Health Professions Education

Volume 6 | Issue 2 Article 19

2020-06-01

Generalizability Theory's Role in Validity Research: Innovative Applications in Health Science Education

Clarence Kreiter
Department of Family Medicine, Office of Consultation and Research in Medical Education, University of Iowa Carver College of Medicine, Iowa City, IA, USA

Nikki Bibler Zaidi
Research.Innovation.Scholarship.Education.(RISE), Michigan Medicine, University of Michigan, Ann Arbor, MI, USA

Recommended Citation

Kreiter, Clarence and Zaidi, Nikki Bibler (2020) "Generalizability Theory's Role in Validity Research: Innovative Applications in Health Science Education," *Health Professions Education*: Vol. 6: Iss. 2, Article 19.
DOI: 10.1016/j.hpe.2020.02.002
Available at: https://hpe.researchcommons.org/journal/vol6/iss2/19

Available online at www.sciencedirect.com

ScienceDirect

Health Professions Education 6 (2020) 282–290

www.elsevier.com/locate/hpe

Generalizability Theory's Role in Validity Research: Innovative Applications in Health Science Education

Clarence Kreiter [a,*], Nikki Bibler Zaidi [b]

[a] *Department of Family Medicine, Office of Consultation and Research in Medical Education, University of Iowa Carver College of Medicine, Iowa City, IA, USA*

[b] *Research.Innovation.Scholarship.Education.(RISE), Michigan Medicine, University of Michigan, Ann Arbor, MI, USA*

Received 12 November 2019; revised 21 January 2020; accepted 6 February 2020
Available online 24 February 2020

Abstract

Purpose: While generalizability (G) theory is widely recognized as a method for estimating the reliability (precision) of measures, its unique approach to partitioning and quantifying variance also yields validity (accuracy) evidence. Yet, G theory's ability to provide validity evidence is much less understood and established in the literature. The purpose of this paper is to demonstrate G theory's potential for addressing a wide array of health sciences education validity questions.
Methods: Using Kane's validity framework, this paper explores the use of G theory in the health sciences literature by presenting a number of validity applications. The G studies investigate validity-related measurement questions and demonstrate how G theory contribute to one or more of the four types of Kane's validity inferences (scoring, generalization, extrapolation, and implication).
Results: Each G study is linked to one of Kane's four types of validity inferences. The studies presented in this paper demonstrate how a G theory analysis of score variance, usually within existing (in vivo) assessment data, simultaneously provides researchers with evidence regarding both reliability and validity and offers a more accurate portrayal of the relationship between the two.
Discussion: Because each application of G theory is unique, the examples provided do not represent the entire range of potential applications, but rather demonstrate the methodological flexibility of G theory in addressing complex validity questions. Further advances will require researchers to develop and share additional innovations in G study design and work together to develop a consensus regarding its role in validity research.

Keywords: Generalizability theory; Reliability; Validity; Assessment; Methodology

1. Introduction

Generalizability (G) theory is a statistical framework used to model and analyze measurements (e.g., multiple-choice tests, direct observations, or ratings). G theory uses generalizability (G) studies to model the composition of assessment scores and decision (D) studies to forecast the reliability of measurements

* Corresponding author. University of Iowa Carver College of Medicine – Office of Consultation and Research in Medical Education.(OCMRE) 1204 MEB, USA.
E-mail address: clarence-kreiter@uiowa.edu (C. Kreiter).
Peer review under responsibility of AMEEMR: the Association for Medical Education in the Eastern Mediterranean Region

https://doi.org/10.1016/j.hpe.2020.02.002

given various conditions (e.g., number of items, number of raters, number of occasions, number of stations, etc.) under which they could be obtained.[1] While a G study's estimate of universe (true) score variance is commonly used to generate estimates of reliability, its ability to quantify multiple, individual sources of measurement error also yields validity evidence. Although classical measurement theory regards reliability research as addressing score precision and validity investigations as yielding evidence of accuracy, G theory renders this distinction as somewhat outdated since a G study can partition and quantify score variance in ways that simultaneously provide information regarding both precision (reliability) and accuracy (validity).

Classical test theory (CTT) suggests that for any assessment we administer, the "true score" is the only desired source of variance; yet, a score will always contain unwanted sources of error—both random and systematic—that will impact both reliability and validity. The more inconsistent the score estimate, the less reliable it is; and, the greater the magnitude of unwanted variance within a score, the less valid the interpretation becomes. Health science educators generally understood the classical relationship with the commonly cited dictum: *reliability is a necessary but not sufficient condition for validity* (technically: reliability sets the upper limit on the correlation between a test measure and a criterion). So, educators tended to expect that increases in score reliability would also increase score validity. However, in many assessment environments, the relationship between reliability and validity is more nuanced. In fact, within health science education, an inverse relationship between reliability and validity has often been observed. For instance, efforts to improve test validity by switching from multiple-choice tests to performance assessments have produced lower reliabilities. Similarly, initiatives to improve reliability with standardized testing procedures is thought to reduce validity. From a CTT perspective, these assessment outcomes might appear counter-intuitive. However, G theory's liberalized perspective of the classical measurement approach does model and predict the real–world relationships between reliability and validity (e.g., the reliability-validity paradox).[1]

Even though the developers of G theory promoted its role in validity research more than 45 years ago, it is still primarily regarded as a method for estimating only the reliability of assessments—while its ability to provide validity evidence is much less understood and established.[2] This may be partially attributed to the unique and often nuanced presentation of the concepts and statistical designs that accompany each new application. Cronbach alluded to this in an early discussion of the topic, stating that G theory has "a protean quality. The procedures and even the issues take a new form in every context."[3]; p.201). It is this protean nature that has made it difficult to provide widely applicable and practical guidance regarding G study validity applications and made extant advice necessarily theoretical and abstract in nature.[4] Similarly, the notion of validity has been contextualized and applied somewhat differently within health science education, which has challenged commonly accepted understandings of validity.[5] Fortunately, however, as the number of G theory validity applications in health science education accumulate, an opportunity is emerging to provide methodological guidance by documenting existing G theory research within a contemporary validity framework. This paper describes G studies specifically designed to investigate validity-related measurement questions and identifies how these G studies contribute to one or more of the four types of validity inference (*scoring, generalization, extrapolation*, and *implication*) as described by Kane.[6] While the examples presented in this paper generally match this contemporary concept of validity, certain examples point to areas where modern validity theory might need to be expanded. It is also clear that while this paper focuses on the quantitative aspects of these inferences, there is also the potential for qualitative evidence to contribute to each of the four inferences. The eclectic set of G theory examples presented here demonstrate how a G theory analysis of score variance, usually within existing (*in vivo*) assessment data, simultaneously provides researchers with evidence regarding both reliability and validity and offers a more accurate portrayal of the relationship between the two.

2. Validity applications of G theory

2.1. Investigating case specificity: modeling and partitioning random error

One of the earliest validity applications of G theory in health science education research involved the investigation of *case specificity*. The term case specificity suggests that clinical problem-solving depends largely on the case presented.[7] The case-specific aspect of performance assessment was brought to the attention of medical educators in a book by[8] summarizing their attempts to measure medical reasoning within

real and simulated clinical encounters. In describing their observations, the authors made special note of the weak correlations (<.25) they observed between performance scores across clinical cases. They surmised that these low correlations implied that clinical performance measures depended more upon case-related knowledge than the ability to apply reasoning skills.[8] Because their interpretation of the correlational evidence questioned the existence of the clinical reasoning construct, their research set the stage for a series of important validity-related research questions. Although simple correlation coefficients provided the initial evidence for case specificity, researchers soon began to employ G theory to examine the variance not shared across clinical cases (i.e., error variance).

In G theory investigations of case specificity, researchers first estimated the magnitude of the *person* (p)-*by*-*case* (c) (pc) variance component.[9] As expected, given the low correlation between clinical cases, the pc interaction variance component was found to be quite large. While this finding did not dramatically enhance our understanding of case specificity, it did motivate researchers to take the next step and develop a more elaborated model of the random error in clinical performance assessments. By carefully modeling a complete replication in the *universe of generalization* (the set of conditions to which a decision maker plans to generalize), researchers were better able to identify and understand the facets contributing to measurement error in performance assessments. This modeling and analysis helped demonstrate that earlier studies had included unrecognized sources of error within the pc variance estimate.[10] Most notably, an *occasion* (o) facet and un-modeled residual error were often confounded in pc variance estimates. Subsequent G studies examining both experimental data (repeating the same case on different occasions – p x c x o) and *in vivo* performance assessment data confirmed that most of the estimated pc variance from G studies originated from random error that was unrelated to the clinical knowledge required for a specific medical case.[10–12] In short, pc variance estimates had reflected hidden or confounded error not explicitly modeled in previous research. While a complete replication perspective modeled with G theory allowed researchers to better understand case specificity and identify the relevant sources of measurement error,[13] Jarjoura et al. used a different application of G theory to provide another perspective on case specificity.

Jarjoura et al.[13] utilized multivariate generalizability (MVG) to model an SP-based OSCE performance assessment that generated multiple scores per case. The assessment data used in their study characterized performances from 96 medical students (p) rotating through 16 clinical cases (c). Each case in that study generated four component skill scores (s): 1.) history, 2.) physical exam, 3.) diagnosis, and 4.) treatment. Although the univariate pc error variance for the overall case score was large, MVG estimates of the correlation (covariance) between the residual errors for the four skill scores was close to zero (r = .08, .08, .04, .07, .01, .21 – average r = .11). This clearly demonstrated that the *student-by-case-by-skill* (pcs) interaction, rather than *student-by-case* (pc) interaction was the primary source of measurement error. The authors further concluded that ignoring the multivariate nature of performance assessments was likely to result in a biased (increased) estimate of pc measurement error in a univariate model, and that variation in case-specific knowledge related to the medical content across cases explained only a small portion of the total error variance. They went on to suggest that given the multidimensional nature of medical cases, the validity of performance assessments could be substantially enhanced, and the measurement error reduced, by applying case specific weights to component skill scores in computing a total case score. These weights would be selected to provide a specific emphasis on certain skill areas, minimizing the overall error variance. A subsequent study by Schuwirth and van der Vleuten[14] also challenged the defensibility of combining scores into a total "trait" measure when scoring OSCEs and the mini-CEX. Aggregating scores across stations or cases assumes, sometime inappropriately, the interchangeability of clinical skills.

On the other hand, some scores are expected to vary (i.e., "state" variables) and should not be considered interchangeable. Although not addressed in this paper, validity applications of G theory should consider whether scores come from tau equivalent tests (i.e., individuals are assumed to have a constant true score over tests, but the error variances may vary across tests) or essentially tau equivalent (i.e., tests differ in their true score means but not true score variance). This distinction impacts scoring inferences.[15]

Although these examples of validity evidence do not fit seamlessly within Kane's validity framework, they do reflect on an *extrapolation* inference. At the most basic level, this research investigates what is being measured by these performance assessments. If the construct being measured is in fact consistent with our understanding of the skills required to provide high quality patient care, then an extrapolation argument is supported.

2.2. Investigating construct irrelevant variance (CIV): partitioning systematic true score variance using sampled in vivo test date

G theory has been shown to be useful in identifying and quantifying sources of construct irrelevant variance (CIV) that threaten validity. CIV represents a source of systematic error that tends to be consistent across observations of an individual. This type of nonrandom error is introduced by a variety of psychological and situational factors and tends to manifest as construct-irrelevant test difficulty and construct-irrelevant contamination in score interpretation.[16] In a G study designed to estimate systematic error, Kreiter et al.[17] investigated the degree to which ratings of clerkship preceptors' instructional skills might be inappropriately assessing characteristics of the clinic in which the preceptor worked. In terms of Kane's validity inferences, this would threaten *extrapolation* as scores might not reflect real-world teaching performance in the clinic.[6] To investigate this possibility, the researchers employed sampling along with G theory to dissect the proportion of score variance reflecting a preceptor's teaching performance from the proportion of variance associated with the clinic in which the preceptor worked. The study allowed a perspective on the validity and fairness of using means from this rating process to measure a teacher's (preceptor's) effectiveness. The objective of the investigation was to assess the degree to which preceptors working in clinics less suited to the educational mission were experiencing CIV in the form of systematic negative rating bias. In other words, the degree to which the same preceptors working in different clinics would receive different scores.

Since for all foreseeable applications of the rating form, preceptors would always be nested within a single clinical site, the study[17] regarded this design feature as an immutable aspect of the rating process. To investigate the impact of this rating design, the researchers used two specialized samples (A & B) of *in vivo* rating data. For sample A, each rating was of a different preceptor nested within a clinic site. This implied that within the G study of sample A, the object-of-measurement was site, and the universe (true) score reflected only the influence of site. In sample B, multiple ratings of the same preceptor within a site were analyzed. The measurement model for sample B treated preceptor as the object-of-measurement despite the fact that preceptors were nested within a single site. This effectively confounded preceptor and site characteristics, producing a universe score reflecting the combined influence of the clinic and preceptor. By examining the difference in the magnitude of universe score variance in samples A and B, the researchers were able to quantify the degree to which preceptor ratings reflected site-related CIV. In this rating process, the researchers found that approximately half the variance that had been previously attributed to the preceptor (teacher), was in fact a reflection of the clinical site in which the preceptor worked. With the type of data used in this study, ANOVA hypothesis testing of site means was not a viable option. From a methodological/statistical perspective, this research offered an example of how validity issues related to CIV and systematic bias could be addressed with a G study analysis of purposively sampled data. While G theory is typically associated with calculating different types of reliability through quantifying the components of random error, this research demonstrated that G theory could also offer insight into the systematic error that impacts validity.

2.3. Investigating characteristics of the measured construct related to scoring/rating procedures

To investigate which student characteristics were assessed by a clerkship rating form, a recent G study examined the degree to which reliability was enhanced by increasing the number of days over which a preceptor was able to observe a medical student within an emergency medicine (EM) clerkship.[18] A G study was conducted to estimate the relative magnitude of rater and occasion effects on clinical evaluations conducted within an EM clerkship rotation at a large Midwestern medical school. The assessment form was standardized, but the rating process was highly unstandardized and the ratings were, to an unknown degree, dependent upon the characteristics of the patients entering the clinic (along with other clinic-related factors). Given that a student's day-to-day exposure to cases within the clerkship was quite variable, it was logical to assume that if ratings were based upon the direct observation of students performing EM clinical skills, each day would afford unique observational opportunities for assessing those skills. Since we know from OSCE score data that the correlation between performance scores across medical cases is low, it is logical to assume that if the clerkship ratings reflected skills similar to those measured by an OSCE, increasing the number of days (which would also increase the number of cases) over which a student was observed should enhance score precision (reliability) on the EM clinical clerkship measure. If the true score variance was not

substantially increased by adding additional days over which a preceptor evaluated a student, it could be inferred that these ratings reflected something substantially different than what an OSCE score measured. In the context of Kane's validity inferences, this relates to *scoring* (Is the scoring criteria appropriate for assessing these student characteristics?), *generalization* (Is this a representative sample of students, raters, and clinical occasions?), and *extrapolation* (Are ratings measuring specific procedural skills required within a real clinical setting, or alternately, a rater's subjective impression of those skills?).[6] To assess these questions, a G study examined a completely nested *occasion-nested-within-rater-nested-within-person* (o:r:p) data collection design. Preceptors completed an evaluation at the end of each clerkship day, and a balanced random sample of evaluations in which raters observed a student on multiple days was analyzed. Although previous G studies provided estimates of reliability contingent upon a varying number of independent ratings (each by a different rater), this was the first study to estimate the amount of information provided by multiple ratings from the same rater (Kreiter, 1998).

The findings demonstrated that increasing the number of days (cases) over which a preceptor rated the same student had little impact on the reliability of the mean rating. In other words, if a preceptor rated the same student on multiple days, rather than on only a single day, there was little increase in the precision of the student's mean score. On the other hand, if a different preceptor rated a student each day, a large increase in reliability was observed. In fact, rater-related variance was more than three times greater than occasion-related variance. Having a single rater repeatedly rate a student on different days (cases) did not generate a reliable mean score. Further, this finding had important validity implications by suggesting that these ratings did not primarily reflect a detailed observation of students performing clinical tasks. Rather, the ratings appeared to be a product of an intuitive appraisal of a student's capabilities, and that the appraisal was formed early in the student–preceptor interaction (e.g., a halo effect). While experimental research has found that under certain controlled conditions (*in vitro*), raters can change their global assessment of a student's clinical skills, this study of *in vivo* ratings demonstrated that preceptors were not likely to change their judgments over multiple repeated observations within an actual clerkship environment, and that a mean clerkship rating was not likely to reflect a detailed observation of a student's clinical performance.[19]

2.4. Investigating the relationship between measured constructs (concurrent validity)

Over the last 40 years, medical educators have developed several computer-based case simulations designed to assess clinical reasoning. A primary goal for the scores generated by these simulations was to provide unique information about examinees beyond that already obtained with multiple-choice question (MCQ) examinations.[20] To address whether this goal was achieved, Clauser et al.[21] employed MVG analyses using scores from the United States Medical Licensing Examination (USMLE) computer-based case simulation (CCS) and the USMLE MCQ exam. This analytical approach estimated the generalizability (reliability) of MCQ and CCS scores, the relationship between them, and the reliability of a composite score. In that study, 2500 examinees sitting for the USMLE Step 3 exam responded to samples of both MCQs and CCSs. Completely crossed and balanced samples provided data for 10 MVG studies examining performance on MCQs ($n = 180$) and CCSs ($n = 4$). They found that the MCQ scores were significantly more reliable than the CCS scores per time interval, but that the universe (true score) correlation between the two tests was just $r = .69$. While the CCS provided less reliable information, the two exams (MCQ & CCS) did measure different skill attributes. These results further suggested that even if one assumed the CCS was more valid for assessing clinical reasoning, the much higher reliability of the MCQ exam and its significant relationship with the CCS justified the inclusion of MCQs in a composite score designed to reflect an examinee's clinical reasoning skills.

In another example, an investigation of the validity and reliability of portfolio assessment in dental schools by Gadbury-Amyot et al.[22] employed G theory to determine how many faculty raters and how many individual components of the scoring rubric for each competency were needed for the reliable scoring of portfolios. Using Kane's (2006) validity argument framework, they evaluated validity evidence based on the extent to which the proposed interpretations and uses of portfolio assessment were plausible and appropriate. One of their primary validity claims asserted that "a primary trait scoring rubric and accompanying traits are relevant for scoring the portfolios" (p. 663). The rubric used in this study included portfolio primary traits (equated to competency), accompanied by trait-level components, which served as the criteria for raters' evaluations. Using a fully crossed two-facet design (portfolio x component x rater), they found only a small

amount of variance was contributed by the components, while rater variability produced the greatest source of error. They surmised that the small proportion of the variance attributable to the components of the scoring rubric implied that raters develop a global impression (the halo effect) that impacted the evaluation of each of the scoring rubric traits, and that this made it impossible to provide a valid appraisal of specific competency components. Given this, any claims based upon analytical scoring (i.e., requiring a separate score for each of the evaluation criteria) would likely be misleading because scores represented a holistic appraisal, rather than an evaluation of each unique component of the traits assessed.

Both examples in this category contribute to a *scoring* validity inference that address whether or not the scores should be combined, and to a *generalization* inference regarding the level of generalizability or reliability displayed by a score (singly and a composite).[6]

2.5. Investigating implications related to diversity

Using established predictive variables with objective algorithmic selection techniques will not typically produce a racially diverse student body.[23] To address this, medical colleges have begun to utilize subjective methods for selection, and the Association of American Medical Colleges (AAMC) has promoted holistic review as a preferred approach for enhancing diversity.[34] It is important to examine how the subjective judgements employed within holistic review may impact selection.[24,25] An important and largely unexamined question relates to how the heterogeneity of those providing the judgements might affect selection diversity. This sort of evidence supports inferences related to *implications* in the framework of Kane's four validity inferences.[26] Such evidence would be useful as it is currently unknown whether enhancing the diversity of decision-makers might also promote the diversity of the student body by capturing a wider spectrum of opinion regarding who should become physicians. In this context, some aspects of rater disagreement might not be considered a source of error, but rather part of true score or universe score variance. For example, Stratton et al.[27] introduced a G theory methodology for measuring the diversity of subjective opinion within admission interviewers at a large medical school. They utilized faculty from two campuses with different educational missions. While this did not specifically introduce racial or ethnic diversity into the interviewer pool, the G study methodology did examine the impact of interviewer diversity. That study used the two campus/institution-types as a fixed facet in a univariate G study and as a two-level variable in a MVG study. The authors were able to gauge the impact of adding interviewer diversity on interview scores and did not treat the differences (disagreements) between raters across campuses as a source of error. Rather, they modeled this disagreement as a source of variance that could potentially enhance validity by increasing the universe of generalization. More importantly, the study demonstrated that MVG was a useful analytic tool for examining validity issues related to the effects of diverse interviewers or admission committee members. This technique could examine the impact of different admission committee demographics (racial or ethnic) that are currently over- or under-represented in the medical student population. These groups would then represent a fixed facet in the G study model, and in this context, variance representing disagreement between rater groups would not be treated as error variance. A measure of the relative importance of using different groups of raters could be derived and G theory could examine how diversity in a holistic review admission committee impacts admissions. It seems reasonable that if holistic review is viewed as important, medical schools might be well-advised to seek out the diverse opinions of their stake holders. Current admission committees are often comprised of medical school faculty that do not reflect the types of diversity targeted by the school. Using G theory to examine the impact of diversifying admission committees and interviewers might provide evidence to support further diversity initiatives.

3. Potential future validity applications of G theory

3.1. Validity evidence based on test content: defining what is being measured

It should be noted that when a G study satisfies the assumption of random sampling from a well-defined and unrestricted universe of admissible observations, it also provides validity evidence based on test content. This is so because random sampling requires carefully defining the observational universe, and in an educational context, often provides a straight-forward approach to establishing evidence based on test content and extrapolation evidence by better defining what is being measured. While working at the National League of Nursing, Kane[28] noted that the assessed

dispositional attributes of an examinee could be defined in terms of the universe of generalization and that an examinee's universe score could define the target construct as the average performance across a very large sample of such observations. In such educational applications, a G coefficient can be interpreted as reflecting both reliability and validity and would support a validity *generalization* and *extrapolation* inference.[6,28] This is especially relevant in achievement testing in health professions education where behaviorally-defined learning objectives (i.e., be able to define, be able to recognize, be able to predict, etc.) can be directly mapped to, and are synonymous with, individual test items. In addition to mapping test items to specific content and subject areas, cognitive levels can also be modeled as a fix facet (e.g., Bloom's taxonomy) and incorporated into G study models.[29]

3.2. Extrapolation: investigating MTMM evidence

Although the techniques and concepts that are part of multitrait-multimethod (MTMM) research have long been considered an important element in providing evidence to construct an *extrapolation* inference, statistical and methodological complexities associated with the analytic techniques leaves them seldom used within health science education.[30] As confirmatory factor analysis (CFA) and other related methods often yield misleading solutions, researchers chose to subjectively evaluate correlational matrices.[31] However, more objective statistical techniques are needed, and G theory may offer a viable alternative. Although G theory-based MTMM research has been successfully used in the assessments of competence in other professions, it has not been applied in health science education.[32] The most obvious MTMM application of G theory is in addressing convergent and divergent validity when different assessment methods or formats are thought to target the same or different abilities. Convergent validity can be supported whenever scores using different measurement procedures (different item formats, different raters, etc.) are relatively invariant, and would be indicated by small variances related to procedure. Discriminant validity would be evident when person-by-procedure error variance is relatively large. This usually requires univariate mixed G study models to examine person-by-format interactions as MTMM convergent/divergent evidence. Large interactions would demonstrate divergence, and the lack of an interaction would signal convergence. Multivariate G studies might also be useful for exploring MTMM questions—including how other newly-introduced assessments formats compare with existing formats. For example, it would be of interest to know how the Multiple Mini-Interview (MMI) compares with the traditional interview.

Also, future research might examine how a Rasch model, various MTMM techniques, and G theory perform in real-world applications and within simulated data with know characteristics.

3.3. Detecting construct irrelevant variance (bias)

CIV can produce an over-or under-estimate of an examination score for an individual. When CIV is group related, test item difficulty varies for specific groups of examinees, and examinees will score relatively higher on one set of test items compared to another set of test items. Although differences in mean scores do not necessarily reflect bias, when means differences between test scores (or some other criterion) are observed for sub-groups (e.g., gender or race), *The Standards for Educational and Psychological Testing* caution that this could reflect a bias, a form of CIV[33] that can impact an *implication* inference. To find out if bias exists, G theory could be employed to model subgroup classification as a facet in a G study model to estimate the variance attributable to this source. For example, if a G study revealed a large interaction for a subgroup facet or subset of items, this would suggest error variance explained by subgroup membership. In this case, if a large proportion of error was found to be attributable to an interaction between subgroups and a subset of test items—it could indicate differences in the meaning of subgroup scores from an item set and indicate a different interpretation of validity for different subgroups. While G theory has been previously used to study group bias in a writing assessment, it has not been employed in assessments used in health science education.[35]

4. Caveats and conclusions

Results from G analyses are only generalizable to the design used (including any limitations related to hidden, nested, and confounded facets) and to the degree to which the *universe of generalization* is truly represented by the sample used for analysis. Consideration of sampling is essential to an accurate interpretation of G study results, yet too often generalizations are not bound by the sample used in the G study model.[36] Generalizations are necessarily limited by constraints in the G study design. Decisions

regarding whether a facet should be treated as random (conditions sampled represent a randomly sampled subset of the universe) versus fixed (conditions selected represent the universe of interest) certainly has implications for reliability estimation, but also for validity. A fixed facet may for example imply that all relevant conditions of interest in a facet are included in the sample.[36] This means that a fixed facet design will generate results that can only be generalized to the specific conditions within the measurement design and is a limiting factor in validity considerations. On the other hand, a random facet represents a sample that should be considered interchangeable with any same-sized, alternative sample representing the same population. Given this interchangeability, results can be generalized to a much larger universe—one that is not bound by the parameters used in the G study. However, when this assumption of interchangeability is not upheld, any subsequent validity estimates are questionable because a *different* or alternative sample may not result in similar findings. With G theory, as with any research, it is inappropriate to generalize beyond the study's parameters.

Ebel[37] asserted, "Validity has long been one of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few" (p. 640). A survey of the health science education assessment literature suggests this is still largely true, and the lack of a consensus in defining validity makes high quality studies rare.[5] While researchers routinely generate reasonably accurate estimates of reliability for the measures they use, validity evidence is often weak or entirely neglected. The dearth of high-quality validity studies almost certainly stems from the lack of robust research designs, unclear definitions and conceptualizations of validity, and impractical statistical methodologies. G theory can help address this problem and provide researchers with a more concise approach to generating meaningful validity evidence. Because G theory can often be applied to *in vivo* data, it provides a viable alternative to experimental designs (*in vitro data)* that are often impractical and/or unethical in health science education.

The examples presented here represent a small subset of potential designs that can provide validity evidence. Because each validity application of G theory has unique design features, further advances within health science education will require research methodologist to further develop innovative research designs, publish their findings, and construct an expert consensus regarding the best use of the theory in validity research. While there is no single, correct way to apply G theory in validity research, guidance can be provided. The presentation and application of the theory will depend upon many factors including data sampling opportunities, the applicable G and D study designs, whether multivariate models can be applied, and familarity with G study methodology. Most importantly, future advances will require that researchers continue to explore and share new applications.

References

1. Brennan RL. Some problems, pitfalls, and paradoxes in educational measurement. *Educ Meas*. 2001;40(4):6–18.
2. Cronbach LJ, Gleser GC, Nanda H, Rajaratname N. *The dependability of behavioral measures*. New York: Wiley; 1972.
3. Cronbach LJ. In: de DNM, ed. New York: Wiley; 1976:199–208. Gruijter, van der Kamp LJT, eds. Advances in psychological and educational measurement.
4. Kane MT. The role of generalizability in validity. In: *Presentation at NCME meeting, 1999 ERIC TM 029 888*. 1999:1–11.
5. St-Onge C, Young M, Eva KW, Hodges B. Validity: one word with a plurality of meanings. *Adv Health Sci Educ*. 2017;22:853–867.
6. Kane M. Validation. In: Brennan R, ed. *Educational measurement*. 4th ed. Westport, CT: American Council on Education and Praeger; 2006:17–64.
7. Kreiter CD, Ferguson K, Lee WC, Brennan RL, Densen P. A generalizability study of a new standardized rating form used to evaluate students' clinical clerkship performances. *Acad Med*. 1998;73(12):1294–1298.
8. Elstein AS, Shulman LS, Sprakfa SA. *Medical problem solving: an analysis of clinical reasoning*. Cambridge, MA: Harvard University Press; 1978.
9. van der Vleuten CPM, Swanson DB. Assessment of clinical skills with standardized patients: the state of the art. *Teach Learn Med*. 1990;2:58–76.
10. Kreiter CD, Bergus GR. Case specificity: empirical phenomenon or measurement artifact. *Teach Learn Med*. 2007;19:378–381.
11. Norman GR, Tugwell P, Feightner JW, Muzzin LJ, Jacoby LL. Knowledge and clinical problem-solving. *Med Educ*. 1985;19:344–356.
12. Shavelson RJ, Baxter GP, Gao X. Sampling variability of performance assessments. *J Educ Meas*. 1993;30:215–232.
13. Jarjoura D, Early L, Androulakakis V. A multivariate generalizability model for clinical skills assessment. *Educ Psychol Meas*. 2004;64:22–39.
14. Schuwirth LWT, van der Vleuten CPM. *How to design a useful test: the principles of assessment, understanding medical education – edited by tim swanwick*. London Deanery London, UK: Wiley-Blackwell; 2010.
15. Vispoel WP, Morris CA, Kilinc M. Practical applications of generalizability theory for designing, evaluating, and improving psychological assessments. *J Pers Assess*. 2018;100(1):53–67.
16. Messick S. The psychology of educational measurement. *J Educ Meas*. 1984;21:215–237.

17. Kreiter CD, James PA, Stansfield RB, Callaway MR. An empirical validity study of a preceptor evaluation instrument. *Acad Med*. 2002;77:S70−S72.
18. Kreiter CD, Wilson AB, Humbert AJ, Wade PA. Examining rater and occasion influences in observational assessments obtained from within the clinical environment. *Med Educ Online*. 2016. https://doi.org/10.3402/meo.v21.29279.
19. Wood TJ, Pugh D, Touchie C, Chan J, Humphrey-Murto S. Can physician examiners overcome their first impression when examining performance changes? *Adv Health Sci Educ*. 2018;23:721−732.
20. Margolis MJ, Clauser BE. A regression-based procedure for automated scoring of a complex medical performance assessment. In: Williamson DM, Bejar II, Mislevy RJ, eds. *Automated scoring of complex tasks in computer-based testing*. London: Lawrence Erlbaum, Associates; 2006:123−167.
21. 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:S80−S82.
22. Gadbury-Amyot CC, McCracken MS, Woldt JL, Brennan RL. Validity and reliability of portfolio assessment of student competence in two dental school populations: a four-year study. *J Dent Educ*. 2014;78:657−667.
23. Kreiter CD. A measurement perspective on affirmative action in U.S. medical education. *Med Educ Online*. 2013;18(1). https://doi.org/10.3402/meo.v18i0.20531.
24. Kreiter CD, O'Shea M, Bruen C, Murphy P, Pawlikowska. A meta-analytic perspective on the valid use of subjective human judgement to make medical school admission decisions. *Med Educ Online*. 2018. https://doi.org/10.1080/meo.v23.1522225.
25.. Monroe A, Quinn E, Samuelson W, Dunleavy DM, Dowd KW. An overview of medical school admission process and use of applicant data in decision making: what has changed since the 1980s? *Acad Med*. 2013;88(5):672−681.
26. Kane MT. Validating the interpretations and uses of test scores. *J Educ Meas*. 2013;50:1−73.
27. Stratton TD, Kreiter CD, Elam CL. Main and regional campus ratings of applicants to a rural physician leadership program: a generalizability analysis. *Journal of Regional Medical Campuses*. 2019;vol. 1(6). https://doi.org/10.24926/jrmc.
28. Kane MT. A sampling model for validity. *Appl Psychol Meas*. 1982;6:125−160.
29. Violato C. *Assessing competence in medicine and other health professions*. BocaRaton: CRC Press; 2018.
30. Campbell DT, Fiske DW. Convergent and discriminant validation by the multitrait- multimethod matrix. *Psychol Bull*. 1959;56:81−105.
31. Lance CE, Woehr DJ, Meade AW. Case study: a Monte Carlo investigation of assessment center construct validity models. *Organ Res Methods*. 2007;10:30−448.
32. Kreiger K, Teachout MS. Generalizability theory as construct-related evidence of the validity of job performance ratings. *Hum Perform*. 1990;3:19−35.
33. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, & Joint Committee on Standards for Educational and Psychological Testing. *Standards for educational and psychological testing*. Washington, DC: AERA; 2014.
34. Association of American Medical Colleges. *Holistic review*. Internet]. Cited; 2019. Available from: https://www.aamc.org/initiatives/holisticreview/.
35. Huang J. Using generalizability theory to examine the accuracy and validity of large-scale ESL writing assessment. *Assess Writ*. 2012;17:123−139.
36. Shavelson RJ, Webb NM. Measurement methods for the social sciences series. In: *Generalizability theory: a primer*. vol. 1. Thousand Oaks, CA, US: Sage Publications; 1991.
37. Ebel R. Must all tests be valid? *Am Psychol*. 1961;16:640−647.