**Advancing the Application of Generalizability Theory to Performance Assessments in Medical Education**

Performance assessments in medical education, of which the designs are often complex play an important role in evaluating clinical competence [Epstein 2007]. 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 interpretation [Howley 2004]; we address the last point in this paper—generalizability theory, or namely G-theory [Cronbach et al.1963]. 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) [Zhang & Roberts 2013], the Korean Medical Licensing Examination [Park 2022], and the United States Medical Licensing Examination (USMLE) [Clauser et al 2002; Harik et al 2009; Raymond et al 2009]; it’s not difficult to picture that the volume of studies of this kind took place in low-stakes setting would be much larger.

Why is G-theory a prevalent choice for investigating performance assessments? Statistically, a performance assessment is regarded as student performance sampled from a complex universe consisted of all possible *facets* (e.g., tasks, occasions, raters, and measurement methods). G-theory allows researchers to present evidence bearing on the generalizability and convergent validity of performance assessments, taking the sampling properties within the complex setting into account [Shavelson et al. 1993]. Methodologically, G-theory provides a framework for estimating, determining, and designing the generalizability of various observations or ratings [Brennan & Johnson 1995]; these features fit the universe assumption for performance assessments well and, therefore, have been widely utilized in the 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 facets contributes to the observed score can be disentangled. On the other hand, a D-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 researchers quantified results, if shrinking settings from the current design is considered (e.g., reducing the number of raters and/or stations). Dimensionality-wise, G-theory has univariate and multivariate versions, where the latter permits subdomains, of which the associations in between are provided. The discrepancy of the dimensionality 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 al [2021] and a tutorial by Monteiro et al [2019].

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

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) [Brennan 2003]. 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 [Brennan 1998], (2) cut-score specific coefficients [Vispoel et al 2018], and (3) confidence intervals (CIs) of the coefficients [Jiang et al 2022]. 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 error is identical across the span of sum (observed) scores. test. The Standards for Educational and Psychological Testing [American Educational Research Association 2018] 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-score because, intuitively, we always prefer setting the cut-score to an area 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 to report CIs for the consistency coefficients. It used to be difficult to derive a versatile formula for computing CIs for both variance/covariance and coefficient estimates. Tong and Brennan [2007] recommended bootstrapping procedures to handle the inquiry, and Jiang et al [2022] 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 al [2009] 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 al [2006] 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. Nevertheless, it remains less seen than the applications of univariate G-theory. An important reason is that many primers and tutorials only mention multivariate G-theory at the end as a further extension. In addition, statistical understanding and software programs for multivariate G-theory set a high threshold to applied researchers. It’s worth noting 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 unidimensional. Typical scenarios include an OSCE and an objective structured assessment of technical skill (OSATS) where subdomains are of concerns, both theoretically and practically. Known properties of multivariate G-theory include: (1) providing correlation estimates between pairs of subdomains for a given facet [Choi, J., & Wilson 2018]; (2) tuning weights of subdomains (i.e., effective weights) to form meaningful composite scores and global generalizability coefficients [Marcoulides 1994]; (3) examining the validity of subscores when subdomain performance outcomes are reported to both individuals and populations [ Jiang & Raymond 2018; Raymond & Jiang 2020]; (4) 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) [Brennan et al 2022].

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. 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 al [2015] and Walsh and Jaye [2013], standardized patients, raters, and stations are costly in OSCEs. Therefore, varying their numbers should be constrained by financial budgets. Marcoulides[1993], Marcoulides and Goldstein [1990; 1992] as well as Meyer et al [2014] derived series of mathematical formulas to find the optimal solution maximizing generalizability within a budge constraint. Given these formulas tend to work on a particular design, its application is relatively limited if one’s design is off the list. Jiang et al [Jiang, Z., Shi, D., & Distefano, C 2021] 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 a 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 generalizability of a particular solution, where the constraint function ensures that the solution’s costs (i.e., the sum of unit costs multiplying the solution’s numbers) remain affordable. Essentially, these cutting-edge approaches can drive the assessments to be more cost-effective.

Raykov and Marcoulides [2006] 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 al [2018] and Jorgensen [2021] further extended the SEM approach to include auxiliary variables in the models and estimate absolute-error components, respectively. Vispoel et al [2022 with Hong] 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 scenarios. 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 become possible via SEM approaches. We refer readers to comprehensive SEM literature, such as Ullman and Bentler [2012] and Hancock and Mueller [2013], for detailed explanations about “what else SEM can offer”, especially the strengths and properties that regular G-theory doesn’t permit.

Three programs are used the most in applied studies, including G\_String [Bloch & Norman 2012; Teker 2019], the GENOVA Suite (e.g., GENOVA, urGENOVA, and mGENOVA)[ Brennan 2003], and the EduG [Cardinet el al 2011]. 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 converted to SEM, software programs specifically for SEM such as the Mplus [], the lavaan package [Rosseel 2012] in the R platform [], the EQS [], the LISREL [], 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 [Raykov & Marcoulides 2011; Vispoel et al 2022 with Xu]. In terms of using random effect modeling, the R platform’s various packages such as the gtheory [Moore 2016], the lme4 [], and the glmmTMB [Brooks et al, 2017] 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 models; related primers for using these software programs to conduct G-theory analyses are presented by Jiang [2018], Jiang et al [2020] and Moore [2016]. 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 [Jiang & Skorupski 2018; LoPilato et al 2015]], 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 the past experience) to the modeling, and (3) constructing credible intervals (similar to confidence intervals) for any customized parameters [Marsman & Wagenmakers 2017]. A recent (interactive) web app named the gTheoryShiny [**Zhang & Jiang 2022**] 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 data) via simple point-and-click.

This paper is meant to be non-technical such that statistical and programming details are omitted; it aims at informing medical education researchers what G-theory can provide more to performance assessments. Therefore, knowing the conceptual properties of the aforementioned points is necessary, for example, SEM approaches allow G-theory to borrow auxiliary variables to reduce the biases caused by missing data. We expect that 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 related to performance assessments in the medical education field can be broadened and deepened.

Choi, J., & Wilson, M. R. (2018). Modeling rater effects using a combination of generalizability theory and IRT. *Psychological Test and Assessment Modeling*, *60*(1), 53-80.

Brennan, R. L., Kim, S. Y., & Lee, W. C. (2022). Extended Multivariate Generalizability Theory With Complex Design Structures. *Educational and Psychological Measurement*, *82*(4), 617-642.

Marcoulides, G. A. (1994). Selecting weighting schemes in multivariate generalizability studies. *Educational and Psychological Measurement*, *54*(1), 3-7.

Jiang, Z., Raymond, M., DiStefano, C., Shi, D., Liu, R., & Sun, J. (2022). A Monte Carlo study of confidence interval methods for generalizability coefficient. *Educational and Psychological Measurement*, *82*(4), 705-718.

Cardinet, J., Johnson, S., & Pini, G. (2011). *Applying generalizability theory using EduG*. Routledge.

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.

Jiang, Z., & Skorupski, W. (2018). A Bayesian approach to estimating variance components within a multivariate generalizability theory framework. *Behavior research methods*, *50*(6), 2193-2214.

Vispoel, W. P., Xu, G., & Schneider, W. S. (2022). Using Parallel Splits with Self-Report and Other Measures to Enhance Precision in Generalizability Theory Analyses. *Journal of personality assessment*, *104*(3), 303-319.

Jiang, Z., Raymond, M., Shi, D., & DiStefano, C. (2020). Using a linear mixed-effect model framework to estimate multivariate generalizability theory parameters in R. *Behavior research methods*, *52*(6), 2383-2393.

Jiang, Z., Shi, D., & Distefano, C. (2021). A Short Note on Optimizing Cost-Generalizability via a Machine-Learning Approach. *Educational and Psychological Measurement*, *81*(6), 1221–1233.

Meyer J. P., Liu X., Mashburn A. J. (2014). A practical solution to optimizing the reliability of teaching observation measures under budget constraints. Educational and Psychological Measurement, 74(2), 280-291.

Brown, C., Ross, S., Cleland, J., & Walsh, K. (2015). Money makes the (medical assessment) world go round: The cost of components of a summative final year Objective Structured Clinical Examination (OSCE). *Medical teacher*, *37*(7), 653-659.

Walsh, K., & Jaye, P. (2013). Cost and value in medical education. *Education for Primary Care*, *24*(6), 391-393.

Goldstein, Z. & Marcoulides, G. (1991). Maximizing the coefficient of generalizability in decision studies. *Educational and Psychological Measurement*, 51, 79-88.

Marcoulides, G. A., & Goldstein, Z. (1990). The optimization of generalizability studies with resource constraints. *Educational and Psychological Measurement,* 50(4), 761-768.

Marcoulides, G. A. (1993). Maximizing power in generalizability studies under budget constraints. Journal of Educational Statistics, 18(2), 197-206.

Marcoulides, G. A., & Goldstein, Z. (1992). The optimization of multivariate generalizability studies with budget constraints. *Educational and Psychological Measurement*, 52(2), 301-308.

Jiang, Z., & Raymond, M. (2018). The use of multivariate generalizability theory to evaluate the quality of subscores. *Applied psychological measurement*, *42*(8), 595-612.

Jiang, Z., Walker, K., Shi, D., & Cao, J. (2018). Improving generalizability coefficient estimate accuracy: A way to incorporate auxiliary information. *Methodological Innovations*, *11*(2), 2059799118791397.

Jiang, Z. (2018). Using the linear mixed-effect model framework to estimate generalizability variance components in R: A lme4 package application. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *14*(3), 133.

Jorgensen, T. D. (2021). How to estimate absolute-error components in structural equation models of generalizability theory. *Psych*, *3*(2), 113-133.

Raymond, M. R., & Jiang, Z. (2020). Indices of subscore utility for individuals and subgroups based on multivariate generalizability theory. *Educational and psychological measurement*, *80*(1), 67-90.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. Journal of Statistical Software, 48(2), 1-36. https://doi.org/10.18637/jss.v048.i02

LoPilato, A. C., Carter, N. T., & Wang, M. (2015). Updating generalizability theory in management research: Bayesian estimation of variance components. *Journal of Management*, *41*(2), 692-717.

Marsman, M., & Wagenmakers, E. J. (2017). Bayesian benefits with JASP. *European Journal of Developmental Psychology*, *14*(5), 545-555.

Raykov, T., & Marcoulides, G. A. (2006). Estimation of generalizability coefficients via a structural equation modeling approach to scale reliability evaluation. *International Journal of Testing*, *6*(1), 81-95.

Vispoel, W. P., Hong, H, Lee, H., & Jorgensen, T. R. (2022). Assessing relative and absolute differences in scores within G-theory designs using structural equation modeling versus conventional procedures. Manuscript submitted for publication

Clauser, Brian E.; Balog, Kevin; Harik, Polina; Mee, Janet; Kahraman, Nilufer. A Multivariate Generalizability Analysis of History-Taking and Physical Examination Scores From the USMLE Step 2 Clinical Skills Examination. Academic Medicine: October 2009 - Volume 84 - Issue 10 - p S86-S89 doi: 10.1097/ACM.0b013e3181b36fda

Margolis, Melissa J.; Clauser, Brian E.; Cuddy, Monica M.; Ciccone, Andrea; Mee, Janet; Harik, Polina; Hawkins, Richard E.. Use of the Mini-Clinical Evaluation Exercise to Rate Examinee Performance on a Multiple-Station Clinical Skills Examination: A Validity Study. Academic Medicine: October 2006 - Volume 81 - Issue 10 - p S56-S60 doi: 10.1097/01.ACM.0000236514.53194.f4

Brennan, R. L., & Johnson, E. G. (1995). Generalizability of performance assessments. Educational Measurement: Issues and Practice, 14(4), 9-12.

Howley, L. D. (2004). Performance assessment in medical education: where we’ve been and where we’re going. *Evaluation & the health professions*, *27*(3), 285-303.

Shavelson, R. J., Baxter, G. P., & Gao, X. (1993). Sampling variability of performance assessments. *Journal of educational Measurement*, *30*(3), 215-232.

Epstein, R. M. (2007). Assessment in medical education. *New England journal of medicine*, *356*(4), 387-396.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137-163.

Zhang, X., & Roberts, W. L. (2013). Investigation of standardized patient ratings of humanistic competence on a medical licensure examination using Many-Facet Rasch Measurement and generalizability theory. *Advances in Health Sciences Education*, *18*(5), 929-944.

Park, J. (2022). 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. *Journal of Educational Evaluation for Health Professions*, *19*.

Raymond, M. R., Clauser, B. E., Swygert, K., & van Zanten, M. (2009). Measurement precision of spoken English proficiency scores on the USMLE Step 2 Clinical Skills Examination. *Academic Medicine*, *84*(10), S83-S85.

Andersen, Steven Arild Wuyts, ; Nayahangan, Leizl Joy, MHCM; Park, Yoon Soo ; Konge, Lars. Use of Generalizability Theory for Exploring Reliability of and Sources of Variance in Assessment of Technical Skills: A Systematic Review and Meta-Analysis. Academic Medicine: November 2021 - Volume 96 - Issue 11 - p 1609-1619 doi: 10.1097/ACM.0000000000004150

Harik, P., Clauser, B. E., Grabovsky, I., Nungester, R. J., Swanson, D., & Nandakumar, R. (2009). An examination of rater drift within a generalizability theory framework. *Journal of Educational Measurement*, *46*(1), 43-58.

Clauser, B. E., Margolis, M. J., & Swanson, D. B. (2002). An examination of the contribution of computer-based case simulations to the USMLE step 3 examination. *Academic medicine*, *77*(10), S80-S82.

Monteiro, S., Sullivan, G. M., & Chan, T. M. (2019). Generalizability theory made simple (r): an introductory primer to G-studies. *Journal of graduate medical education*, *11*(4), 365-370

Brennan, R. L. (2003). Coefficients and indices in generalizability theory. *Center for advanced studies in measurement and assessment, CASMA research report*, *1*, 1-44.

American Educational Research Association. (2018). *Standards for educational and psychological testing*. American Educational Research Association.

Brennan, R. L. (1998). Raw-score conditional standard errors of measurement in generalizability theory. *Applied Psychological Measurement*, *22*(4), 307-331.

Tong, Y., & Brennan, R. L. (2007). Bootstrap estimates of standard errors in generalizability theory. *Educational and Psychological Measurement*, *67*(5), 804-817.

Teker, G. T. (2019). Coping with unbalanced designs of generalizability theory: G string V. *International Journal of Assessment Tools in Education*, *6*(5), 57-69.

Brennan, R.L. Generalizability Theory. New York, Springer, 2003

Bloch R., Norman G. Generalizability theory for the perplexed: a practical introduction and guide: AMEE Guide No. 68. Med Teach. 2012;34(11):960-92

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, *23*(1), 1.

Ullman, J. B., & Bentler, P. M. (2012). Structural equation modeling. *Handbook of Psychology, Second Edition*, *2*.

Hancock, G. R., & Mueller, R. O. (Eds.). (2013). *Structural equation modeling: A second course*. Iap.