**Recent Advances in Generalizability Theory to Improve Assessments in Medical Education and Practice**

**Authors and Author affiliations**

Zhehan Jiang 1,2 Jihong Zhang 3\* Lingling Xu 1,2 Yuting Han 1,2 Fen Cai1,2 Haiying Liang 1,2  Jinying Ouyang 1,2\*

1 Institute of Medical Education, Health Science Center, Peking University, Beijing, Peking, China

2 National Center for Health Professions Education Development, Peking University, Beijing, Peking, China

3 College of Education and Health Professions, University of Arkansas, Fayetteville, AR, USA

**Corresponding author's information**

Jihong Zhang

Postal address: West Maple Street 751, Fayetteville, Ar 72701, USA

E-mail address: z.john92@gmail.com

Jinying Ouyang

Postal address: Peking University Health Science Center, 38 Xueyuan Rd, Haidian District, Beijing, China

E-mail address: ouyangjinying@bjmu.edu.cn

**Recent Advances in Generalizability Theory to Improve Assessments in Medical Education and Practice**

**Abstract**

Performance assessment in medical education is a way of evaluating medical students' clinical competence; among studies related to performance assessment, generalizability theory (G-theory) is a prevalent approach for quantifying and investigating the reliability of the assessment. It looks at how consistent results are across different conditions, such as different raters or tasks, to navigate the investigation on the reliability of assessment measures. However, the progress of G-theory's application to the assessment in medical education remains bleak. This paper introduces many less-known, yet advanced and valuable, G-theory properties including consistency indexes for absolute decisions, cost-effectiveness optimization, multivariate extensions of current results, confidence intervals of relevant coefficients, estimation methods for different inquiries, open-source software options, and others.

**Key messages**

* Additional G-theory results are recommended to report whenever needed: (1) conditional standard errors of measurement, (2) cut-score specific coefficients, and (3) confidence intervals (CIs) of the coefficients.
* Advocation for Multivariate G-theory: (1) tuning weights of subdomains; (2) examining the quality of subscores; (3) offering solutions to complex situations.
* Using G-theory with monetary constraints for cost-effectiveness optimization is practically valuable in planning medical education assessment.
* Implementing G-theory in a structural equation modelling framework can incorporate properties of multiple regression, path analysis, and confirmatory factor analysis that involve both observed and latent variables, yield measures of global fit, tests for the model of interest, and better handling of missing data.

**Keywords**: Generalizability theory; performance assessment; reliability; clinical competence

**Introduction**

Performance assessment in medical education, of which the designs are often complex, plays an essential role in evaluating clinical competence.1 Solid performance assessments are expected to include (1) a well-defined statement of purpose, (2) a detailed description of target constructs (e.g., abilities and/or skills), (3) programmatic instructions for feasible administration and scoring, and (4) scientific procedures for data analyses and interpretation2; we address the last point in this paper—generalizability theory, or namely G-theory3, which is used to estimate “reliability” to show how consistently a method measures targets. Reliability per se is not observable; it demands modeling and estimation. High reliability indicates that when one applies the same method to the same sample under the same conditions, he/she should find similar (even identical) results. If not, the method of measurement may be unreliable or bias may have crept into the research. As one shall find in this paper, G-theory provides beyond conventional indices for reliability estimation and its advancement (e.g., the multivariate version) allows researchers to discover complex hidden information. To illustrate, using multivariate G-theory can yield better estimates of score consistency for composite scores and correctly estimate correlation coefficients among subdomains for measurement errors when corresponding G-theory design demands assumptions for the existence of multiple subdomains. Simply counting the number of publications (from PubMed, Embase, Ovid MEDLINE, Cochrane Library, PLOS Medicine, BioMed Central, OpenGrey, Google Scholar, Directory of Open Access Journals, Scopus, and Web of Science) related to the applications G-theory to performance assessments in medical education, one would realize the importance of this topic. To name a few, G-theory was used to investigate the Comprehensive Osteopathic Medical Licensure Examination-USA (COMLEX-USA)4, the Korean Medical Licensing Examination5, and the United States Medical Licensing Examination (USMLE).6-8 It’s also not difficult to picture that studies of this kind took place in low-stakes settings have been conducted substantially; A typical instance is a mini-clinical evaluation exercise (mini-CEX) for residency training, where G-theory can be used, for example, to reflect the variability in ratings that would occur if each resident were graded by a large number of evaluators while seeing a large number of patients (Norcini et al., 2003).

**Generalizability Theory**

G-theory is a prevalent choice for investigating performance assessments. Statistically, a performance assessment is regarded as student performance sampled from a complex universe consisting of all possible *facets* (e.g., tasks, occasions, raters, and measurement methods), which may lead to non-negligible deviation from the individual ability targeted in medical examination. G-theory allows researchers to present evidence bearing on the generalizability of performance assessments, considering the sampling properties within the complex setting.9 Methodologically, G-theory provides a framework for estimating, determining, and designing the generalizability of various observations or ratings10; these features fit the universe assumption for performance assessments well and, therefore, have been widely utilized in medical education and related studies.

Essential G-theory consists of G- and D- studies. One can use the former, a G-study, to investigate the composition of assessment scores; that said, the variance of each facet contributes to the observed score can be disentangled. On the other hand, a D-study (i.e., decision study), containing information for decision makings, is utilized to improve (sometimes optimize) the target performance assessment, such that its generalizability can reach a desirable level; likewise, the study can also inform of researchers quantified results, if shrinking the original setting from the current design is considered (e.g., reducing the number of raters and/or stations). G-theory has univariate and multivariate versions, where the latter permits subdomains, of which the associations in between are provided. The discrepancy of its univariate/multivariate assumption, however, does not alter the properties of G- and D- studies. Most applied works using G-theory cover the aforementioned aspects; in fact, recent literature reviews and primers on G-theory for medical educators and researchers almost all lead to this conclusion; representative examples include a recent meta-analysis by Andersen et al11 and a tutorial by Monteiro et al12.

Although it’s evident that, from both theoretical and empirical perspectives, the application of G-theory has been trending in performance assessment research, substantial studies of this kind fall into the “trap of clichés,” meaning that many useful and/or new properties of G-theory are in fact neither recognized nor adopted by applied researchers who, perhaps, tend to follow typical research routines of the articles that are highly cited. Surprisingly, as detailed in the following parts of this paper, many researchers have contributed to the methodological development in recent years.

To inform researchers what could have achieved more with G-theory, this paper summarizes underrated utilities and recent developments in the methodology.

**Advancement of G-theory Application**

Generally, reporting G-theory results often involve (1) variance and/or covariance estimates as well as their proportions for each facet present in target performance assessments (2) estimates of (relative and/or absolute) standard errors of measurement, and (3) two consistency coefficients (i.e., generalizability and dependability coefficients).13 In addition to these basic parts, depending on the assessment purpose, the results can be enriched by extra information recommended in this paper: (1) conditional standard errors of measurement,14 (2) cut-score specific coefficients,15 and (3) confidence intervals (CIs) of the coefficients.16 Above all, conditional standard errors of measurement characterize measurement error spans for any given sum (observed) score, making the analytical perspective different from traditional standard errors of measurement, which assume the measurement errors are identical across the span of sum (observed) scores. The Standards for Educational and Psychological Testing17 recommends that conditional standard errors of measurement be reported, as they provide reliability information to a specific level of sum (observed) scores; this is particularly useful in judging candidates of cut-scores because, intuitively, we always prefer setting the cut-score to a point/value containing less measurement errors. The second piece of the recommended information, cut-score specific coefficients, provides customized dependability coefficients for a specific sum (observed) score, such that absolute (criterion- or domain-referenced) decisions can be better made. Finally, since CIs measure the degree of uncertainty in a sampling method, it’s natural to consider reporting CIs for the consistency coefficients. It was difficult to derive a versatile formula for computing CIs for both variance/covariance and coefficient estimates, but the use of bootstrapping circumvents the difficulty. Tong and Brennan18 recommended bootstrapping procedures to handle the inquiry, and Jiang et al16 further proved that a bootstrapping genre called “parametric methods with spherical random effects” works the best for yielding CIs for the coefficients.

Multivariate G-theory has been utilized in relevant studies, for example, Clauser et al19 investigated four components of the USMLE Step 2 Clinical Skills examination (i.e., communication and interpersonal skills, spoken English proficiency, data gathering, and documentation) via multivariate G-theory, and Margolis et al20 used the framework to study the mini-Clinical Skills Examination (CEX) rating’s seven competencies: medical interviewing, physical examination, humanistic qualities, clinical judgment, organization/efficiency, counseling, and overall clinical competence. Although the volume of publications related to multivariate G-theory is sufficient for meta-analyses, it doesn’t alter the fact that, compared with univariate G-theory or classical test theory, the applications of multivariate G-theory in medicine educational research are relatively small. An important reason is that many primers and tutorials only mention multivariate G-theory at the end as a further extension. Also, samples required for multivariate G-theory would be even larger, making many applied studies not qualified for using the framework. In addition, statistical understanding and software programs for multivariate G-theory set a high threshold for applied researchers. But one should know that, in many realistic situations, multivariate G-theory is much more appropriate, as it doesn’t restrict the ability/skill (i.e., domain) to be conceptually single/univariate. Typical scenarios include an OSCE and an objective structured assessment of technical skill (OSATS) where subdomains are of concern, both theoretically and practically. A highly well-known property, if not the most, of multivariate G-theory is providing correlation estimates between pairs of subdomains for a given facet21. However, simply obtaining the correlation estimates does not always intuitive; the values do not possess straightforward interpretations per se. In addition, the multivariate estimates of the target domains (e.g., “History Taking”, “Physical Examination”, and “Documentation” in the USMLE Step 2 Clinical Skills Examination) are always combined to form a single score for further usages (e.g., score reporting or decision-making), while simply summing the estimates is frequently adopted. Based on the two aforementioned issues, the advocating in this section entails (1) tuning weights of subdomains (i.e., effective weights) to form meaningful composite scores and global generalizability coefficients22; (2) examining the quality of subscores (i.e., score of subscales) when subdomain performance outcomes are reported to both individuals and populations23, 24; (3) offering solutions to complex situations in which the design of a test and the resulting data structure are not definable by a single design (e.g., mixed-format assessments composed of multiple-choice and free-response items and special assessments containing both testlets and stand-alone sets of items).25 The 1st point about weighting is deploying statistical processes to assign proportions to different subdomains (i.e., the importance is unequal across the subdomains) in order to meet certain criteria, such as yielding the highest consistency index for the composite scores. The 2nd point relates to the correlation estimates between pairs of subdomains. These correlation estimates are taken to form subscore-related indexes, so that researchers can use them to examine the quality of subscores. It’s intuitive that subscores can provide potential diagnostic value for examinees interested in knowing their strengths and weaknesses in specific content areas, so that one can plan for future remedial work (Haladyna & Kramer, 2004). However, reporting subscores is not always meaningful and reliable. The Standards for Educational and Psychological Testing 17 emphasizes that scores, including subscores in the present context, should not be reported unless the validity, comparability, and reliability of such scores have been established, and the standard applies to subscores as well: If a test provides more than one score, the distinctiveness of the separate scores should be demonstrated. Derived from multivariate G-theory, two indexes-**Ɠ** and **RPV**-are used to demonstrate the quality of subscores for population and individuals, respectively. That said, if the indexes aren't sufficiently high (e.g.,0.8), it will be likely to hint that the subscore result is misleading, and therefore should not be reported. The 3rd point is about using multivariate G-theory to accommodate assessment designs: separately modeling multiple choice and free-response data, for example, ignores dependence between the two analyses; forwarding G-theory to these "hyper" settings not only makes the analysis aligned better with the design but also reduce biases and errors possibly caused by the independence practice (i.e., separately modeling).

Commonly seen in published papers using G-theory, D- studies are utilized to determine the number of facet(s) needed, in order to achieve a satisfactory level of consistencycoefficients (e.g., how to increase the dependability coefficient from 0.6 to 0.8). Researchers vary the numbers of levels of facet(s) and, therefore, build combinations from these numbers to form possible “solutions”. Each solution can yield a set of coefficients for decision-making. However, these solutions tend to be subjective, sometimes even highly arbitrary. As a result, finding an optimized one from the solutions relies on human efforts and, more importantly, the task becomes more difficult when constraints are present. Monetary constraints are highly realistic in medical education assessment. As evidenced by Brown et al26 and Walsh and Jaye27, standardized patients, raters, and stations are costly in OSCEs. Therefore, varying their numbers should be constrained by financial budgets. Marcoulides28, Marcoulides and Goldstein29, 30 as well as Meyer et al31 derived a series of mathematical formulas to find the optimal solution that can maximize generalizability within a budget constraint. Given these formulas tend to work on a particular design, its application is relatively limited if an applied researcher’s design is off the list. Jiang et al32 proposed using machine-learning approaches to handle the conditional optimization inquiry. Without working on mathematical problems, the machine-learning approaches are driven by modern algorithms that research only needs to define target (cost) function and constraint functions. In the present context, a target (cost) function can be a D-study’s formula for estimating the generalizability of a particular solution, where the constraint function ensures that the solution’s costs (i.e., the sum of unit costs multiplying by the solution’s numbers) remain affordable. Essentially, these cutting-edge approaches can drive the assessments to be more cost-effective.

Raykov and Marcoulides33 derived structural equation modeling (SEM) approaches to estimate G-theory indexes in settings where sampling of subjects and conditions in one- and two- facet crossed designs of univariate G-theory, where Jiang et al34 and Jorgensen35 further extended the SEM approach to include auxiliary variables in the models and estimate absolute-error components, respectively. Vispoel et al36 demonstrated that multivariate G-theory can be fitted in SEM as well when designs are *p* •×*i* (persons and items are fully crossed for all subdomains), *p* •×*i* •×o (persons, items, and occasions are fully crossed for all subdomains), and *p* •×*i* º×o • (persons, items, and occasions are fully crossed but subdomains are measured with different items). All these efforts lead to an easier implementation of G-theory in SEM framework, which can not only incorporate properties of multiple regression, path analysis, and confirmatory factor analysis that involve both observed and latent variables, but also yield measures of global fit, tests for the model of interest, and better handling of missing data. To make these benefits concrete, one can imagine the (latent) ability estimates of an OSCE at undergraduate phrases are assumed to be able to predict multiple-choice scores in future licensing exams, while this prediction (i.e., effect) is also assumed to be mediated by self-stigma. These assumptions can stem from theories and/or well-educated guesses, and testing as well as, if necessary, correcting them becomes possible via SEM approaches. We refer readers to comprehensive SEM literature, such as Ullman and Bentler37 and Hancock and Mueller38 for detailed explanations about “what else SEM can offer,” especially the strengths and properties that regular G-theory doesn’t permit. It should be noted that, the vein of SEM for G-theory estimation does not imply that they can be converted from one to another; it becomes possible because the analytical solutions can be found using the same estimation tool, while G-theory and SEM remain to be distinct entities.

Three programs are used the most in applied studies, including G\_String39, 40, the GENOVA Suite (e.g., GENOVA, urGENOVA, and mGENOVA)13, and the EduG41. To date, more software programs are available for G-theory due to the development of estimation frameworks such as SEM approaches and random effect modeling. It’s evident that, since G-theory can be estimated within SEM framework, software programs specifically for SEM such as the Mplus42, the lavaan package43 in the R platform44, the EQS45, the LISREL46, and others become applicable to G-theory estimation. From the perspective of function update frequency, we recommend the Mplus as well as the lavaan of which related tutorials for conducting G-theory analyses are also available.47, 48 In terms of using random effect modeling, various R platform’s packages (e.g., the gtheory49, the lme450, and the glmmTMB51) can be used for different G-theory designs; with random effect modeling, researchers can not only use restricted maximum likelihood (REML) to yield unbiased estimates for small sample scenarios, but also obtain the flexibility of adding other sources of variables (e.g., effects such as genders and pre-test scores) to enrich the G-theory models; related primers for using these software programs to conduct G-theory analyses are presented by Jiang52, Jiang et al53 and Moore49. Following the vein of random effect modeling, Bayesian estimations through the BUGS/JAGS software programs for both univariate and multivariate G-theory were also presented,54, 55 while these programs can be called in the R platform, which essentially wraps data imports, cleaning, and a series of analyses together as a “onesie”. Bayesian properties bring additional advantages to G-theory estimations, for example, (1) treating missing responses as parameter estimates to lessen the hazards of data incompleteness, (2) incorporating prior information (from the literature or past experience) into the modeling, and (3) constructing credible intervals (similar to confidence intervals) for any customized parameters.56 A recent (interactive) web app named the gTheoryShiny57 provides a user-friendly interface for applied researchers without solid programming skills; users can execute advanced functions (e.g., detecting and recommending structures of G-theory designs, imputing missing responses, extracting latent ability estimates, replacing link functions for binary and count responses) via simple point-and-click.

**A Walkthrough Example**

**Conclusions**

Generalizability theory is widely used to evaluate the generalizability in performance assessments in various testing scenarios.4-8,58-62 A solid psychometric foundation is essential to high-stakes decision-making in medical education and assessment, which influences the functionality of health care system. 62 Advanced methods were developed recently to deal with new challenges and extend the application scope for increasing demand. This article covers new progress in generalizability such as cost-effectiveness optimization, multivariate extensions of current results, confidence intervals of relevant coefficients, estimation methods for different inquiries, and open-source software options.

This paper is meant to be non-technical such that statistical and programming details are omitted; it aims at informing medical education researchers about what G-theory can provide more to performance assessments. Therefore, knowing the conceptual properties of the points mentioned above is necessary. For example, SEM approaches allow G-theory to borrow auxiliary variables to reduce the biases caused by missing data. We expect applied researchers to collaborate with statisticians when the advanced applications mentioned above are demanded. With the additional functions and features addressed in this paper, future studies on performance assessments in the medical education field can be broadened and deepened. While G-theory is recognized as a method for estimating the precision of measures, its uniqueness in partitioning and quantifying variance can also yield validity evidence. Using Kane’s validity framework, the use of G-theory can investigate validity-related measurement questions, contributing to one or more of the four types of Kane’s validity inferences (scoring, generalization, extrapolation, and implication). For example, if one uses G-theory to determine how many faculty raters and how many individual components of the scoring rubric for each competency were demanded to ensure reliable scoring of portfolios, he/she essentially evaluates validity evidence based on the extent to which the proposed interpretations and uses of portfolio assessment were viable and appropriate; a sample validity claim can be “a primary trait scoring rubric and accompanying traits are relevant for scoring the portfolios”. Therefore, the rubric could include portfolio primary traits (equated to competency), accompanied by trait-level components serving as the criteria for raters’ evaluations; this can easily be an example about scoring validity inference that address whether or not the scores should be combined.

**Disclosure statement**

No potential conflict of interest was reported by the author(s).

**Author Contributions statement**

ZJ and JZ developed the study concept and drafted the manuscript. ZJ and JO conducted the literature review and discussion. LX, YH, FC and HL were involved in drafting and revising the manuscript. All authors agree to be accountable for all aspects of the work.

**Data Availability Statement**

Data sharing is not applicable to this article as no new data were created or analyzed in this study

**Additional information**

**Funding**

This work was supported by the National Natural Science Foundation of China for Young Scholars under Grant 72104006; Peking University Health Science Center under Grant BMU2021YJ010.

**Reference**

1. Epstein RM. Assessment in medical education. *N Engl J Med*. 2007;356(4):387–396.

2. Howley LD. Performance assessment in medical education. *Eval Health Prof*. 2004;27(3):285–303.

3. Cronbach LJ, Rajaratnam N, Gleser GC. Theory of generalizability: A liberalization of reliability theory. *Br J Stat Psychol*. 1963;16(2):137–163.

4. Zhang X, Roberts WL. Investigation of standardized patient ratings of humanistic competence on a medical licensure examination using Many-Facet Rasch Measurement and generalizability theory. *Adv Health Sci Educ*. 2012;18(5):929–944.

5. Park J. Possibility of using the yes/no Angoff method as a substitute for the percent Angoff method for estimating the cutoff score of the Korean Medical Licensing Examination: A simulation study. *J Educ Eval Health Prof*. 2022;19:1–9.

6. Clauser BE, Margolis MJ, Swanson DB. An examination of the contribution of computer-based case simulations to the USMLE step 3 examination. *Acad Med*. 2002;77(Supplement):S80–S82.

7. Harik P, Clauser BE, Grabovsky I, Nungester RJ, Swanson D, Nandakumar R. An examination of rater drift within a generalizability theory framework. *J Educ Meas*. 2009;46(1):43–58.

8. Raymond MR, Clauser BE, Swygert K, van Zanten M. Measurement precision of spoken english proficiency scores on the USMLE step 2 clinical skills examination. *Acad Med*. 2009;84(Supplement):S83–S85.

9. Shavelson RJ, Baxter GP, Gao X. Sampling variability of performance assessments. J *Educ Meas*. 1993;30(3):215–232.

10. Brennan RL, Johnson EG. Generalizability of performance assessments. *Educ Meas: Issues Pract*. 1995;14(4):9–12.

11. Andersen SAW, Nayahangan LJ, Park YS, Konge L. Use of generalizability theory for exploring reliability of and sources of variance in assessment of technical skills: A systematic review and meta-analysis. *Acad Med*. 2021;96(11):1609–1619.

12. Monteiro S, Sullivan GM, Chan TM. Generalizability theory made simple(r): An introductory primer to G-Studies. *J Grad Med Educ.* 2019;11(4):365–370.

13. Brennan RL. Coefficients and indices in generalizability theory. *CASMA Research Report*. 2003;1:1–44.

14. Brennan RL. Raw-score conditional standard errors of measurement in generalizability theory. *Appl Psychol Meas.* 1998;22(4):307–331.

15. Vispoel WP, Morris CA, Kilinc M. Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychol Methods*. 2018;23(1):1–26.

16. Jiang Z, Raymond M, DiStefano C, Shi D, Liu R, Sun J. A monte carlo study of confidence interval methods for generalizability coefficient. *Educ Psychol Meas*. 2022;82(4):705–718.

17. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. *Standards for Educational and Psychological Testing.* Washington: American Educational Research Association; 2014.

18. Tong Y, Brennan RL. Bootstrap estimates of standard errors in generalizability theory. *Educ Psychol Meas*. 2007;67(5):804–817.

19. Clauser BE, Balog K, Harik P, Mee J, Kahraman N. A multivariate generalizability analysis of history-taking and physical examination scores from the USMLE step 2 clinical skills examination. *Acad Med*. 2009;84(Supplement):S86–S89.

20. Margolis MJ, Clauser BE, Cuddy MM, et al. Use of the mini-clinical evaluation exercise to rate examinee performance on a multiple-station clinical skills examination: A validity study. *Acad Med*. 2006;81(Suppl):S56–S60.

21. Choi J, Wilson MR. Modeling rater effects using a combination of generalizability theory and IRT. *Psychol test Assess model.* 2018;60(1):53–80.

22. Marcoulides GA. Selecting weighting schemes in multivariate generalizability studies. *Educ Psychol Meas*. 1994;54(1):3–7.

23. Raymond MR, Jiang Z. Indices of subscore utility for individuals and subgroups based on multivariate generalizability theory. *Educ Psychol Meas.* 2020;80(1):67–90.

24. Jiang Z, Raymond M. The use of multivariate generalizability theory to evaluate the quality of subscores. *Appl Psychol Meas.* 2018;42(8):595–612.

25. Brennan RL, Kim SY, Lee W-C. Extended multivariate generalizability theory with complex design structures. *Educ Psychol Meas*. 2022;82(4):617–642.

26. Brown C, Ross S, Cleland J, Walsh K. Money makes the (medical assessment) world go round: The cost of components of a summative final year objective structured clinical examination (OSCE). *Med Teach*. 2015;37(7):653–659.

27. Walsh K, Jaye P. Cost and value in medical education. *Educ Prim Care*. 2013;24(6):391–393.

28. Marcoulides GA. Maximizing power in generalizability studies under budget constraints. *J Educ Stat*. 1993;18(2):197–206.

29. Marcoulides GA, Goldstein Z. The optimization of generalizability studies with resource constraints. *Educ Psychol Meas*. 1990;50(4):761–768.

30. Marcoulides GA, Goldstein Z. The optimization of multivariate generalizability studies with budget constraints. *Educ Psychol Meas.* 1992;52(2):301–308.

31. Meyer JP, Liu X, Mashburn AJ. A practical solution to optimizing the reliability of teaching observation measures under budget constraints. *Educ Psychol Meas.* 2014;74(2):280–291.

32. Jiang Z, Shi D, Distefano C. A short note on optimizing cost-generalizability via a machine-learning approach. *Educ Psychol Meas.* 2021;81(6):1221–1233.

33. Raykov T, Marcoulides GA. Estimation of generalizability coefficients via a structural equation modeling approach to scale reliability evaluation. *Int J Test.* 2006;6(1):81–95.

34. Jiang Z, Walker K, Shi D, Cao J. Improving generalizability coefficient estimate accuracy: A way to incorporate auxiliary information. *Method Innov*. 2018;11(2):1–14.

35. Jorgensen TD. How to estimate absolute-error components in structural equation models of generalizability theory. *Psych.* 2021;3(2):113–133.

36. Vispoel WP, Hong H, Lee H, Jorgensen TR. Assessing relative and absolute differences in scores within g-theory designs using structural equation modeling versus conventional procedures. 2022. Unpublished.

37. Ullman JB, Bentler PM. Structural equation modeling. In: Schinka JA, Velicer WF, eds. *Handbook of Psychology Volume 2: Research Methods in Psychology*. 2nd ed. Hoboken, NJ: John Wiley & Sons, Inc.; 2012.

38. Hancock GR, Mueller RO. *Structural Equation Modeling: A Second Course*. Charlotte, NC: Information Age Publishing; 2013.

39. Bloch R, Norman G. Generalizability theory for the perplexed: A practical introduction and guide: Amee guide no. 68. *Med Teach.* 2012;34(11):960–992.

40. Taşdelen Teker G. Coping with unbalanced designs of generalizability theory: G String V. *Int J Assess Tool Educ*. 2019;6(5):57–69.

41. Cardinet J, Johnson S, Pini G. *Applying Generalizability Theory Using EduG*. New York, NY: Routledge; 2011.

42. Muthén LK, Muthén BO. *Mplus User’s Guide.* 8th ed. Los Angeles, CA: Muthén & Muthén; 1998–2017.

43. Rosseel Y. lavaan: An R package for structural equation modeling. *J Stat Softw.* 2012;48(2):1–36.

44. R Core Team. *R: A language and environment for statistical computing*. Vienna, AU: R Foundation for Statistical Computing; 2022.

45. Bentler PM. *EQS structural equations program manual.* Encino, CA: Multivariate Software; 1995.

46. Jöreskog KG, Sörbom D. *LISREL 11: Structural equation modeling*. Chapel Hill, NC: Scientific Software International, Inc; 2021.

47. Raykov T, Marcoulides GA. *Introduction to Psychometric Theory*. New York, NY: Routledge; 2011.

48. Vispoel WP, Xu G, Schneider WS. Using parallel splits with self-report and other measures to enhance precision in generalizability theory analyses. *J Pers Assess.* 2022;104(3):303–319.

49. Moore CT. gtheory: Apply generalizability theory with R. 2016.

50. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. *J Stat Softw*. 2015;67(1):1–48.

51. Brooks ME, Kristensen K, Van Benthem KJ, et al. Glmmtmb balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *R J*. 2017;9(2):378–400.

52. Jiang Z. Using the linear mixed-effect model framework to estimate generalizability variance components in R. *Methodology*. 2018;14(3):133–142.

53. Jiang Z, Raymond M, Shi D, DiStefano C. Using a linear mixed-effect model framework to estimate multivariate generalizability theory parameters in R. *Behav Res Methods*. 2020;52(6):2383–2393.

54. Jiang Z, Skorupski W. A bayesian approach to estimating variance components within a multivariate generalizability theory framework. *Behav Res Methods*. 2018;50(6):2193–2214.

55. LoPilato AC, Carter NT, Wang M. Updating generalizability theory in management research. *J Manage*. 2015;41(2):692–717.

56. Marsman M, Wagenmakers E-J. Bayesian benefits with JASP. *Eur J Dev Psychol*. 2017;14(5):545–555.

57. Zhang J, Jiang Z. gTheoryShiny: An online application for interactive g-theory inference. *PsyArxiv.* Preprint posted online November 30, 2022. doi:10.31234/osf.io/d6mtk.

58. Trejo-Mejía J A, Sánchez-Mendiola M, Méndez-Ramírez I, Martínez-González A. Reliability analysis of the objective structured clinical examination using generalizability theory. *Med Educ Online*. 2016;21(1):31650.

58. Baig L, Violato C. Temporal stability of objective structured clinical exams: A longitudinal study employing item response theory. *BMC Med Educ*. 2012;12:121.

60. Vallevand 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.

61. Hatala R, Marr S, Cuncic C, Bacchus CM. Modification of an OSCE format to enhance patient continuity in a high-stakes assessment of clinical performance. *BMC Med Educ*. 2011;11:23.

62. Schauber S K, Hecht M, Nouns Z M. Why assessment in medical education needs a solid foundation in modern test theory. Adv Health Sci Educ. 2018;23(1):217–232.

Haladyna, S. J., & Kramer, G. A. (2004). The validity of subscores for a credentialing test. Evaluation and the health professions, 24 (7), 349–368.

Norcini, J. J., Blank, L. L., Duffy, F. D., & Fortna, G. S. (2003). The mini-CEX: a method for assessing clinical skills. *Annals of internal medicine*, *138*(6), 476-481.