in this study, even though they are potential sources of error; we recommend considering these variables in subsequent studies.

This is the first study to explore the use of OSCE in Latin America, using G-theory as a mathematical model for evaluating its reliability. One of the most important issues in modern research is the reproducibility of published studies, which has been debated exhaustively in many forums (29), and the situation is not improving due to a multitude of causes (e.g., publication bias, funding and emphasis on absolute originality, to name a few). The reproducibility problem has been studied in education research (30), where only 0.13% of education studies are replications. The reproducibility situation in medical education research has not been studied, but there is no reason to suspect it is any different from clinical, basic science, or general education research. The publication of results from different contexts, such as Latin American medical schools, can be relevant to the international medical community because it provides reproducibility evidence of the implementation and mathematical analysis of logistically complex and expensive assessment methods like the OSCE.

Conclusions
The use of OSCE method in evaluating clinical competence has shown its usefulness. This assessment methodology has adequate reliability in our settings, and it could be of great importance to students, teachers, and medical schools using formative and summative OSCEs.

Equivalent versions of the examination, an appropriate planning, and rigorous implementation are factors that produced acceptable results. Our OSCE had good reliability as measured with G-theory.

Ensuring high quality of clinical competence assessment of students is the responsibility of all the relevant stakeholders, including the clinical teachers. The medical school authorities have an important responsibility in supporting and promoting the use of valid assessment methodologies.

Authors’ contributions
JAT, MS, and AM designed and conducted the study. JAT and IM collected the data and conducted the statistical analysis. All the authors contributed to the writing of the paper and have read and approved the final version.

Conflict of interest and funding
The authors have not received any funding or benefits from industry or elsewhere to conduct this study.
References

1. Van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. Teach Learn Med 1990; 2: 58–76.
2. Harden RM, Stevenson WM, Downie W, Wilson GM. Assessment of clinical competence using an objective structured clinical examination (OSCE). Br Med J 1975; 1: 447–51.
3. Hubbard JP. Measuring medical education. Philadelphia: Lea & Febiger; 1971.
4. Elstein AS, Shulman LS, Sprafka SA. Medical problem solving. Cambridge, MA: Harvard University Press; 1978, pp. 292–4.
5. Downing S. Reliability: on the reproducibility of assessment data. Med Educ 2004; 38: 1006–12.
6. Petrusa ER. Clinical performance assessments. In: Norman GR, van der Vleuten CPM, Newble DI, eds. International handbook for research in medical education. Dordrecht, the Netherlands: Kluwer Academic Publishers; 2002, pp. 673–709.
7. Newble DI, Swanson DB. Psychometric characteristics of the objective structured clinical examination. Med Educ 1988; 22: 325–34.
8. Downing SM. Validity: on the meaningful interpretation of assessment data. Med Educ 2003; 37: 830–7.
9. Brennan RL. Generalizability theory. New York: Springer-Verlag; 2010.
10. Van der Vleuten C. Validity of final examinations in undergraduate medical training. Br Med J 2000; 321: 1217–19.
11. Cronbach L, Gleser G, Nanda H, Rajaratnam N. The dependability of behavioral measurements: theory of generalizability for scores and profiles. New York: Wiley; 1972.
12. Streiner DL, Norman G. Reliability. In: Streiner DL, Norman G, eds. Health measurement scales: a practical guide to their development and use. Oxford: Oxford University Press; 2003, pp. 126–52.
13. Cronbach L, Shavelson RJ. My current thoughts on coefficient alpha and successor procedures. Educ Psychol Meas 2004; 64: 391–18.
14. Bloch R, Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Med Teach 2012; 34: 960–92.
15. Shavelson J, Webb N. MMSS generalizability theory. A primer. Thousand Oaks, CA: Sage; 1991.
16. Baig L, Violato C. Temporal stability of objective structured clinical exams: a longitudinal study employing item response theory. BMC Med Educ 2012; 12: 121. [cited 29 July 2016] Available from: http://www.biomedcentral.com/1472-6920/12/121
17. Lawson DM. Applying generalizability theory to high-stakes objective structured clinical examinations in a naturalistic environment. J Manipulative Physiol Ther 2006; 29: 463–7.
18. Valleyand A, Violato C. A predictive and construct validity study of a high-stakes objective clinical examination for assessing the clinical competence of international medical graduates. Teach Learn Med 2012; 24: 168–76.
19. Auewarakul C, Downing SM, Jaturatamrong U, Praditsuwan R. Sources of validity evidence for an internal medicine student evaluation system: an evaluative study of assessment method. Med Educ 2005; 39: 276–83.
20. Hatala R, Marr S, Cunic C, Bacchus CM. Modification of an OSCE format to enhance patient continuity in a high-stakes assessment of clinical performance. Med Educ 2011; 45: 517–27.
21. Boulet J, McKinley D, Whelan G, Hambleton R. Quality assurance methods for performance-based assessments. Adv Health Sci Educ 2003; 8: 47.
22. Donnon T, Paolucci E. A generalizability study of the medical judgment vignettes interview to assess students’ noncognitive attributes for medical school. BMC Med Educ 2008; 8: 58. doi: http://dx.doi.org/10.1186/1472-6920-8-58
23. Hull AL, Hodder S, Berger B, Ginsberg D, Lindheim N, Quan J, et al. Validity of three clinical performance assessments of internal medicine clerks. Acad Med 1993; 70: 517–22.
24. Wilkinson TJ, Newble DJ, Wilson PD, Cater JM, Helms RM. Development of a three-centre simultaneous objective structured clinical examination. Med Educ 2000; 34: 798–807.
25. Sánchez M, Durante I, Morales S, Lozano R, Martínez A, Graue E. Plan de Estudios 2010 de la Facultad de Medicina de la Universidad Nacional Autónoma de México. Gac Med Mex 2011; 147: 152–8.
26. Auewarakul C, Downing SM, Jaturatamrong U, Praditsuwan R. Sources of validity evidence for an internal medicine student evaluation system: an evaluative study of assessment method. Teach Learn Med 2012; 24: 168–76.
27. Brannick MT. A systematic review of the reliability of objective structured clinical examination scores. Med Educ 2011; 45: 1181–9.
28. Fan X, Chen M. Published studies of inter-rater reliability often overestimate reliability: computing the correct coefficient. Educ Psychol Meas 2000; 60: 532–42.
29. Open Science Collaboration. Estimating the reproducibility of psychological science. Science 2015; 349: aac4716. doi: http://dx.doi.org/10.1126/science.aac4716
30. Makel MC. Facts are more important than novelty. Replication in the education sciences. Educ Res 2014; 43: 304–16. doi: http://dx.doi.org/10.3102/0013189X14545513
Appendix A

Formulas for Computing Coefficient G and Coefficient D

FORMULA FOR COEFFICIENT G

\[
g = \frac{S_p^2}{S_p^2 - \frac{s^2_{e\text{-inter}}}{n_s \times n_e} + \frac{s^2_{v\text{-inter}}}{n_v \times n_e}}
\]

where:

- \( S_p^2 \) = student variance
- \( n_s \) = number of sites
- \( n_v \) = number of versions
- \( n_e \) = number of stations
- \( n_s \times n_v \times n_e \) = number of stations \( \times \) number of versions \( \times \) number of sites

FORMULA FOR COEFFICIENT D

\[
\Phi = \frac{S_p^2}{S_p^2 + \frac{s^2_{e\text{-inter}}}{n_s} + \frac{s^2_{v\text{-inter}}}{n_v} + \frac{s^2_{e\text{-within}}}{n_e} + \frac{s^2_{v\text{-within}}}{n_v} + \frac{s^2_{pe\text{-inter}}}{n_s} + \frac{s^2_{pv\text{-inter}}}{n_v}}
\]