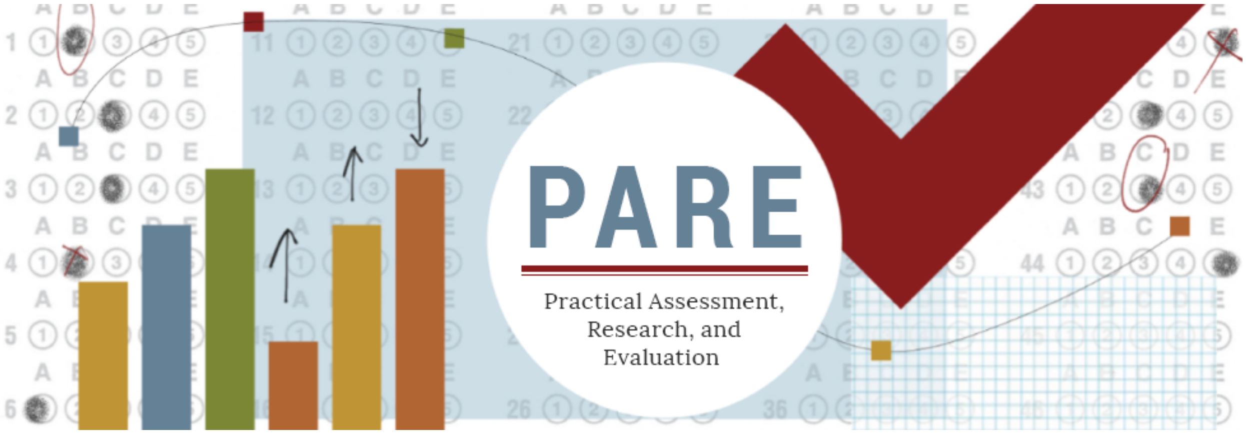


A peer reviewed, open-access electronic journal: ISSN 1531-7714

Mixed Model Generalizability Theory: A Case Study and Tutorial

Alan Huebner, *University of Notre Dame*
Gustaf B. Skar, *Norwegian University of Science and Technology*
Mengchen Huang, *Insights @ Riot Games*

Abstract: Generalizability theory is a modern and powerful framework for conducting reliability analyses. It is flexible to accommodate both random and fixed facets. However, there has been a relative scarcity in the practical literature on how to handle the fixed facet case. This article aims to provide practitioners a conceptual understanding and computational resources to deal with designs with a fixed facet in both univariate and multivariate generalizability theory settings. The analyses feature a real data set, which is available to readers along with the code to reproduce all analyses.

Keywords: reliability, generalizability theory, writing assessment

Introduction

G-theory is a powerful, modern framework for the decomposition of variance and variance components. The main purposes of the method are to (1) disentangle the error introduced by multiple sources of variation (referred to as "facets" in G-theory) and (2) aid in the selection of optimal measurement procedures. These steps are referred to as G and D studies, respectively. G-theory accommodates various study designs, for example in its treatment of random versus fixed facets: if items from an included facet can be thought of as representative for a larger pool of items (a random facet) or exhausting the pool of items (a fixed facet). This distinction is important because it helps the researcher understand to what extent a students' performance is generalizable to a domain of knowledge (we discuss this below in terms of "Universe of Generalization" and "Universe of Admissible Observations"); if a facet is fixed and the reliability acceptable it is more likely that a score represents domain knowledge than if a facet is random simply because of sampling errors. The distinction is also important statistically; fixed and random facets need different treatment in statistical analyses for the results to be valid. In addition, fixed facets can be handled using either univariate or multivariate G-theory (MG-theory) models, which will be explained below. The variety of approaches within the G-theory framework is a testament to the power and flexibility of the method; however, the complexities of choosing and performing the procedures can present a formidable challenge to novice users. Thus, this tutorial aims to guide practitioners in choosing and implementing the appropriate methods.

There are currently several primary authoritative resources on G-theory, including: the seminal paper by Gleser et. al. (1965), Shavelson and Webb (1991) who offer a compact introduction to G-theory, and Brennan (2001a) who provides a more mathematically detailed account that includes several advanced topics. There have also been other guides for implementing G-theory geared toward practitioners, e.g., Webb et al. (2006). However, less attention has been given to special issues, including the issue of fixed versus random facets in G-theory, especially in practical applications of the method. In particular, Jiang (2018) and Huebner and Lucht (2019) illustrate performing computations for G and D studies using the statistical software environment R (R Core Team, 2020), but both papers only consider designs in which all facets are random. Vispoel et al. (2018) include some mention of fixed facets, but their focus is mainly on connections between G-theory, structural equation modeling, and classical test theory. Furthermore, these contributions have focused on univariate G-theory exclusively, when, in fact, MG-theory can also be used to handle a fixed facet. This tutorial aims to fill this gap using real data from writing assessment.

The current paper uses a real data set to provide a practical tutorial for conducting a G-theory analysis with a fixed facet using both univariate and MG-theory. Specifically, we present results from (1) a univariate analysis with all facets considered as random, (2) a univariate analysis with one random facet and one fixed facet, and (3) a MG-theory model. We also discuss the practical implications of the approaches, including the real-world interpretations in the context of the current study. At least one previous study has compared results from different designs for univariate and multivariate G-theory (Keller et. al., 2010), but the current paper aims to aid practitioners in implementing the methods for their own studies. Specifically, the current tutorial aims to make the following contributions:

- Outside of the texts mentioned above, there are very few instructional resources available that focus on random vs. fixed facets using a real data case study.
- Furthermore, there are few instructional resources available for multivariate G-theory using real data.
- The data for the current study is relatively large and available freely on GitHub, along with the code for the analyses.

Moreover, the application of G-theory presented here targets an important, but to the best of our knowledge, under-researched aspect of writing assessment: namely, the consequences of assuming that rating scales either exhaust ways of judging qualities of a text, or that rating scales in fact are random, as it were, items from a larger pool of possible items.

This tutorial is organized as follows. In the next section, data gathered for a writing assessment application is described. Then, the fundamentals of G-theory are reviewed, including basic terminology and notation, crossed versus nested designs, and fixed vs. random facets. Then, the data is analyzed three ways: (i) a univariate model with both facets random, (ii) a univariate model with one facet regarded as fixed, and (iii) a multivariate model. It is suggested that readers should have a basic understanding of G and D studies at the level of the presentation of Shavelson and Webb (1991) to take full advantage of the subsequent tutorial. The data as well as code, including source files for the estimation algorithms and scripts to run the analyses, are posted on GitHub: https://github.com/alanhuebner10/Gtheory-codes-PARE2025

Data Set: Ratings of Text Quality

Data for this tutorial originate from a large-scale writing intervention project (Skar, Aasen, et al., 2020). Elementary grade students were administered “discursive writing tasks,” which prompted students to write, for example, a letter to researchers about favorite activities during recess time, or to write a piece detailing what the student would do if s/he was the prime minister of Norway. The ensuing student texts were collected by researchers and subjected to ratings of text quality. These ratings were performed by trained raters using validated rating scales (Skar, Jølle, et al., 2020). The raters were university staff and graduate students at the university of the second author and were trained by the second author in sessions that included supervised ratings and rater discussions. In the operational rating, a student's final score on each scale was the average score from the two raters.

The rating scales were designed to capture the most important aspects of text quality. These were the eight rating scales: audience awareness, organization, content relevance, vocabulary, sentence construction, spelling, legibility, and punctuation. Each rating scale had descriptions of quality for five levels. The data set used for this tutorial consists of 6,704 ratings (2 raters × 8 rating scales × 419 texts). As mentioned above, the same two raters assessed all eight scales for all texts. Thus, the design was fully crossed.

Understanding the Data in a G-theory Context

Before reviewing the fundamentals of G-theory, we briefly shed light on the data set using three key concepts in G-theory: the universe of generalization (UG), universe score (US), and the universe of admissible observations (UA). In the words Kane et al. (1999, p. 8) UG is "the subdomain for which it is plausible to consider the observed performances to be a random or representative sample" and US is "an individual's expected score over the universe of generalization." The UA is "the set of all possible observations involving combinations of conditions of the various facets in the G study" (Kane, 2002, p. 166), with *admissible* denoting that "the observations […] are eligible for inclusion in the G study and in future D study" (Kane, 2002, p. 166).

In our case, a student's text quality score would be his/her US (i.e., his/her expected score over all possible assessment instances) in the UG of discursive writing in school. A key question would be to what extent the *observed score* (OS) would be representative for the US. The association between OS and US would weaken if the writing tasks, the raters, or the rating scales are not randomly sampled, or at least representative (cf., Kane et al., 1999). Likewise, the association would strengthen if the OS would be based on a complete measurement, i.e., one that incorporated all conditions of the UA.

When conditions of measurement are exhausted, a G-theory facet (e.g., raters, occasions) is fixed (Shavelson & Webb, 1991, p. 65), which will yield higher reliability estimates than treating a facet as random. Determining the UA is, however, difficult (see Kane, 2002, for a general discussion). The rating scales outlined above were designed to yield a comprehensive score of text quality but employing a different theoretical lens might have yielded very different rating scales. As a concrete example, consider writing assessment within the context of mother tongue (MT) education versus writing assessment in the context of second language acquisition (SLA). It is quite possible that rating scales in SLA would include very nuanced descriptions of grammar and other mechanics to a greater extent than in the former, while this could easily be glossed over in the context of MT education. If MT and SLA would be considered as distinct and separate universes, rating scales probably would be considered as fixed facets, but if the UG would be writing in general (a plausible idea, since students would use writing regardless if learning to write as part of the MT education or SLA), then each set of rating scales would perhaps more accurately be considered to be a random sample of possible rating scales. There are previous examples of treating the rating scale facet as fixed citing practical reasons: "replication of the measurement must contain the same [rating scale]" (Otha et al., 2018) and theoretical: "One could argue that there are a limited number of discernible aspects that can be rated, which would imply that trait is a fixed facet" (Schoonen, 2005). In many instances, though, a researcher may prefer to treat rating scales (and writing tasks) as random because fixed models "do not have the aim of generalizing beyond the condition of each facet" (In'nami & Koizumi, 2016), which limits the application of the results, making claims about students' writing proficiency limited to the conditions of the measurement.

Generalizability Theory

As mentioned above, G-theory refers to sources of measurement error, or variation, as facets (Shavelson & Webb, 1991). In G-theory, an experiment may have any number of facets, but in practice, there are usually at most three or four facets. A common example might involve a sample of students being administered an essay test that is rated by several raters on two different occasions. In this case, the raters and occasions are the facets in the experiment, and the specific raters and occasions are the conditions of each facet. If all raters rate all the essays on all occasions, then the facets are crossed. However, if each occasion uses a different set of raters, then raters are nested within occasions. (We provide an Appendix discussing how the current data may be handled if it was nested; the reader may also see Shavelson & Webb [1991] for additional examples of crossed and nested designs).

In the current application, the raters and rating scales are both facets, as they are both potential sources of variation. Here, all raters rated the student text on all the rating scales. Thus, for this design, the rater facet is crossed with rating scales. Following the conventions of Brennan (2001a), mathematical notation to summarize the design is as follows. The students, or people, are denoted as p, and the raters and rating scales facets are denoted as r and s, respectively. Then, the sample sizes are denoted as $n_p = 419$, $n_r = 2$ and $n_s = 8$, and the design is notated as $p \times r \times s$.

G-theory is essentially a linear random effects model, and thus at least one facet must be random (Shavelson & Webb, 1991). For the current data example, if the raters and rating scales are drawn from very large sets of possible raters and rating scales, they are referred to as random facets, and the resulting G-theory model is referred to as a random model. In other words, the G-theory results are applicable to all possible raters and rating scales. However, if we only wish to generalize to the raters (or rating scales) used in the study, then raters (or rating scales) are said to be a fixed fact. Then, the resulting G-theory model is referred to as a mixed model.

Univariate G study

The G study phase estimates variance components for main effects and interactions. For example, the variance component for persons is denoted as $\sigma^2(p)$, the variance component for the interaction between persons and raters is denoted as $\sigma^2(pr)$, and so on. Under the two-facet crossed design, there are seven total variance estimable variance components, $\sigma^2(p)$, $\sigma^2(r)$, $\sigma^2(s)$, $\sigma^2(pr)$, $\sigma^2(ps)$, $\sigma^2(rs)$, and $\sigma^2(prs)$, where the last component is confounded with the residual (Shavelson & Webb, 1991). The variance components are derived from the traditional analysis of variance (ANOVA) estimates via the expected mean square (EMS) procedure. Chapter 3 of Brennan (2001a) gives a general treatment of this procedure as well as examples for several designs. Applied practitioners can obtain the ANOVA estimates and G-theory variance components in R for several designs by following the examples illustrated in Huebner and Lucht (2019).

Univariate D study

The D study phase provides coefficients to summarize reliability under the actual sample sizes as well as other potential sample sizes. Thus, in the context of the current application, G-theory provides projected reliability estimates that would result under different numbers of raters or rating scales used. The sample sizes for the D study are notated with a "$'$" in the superscript, indicating that the D study sample size is not necessarily the same as the actual sample size. For the current application, the D study sample sizes for raters and rating scales in the analysis below are equal to the actual sample sizes, which are notated as $n'_r = n_r = 2$ and $n'_s = n_s = 8$, respectively.

There are different reliability estimates that can be constructed from the results within G-theory, but the two main coefficients are the generalizability coefficient $E\rho^2$ and the index of dependability Φ. These are also referred to as coefficients for relative and absolute decisions, respectively (Shavelson & Webb, 1991). In the context of the current study, this means that $E\rho^2$ is appropriate if the practitioner is interested in merely ranking students on their writing scores, while Φ would be used if there were interest in the actual score achieved by an individual. The latter would most likely be true because of the researcher's interest to make inferences about the performance level of students at different developmental stages. Thus, in the following presentation only Φ is considered, but the general concepts apply to $E\rho^2$ as well.

The index of dependability is given by

$$\Phi = \frac{\sigma^2(\tau)}{\sigma^2(\tau) + \sigma^2(\Delta)} \qquad (1)$$

where $\sigma^2(\tau)$ is the universe score variance and $\sigma^2(\Delta)$ is the absolute error variance. The quantity $\sigma^2(\tau)$ is the variance of the mean score of every instance of measurement in the universe of generalization, while $\sigma^2(\Delta)$ is the error variance when generalizing from subjects' observed mean scores to the universe scores (Brennan 2001a; Shavelson & Webb, 1991). It will be seen below that for a two-facet design, $\sigma^2(\Delta)$ is smaller when one facet is regarded as fixed than when both facets are random.

Univariate Random vs. Mixed Models

The calculation of the universe score variance and absolute error variance, $\sigma^2(\tau)$ and $\sigma^2(\Delta)$, respectively, differ under random and mixed G-theory models, and understanding these differences reveals the meaning of random versus fixed facets. In the following, we use Brennan's (2001a) notational convention of denoting facets with uppercase letters for the D study phase, and as previously mentioned, the D study sample sizes were $n'_r = n_r = 2$ and $n'_s = n_s = 8$. For the $p \times R \times S$ design in which both raters and rating scales are random, the universe score variance is simply $\sigma^2(\tau) = \sigma^2(p)$. Furthermore, for this design the absolute error variance is given by

$$\begin{aligned}\sigma^2(\Delta) &= \sigma^2(R) + \sigma^2(S) + \sigma^2(pR) + \sigma^2(pS) + \sigma^2(RS) + \sigma^2(pRS) \\ &= \frac{\sigma^2(r)}{n'_r} + \frac{\sigma^2(s)}{n'_s} + \frac{\sigma^2(pr)}{n'_r} + \frac{\sigma^2(ps)}{n'_s} + \frac{\sigma^2(rs)}{n'_{rs}} + \frac{\sigma^2(prs)}{n'_{rs}} \qquad (2)\end{aligned}$$

(Brennan, 2001a).

On the other hand, for the $p \times R \times S$ design with S fixed, the formulas differ. Section 4.3.1 of Brennan (2001a) provides rules for obtaining formulas for $\sigma^2(\tau)$ and $\sigma^2(\Delta)$ in mixed models. Using Rule 1, the universe score variance is

$$\sigma^2(\tau) = \sigma^2(p) + \sigma^2(pS) = \sigma^2(p) + \frac{\sigma^2(ps)}{n'_s} \qquad (3)$$

When the S facet is fixed, the variance component for the pS interaction is included. This is because in this case the universe score is based on a student's average score over the finite levels of S (i.e., the rating scales). Thus, the universe score variance $\sigma^2(\tau)$ is the variance over students' mean scores for *just those rating scales.* Any student effects specific to those particular levels therefore become part of $\sigma^2(\tau)$ (Webb et al., 2006). Also, when S is fixed, the $\sigma^2(S)$ and $\sigma^2(pS)$ terms are removed from the calculation of $\sigma^2(\Delta)$, i.e., they are omitted from Equation 2 (though $\sigma^2(pS)$ is moved to $\sigma^2(\tau)$ as was explained above). This is because when s is fixed, the variation due to rating scales and the pS interaction do not count toward error, since we are not generalizing beyond the rating scales in the study. The absolute error variance $\sigma^2(\Delta)$ is larger for the random model than for the mixed, as more error results when generalizing to a large set of rating scales. Thus, when s is fixed, there is no sampling variability for the rating scales -- every instance of the measurement procedure would involve the same rating scales. In other words, if we were to replicate the study, we would use different raters but the *same* rating scales (Brennan, 2001a). Similarly, the dependability coefficient $\widehat{\Phi}$ is larger for the mixed model, as the universe of generalization only consists of the rating scales in the study; i.e., the universe of generalization with a fixed facet is narrower than with both facets random (Brennan, 2001a).

Multivariate G-theory

The univariate G-theory design used above, $p \times r \times s$, may be considered as a multivariate G-theory model by fixing the s facet and estimating a $p \times r$ design for each level of s. This MG-theory design is notated as $p^\bullet \times r^\bullet$, where the filled circle in the superscripts indicate that subjects and raters are crossed with rating scales (Brennan, 2001a). Since there are eight rating scales for the current application, there are eight $p \times r$ designs, each with its own set of variance components for p, r, and the pr residual. The person, rater, and residual variance components for rating scale 1 are notated as, respectively, $\sigma_1^2(p)$, $\sigma_1^2(r)$, and $\sigma_1^2(pr)$; the variance components for rating scale 2 are notated as $\sigma_2^2(p)$, $\sigma_2^2(r)$, and $\sigma_2^2(pr)$, and so on for the rest of the eight rating scales. Furthermore, the multivariate design allows for the estimation of covariance components for p, r , and pr. The covariance component for persons for rating scales 1 and 2 is notated as $\sigma_{12}(p)$, and so on. The variance-covariance matrices for p, r, and pr are notated as $\mathbf{\Sigma}_p$, $\mathbf{\Sigma}_r$, and $\mathbf{\Sigma}_{pr}$, respectively. These matrices are symmetric, with the variances on the diagonals and the covariances on the off-diagonals.

The covariance components provide a measure of the linear association between persons, raters, and residuals, but it is often more intuitive to deal with correlations. One advantage of MG-theory is that it provides the "disattenuated" correlations of the rating scales. While Pearson coefficients can be attenuated toward zero due to measurement error, disattenuated correlations describe the relation between the variables as if the reliability was perfect (Institute for Objective Measurement, 2024). Brennan (2001a) states the estimated disattenuated correlation between universe scores for persons on rating scales v and v' is as

$$\hat{\rho}_{vv'}(p) = \frac{\hat{\sigma}_{vv'}(p)}{\sqrt{\hat{\sigma}_v^2(p)\hat{\sigma}_{v'}^2(p)}}$$

and similar for $\hat{\rho}_{vv'}(i)$ and $\hat{\rho}_{vv'}(pr)$. If these correlations are very high, a researcher may wish to investigate if they are measuring very similar aspects of the construct.

The D study uses the $\mathbf{\Sigma}_p$, $\mathbf{\Sigma}_r$, and $\mathbf{\Sigma}_{pr}$ matrices obtained in the G study to assess the reliability of the scores using the actual number of raters $n_r = 2$ as well as the reliability under other potential sample sizes n'_r. The variance covariance matrix for universe scores, notated as $\mathbf{\Sigma}_\tau$, is simply equal to $\mathbf{\Sigma}_p$ for the current design. Then, the variance-covariance matrix for the absolute error is given by

$$\mathbf{\Sigma}_\Delta = \frac{\mathbf{\Sigma}_r}{n'_r} + \frac{\mathbf{\Sigma}_{pr}}{n'_r} \qquad (4)$$

where the notation $\frac{\mathbf{\Sigma}_r}{n'_r}$ means that every element of the $\mathbf{\Sigma}_r$ matrix is divided by n'_r.

Furthermore, as stated above, a multivariate D study allows for levels of the fixed facet to be weighted differently, and these weights factor into the computation of the quantities $\sigma_C^2(\tau)$ and $\sigma_C^2(\Delta)$, where the "C" subscript is for the "composite" universe score and composite absolute error variance, respectively. Weighting may be useful in situations where many related aspects are rated and where professionals and other stakeholders do not perceive all aspects to be equally important. For example, in a situation where students are to write a letter to the principal, stakeholder may assess a students' ability to use an appropriate voice as more important than the students' ability to correctly use commas. For $n_v = 8$ rating scales, there are the same number of weights, notated as $w_1, w_2, \ldots, w_8$. The weights are assigned such that they convey the relative importance of each level and sum to one. For example, suppose for the current application rating scales 7 and 8 are deemed half as important as the other rating scales. Then, rating scales one through six would be assigned weights equal to $1/7$, while rating scales 7 and 8 would be assigned weights of $1/14$. It is easily checked that $6(1/7) + 2(1/14) = 1$.

To compute the $\sigma_C^2(\tau)$, the elements of $\mathbf{\Sigma}_p$ are multiplied by the corresponding weights. For example, the first variance component on the diagonal, $\sigma_1^2(p)$, is multiplied by w_1^2, the covariance component $\sigma_{12}(p)$ for persons for rating scales 1 and 2 is multiplied by $w_1 w_2$, and so on. Then, the weights are used in a similar fashion to with $\mathbf{\Sigma}_\Delta$ to obtain $\sigma_C^2(\Delta)$. Finally, the multivariate index of dependability is given by

$$\Phi = \frac{\sigma_C^2(\tau)}{\sigma_C^2(\tau) + \sigma_C^2(\Delta)} \qquad (5)$$

Example Analysis

Univariate G Study Results

The G study is conducted treating all facets as random, regardless of whether some facets will be considered fixed subsequently for the D study (Shavelson and Webb, 1991; Brennan, 2001a). As seen above, the differences between fixed and random facets become relevant when computing the D study quantities such as the universe score variance and absolute error variance. The analysis of variance (ANOVA) results and variance components for the G study are shown in Table 1. The ANOVA was produced using the `aov()` function in R, and the variance components were obtained using the `gstudy()` function in the `gtheory` R package (Moore, 2016). Huebner and Lucht (2019) explain performing both procedures using R. Referring to the first row of Table 1, the quantity $\sigma^2(p)$ is the variance component for persons, which is the amount of systematic variability between students in their text quality scores. The estimated variance component is $\hat{\sigma}^2(p) = 0.56$. Since it is difficult to judge the absolute magnitude of variance components, results are usually reported as a percentage of the total variance. Thus, variation among persons accounts for about 51% of the total variation, meaning that are substantial differences between subjects in their text quality scores. Table 2 provides interpretations for all variance components. In summary, the largest variance sources were persons and the persons-rating scale interaction, whereas the lowest variances were due to raters and the person-rater interaction.

Univariate D Study Results

We compute D-study results for both a random and mixed model. First, for the random model in which both raters and rating scale are random, the estimated universe score variance is simply $\hat{\sigma}^2(\tau) = \hat{\sigma}^2(p) = 0.56$. Then, the absolute error variance is obtained by substituting the values in Table 2 into Equation 2:

$$\hat{\sigma}^2(\Delta) = \frac{0.0}{2} + \frac{0.15}{8} + \frac{0.03}{2} + \frac{0.16}{8} + \frac{0.04}{2*8} + \frac{0.15}{2*8} = 0.07$$

Finally, substituting $\hat{\sigma}^2(\tau)$ and $\hat{\sigma}^2(\Delta)$ into Equation 1, the index of dependability is calculated to be

$$\hat{\Phi} = \frac{0.56}{0.56+0.07} = 0.89,$$

which would be considered to be a good value given the complex interaction between a rater interpreting the student texts in light of her understanding of both tasks and rating scales (McNamara, 1996).

Table 1. G study results. Both facets are treated as random.

Source	DF	SS	MS	Variance	Percent of Total
p	418	4020	9.62	0.56	51.3
r	1	16	16.34	0.00	0.0
s	7	976	139.44	0.15	13.5
pr	418	173	0.41	0.03	3.0
ps	2926	1362	0.47	0.16	14.4
rs	7	120	17.15	0.04	3.7
prs	2926	450	0.15	0.15	14.2

Note: DF=degrees of freedom, SS=sum of squares, MS=mean square.

On the other hand, as mentioned above, the formulas differ when S is fixed. In this case, the universe score variance is obtained using Equation 3:

$$\hat{\sigma}^2(\tau) = 0.56 + \frac{0.16}{8} = 0.58.$$

As explained above, $\hat{\sigma}^2(\tau)$ is larger for the mixed model than for the random model, because the mixed model includes the $\sigma^2(pS)$ term in the calculation of $\sigma^2(\tau)$. Then, for the mixed model case $\hat{\sigma}^2(\Delta)$ is calculated by omitting the $\sigma^2(S)$ and $\sigma^2(pS)$ terms, resulting in

$$\hat{\sigma}^2(\Delta) = \frac{0.0}{2} + \frac{0.03}{2} + \frac{0.04}{2*8} + \frac{0.15}{2*8} = 0.03.$$

(This value is slightly smaller than $\hat{\sigma}^2(\Delta)$ for the random model but the same to two decimals). Thus, the index of dependability for the mixed model is given by

$$\hat{\Phi} = \frac{0.58}{0.58 + 0.03} = 0.95.$$

Table 3 provides interpretations of the D study quantities, which are for the sample sizes $n_r' = 2$ and $n_s' = 8$, as stated above. However, for a full D study, a practitioner would likely be interested in comparing the reliability resulting from different values are combinations of n_r' and n_s' (e.g., what is the reliability resulting from increasing the number of raters or decreasing number or rating scales). These results are obtained by simply plugging in different sample size values for n_s' and n_t' in the formulas for $\hat{\sigma}^2(\tau)$ and $\hat{\sigma}^2(\Delta)$ above. Figure 1 displays the dependability coefficient values for one to five raters and four, six and eight rating scales, for both the mixed and random G-theory models.

Table 2. Interpretation for G study variance components in the context of the current study.

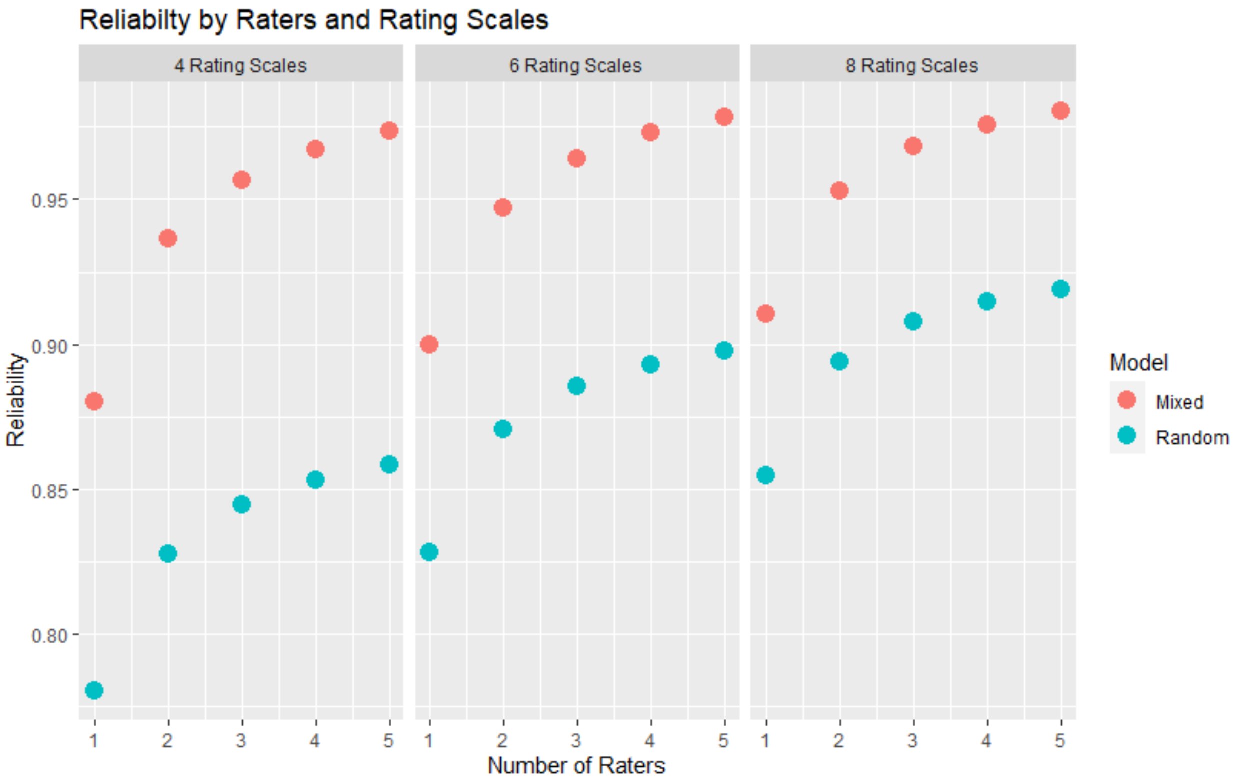
Effect	%	Interpretation
$\hat{\sigma}^2(p)$	51.3%	The variance component for persons quantifies the amount of systematic variability between subjects in their writing scores. The large percentage of total variance indicates substantial differences between students.
$\hat{\sigma}^2(r)$	0.0%	The variance component for raters is virtually zero, indicating the two raters showed very little difference in their measurement standard, averaging over subjects and rating scales.
$\hat{\sigma}^2(s)$	13.5%	The variance component for rating scales is small yet nontrivial, signifying that the scores vary somewhat from scale to scale. In general, some scale scores tend to be higher than others.
$\hat{\sigma}^2(pr)$	3.0%	The variance component for the interaction between persons and raters is relatively small but nonzero, suggesting that the relative standing of subjects differed from one rater to the other to a small degree.
$\hat{\sigma}^2(ps)$	14.4%	The variance component for the interaction between person and rating scale means that the relative standing of subjects differed from one rating scale to another to a considerable degree.
$\hat{\sigma}^2(rs)$	3.7%	The variance component for the rater-rating scale interaction shows that the scores assigned by raters differed over rating scale to a small degree.
$\hat{\sigma}^2(prs)$	14.2%	The variance component for the residual shows that a considerable proportion of the total variance was due to the interaction between subjects, raters, and rating scales and/or unsystematic or systematic sources of variation that were not measured in this study.

Table 3. Interpretation of D study quantities in the context of the current force plate study for $n'_r = 2$ and $n'_s = 2$.

Quantity	Random	Mixed	Interpretation
$\hat{\sigma}^2(\tau)$	0.56	0.58	The universe score variance is the variance of the mean of all raters on rating scales in the universe of generalization.
$\hat{\sigma}^2(\Delta)$	0.07	0.03	In the random design, the absolute error variance includes all variance component in the entire design except universe-score variance $\sigma^2(p)$.
$\hat{\Phi}$	0.89	0.95	The dependability coefficient is the proportion of total variance that is accounted for by universe-score variance.

Note: The Random and Mixed columns refer to the results when considering rating scales (S) as a random versus fixed facet, respectively.

Figure 1. Dependability coefficient values for one to five raters and four, six and eight rating scales, for both the mixed and random G-theory models.



Multivariate Results

R functions were written by the first author to obtain the multivariate G study variance component estimates based on the derivations in section 9.4 of Brennan (2001a). The script in the GitHub repository performs the estimation and uses the G study results to obtain the D study conclusions.

G study. Tables 4, 5, and 6 display the variances, covariances, and correlations for p, r, and pr, respectively, i.e., the $\mathbf{\Sigma}_p$, $\mathbf{\Sigma}_r$, and $\mathbf{\Sigma}_{pr}$, respectively. Variances are in bold font along the diagonals, covariance terms are in the lower triangle, and correlations are italicized in the upper triangle. Table 4 shows that rating scale 5 has the largest persons variance component, $\hat{\sigma}_5^2(p) = 0.96$, while rating scale 7 has the smallest variance component, $\hat{\sigma}_7^2(p) = 0.42$. This means that subjects performed most variably for rating scale 5, while they performed most similarly for rating scale 7. Many of the disattenuated correlations are very high. For example, the disattenuated correlation for universe scores for persons on rating scales 1 and 4 is 0.95; this is obtained by computing

$$\hat{\rho}_{14}(p) = \frac{\hat{\sigma}_{14}(p)}{\sqrt{\hat{\sigma}_1^2(p)\hat{\sigma}_4^2(p)}} = \frac{0.75}{\sqrt{.86 * .72}} \approx .95$$

where $\hat{\sigma}_{14}(p) = 0.75$ is taken from the fourth row, first column of the table and 0.86 and 0.72 are the first and fourth diagonals, respectively. The disattenuated correlation between rating scales 2 and 5 is estimated as 1.00, which means that the subjects were rank-ordered in the same way by the two scales.

Table 4. Variance/covariance/correlation matrix for persons. (Variances are in bold font along the diagonal, covariance terms are in the lower triangle, and correlations are italicized in the upper triangle)

Rating Scale	1	2	3	4	5	6	7	8
1	**0.86**	*0.98*	*0.85*	*0.95*	*0.93*	*0.58*	*0.91*	*0.76*
2	0.74	**0.66**	*0.89*	*0.99*	*1.00*	*0.60*	*0.91*	*0.73*
3	0.69	0.63	**0.77**	*0.95*	*0.89*	*0.59*	*0.73*	*0.59*
4	0.75	0.69	0.71	**0.72**	*0.92*	*0.62*	*0.88*	*0.74*
5	0.84	0.81	0.76	0.76	**0.96**	*0.57*	*0.83*	*0.68*
6	0.48	0.43	0.46	0.47	0.50	**0.79**	*0.58*	*0.65*
7	0.54	0.48	0.41	0.48	0.53	0.34	**0.42**	*0.88*
8	0.51	0.43	0.37	0.45	0.47	0.41	0.41	**0.51**

Table 5 displays the variances and covariances due to the raters. The variances in the diagonals are all close to zero; this means that for all rating scales, there is very little variation due to raters. Rating scale 4 has the largest variation due to raters, $\hat{\sigma}_4^2(r) = 0.04$, but this is still relatively very small. Since there is very little variation, there is also very little covariation, as seen by the off-diagonal covariance values. This pattern of observing very small variances and covariances for raters is also seen in the MG-theory analysis presented by Keller et. al. (2010).

D study. The MG-theory D study uses the weights described above-- scales 7 and 8 are weighted half as important as the other rating scales, resulting in weight values 1/7 for scales 1 through 6 and weight values (1/14) for scales 7 and 8. We wish to compute the multivariate coefficient Φ shown in Equation 5; we begin by computing the matrix $\boldsymbol{\Sigma}_\Delta$ shown in Equation 4, which is obtained by dividing the matrices $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_{pr}$ by n_r' . Next, we need to compute the composite absolute error term $\sigma_C^2(\Delta)$, which is the sum of the products of the weights and elements of $\boldsymbol{\Sigma}_\Delta$. To illustrate, we will demonstrate the computation using $n_r' = 4$, rather than the actual number of raters. The elements in the first row/column (i.e., the [1, 1] elements) of $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_{pr}$ are 0.01 and 0.19, respectively, as shown in Tables 5 and 6. Dividing by $n_r' = 4$ yields 0.0025 and 0.0475, and then the sum is 0.05. Then, since this is the [1, 1] element, the value gets multiplied by (1/7)*(1/7). As another example, the [8, 7] elements of $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_{pr}$ are 0.04 and 0.06, respectively. Dividing these values by $n_r' = 4$ yields 0.01 and 0.015, and the sum is 0.025. This value gets multiplied by (1/14)*(1/14), the weights corresponding to rating scales 7 and 8. This process continues for all elements of the matrix, and then the values are summed to obtain $\sigma_C^2(\Delta)$. The quantity $\sigma_C^2(\tau)$ is obtained by applying the weights to the matrix $\boldsymbol{\Sigma}_p$ in a similar fashion, and then the values of $\sigma_C^2(\tau)$ and $\sigma_C^2(\Delta)$ are plugged into Equation 5.

Using the R script and functions in the GitHub repository to perform the computations, the dependability values for number of raters $n_r' = 1, 2, 3, 4$ yields $\widehat{\Phi} = 0.92, 0.96, 0.97$ and 0.98, respectively. In this case, the rater variability is very small, and thus there is very high reliability even with only $n_r' = 1$ rater. While adding raters does increase reliability, there may be little to gain in this instance.

Table 5. Variance/covariance matrix for raters. (Variances are in bold font along the diagonal and covariance terms are in the lower triangle)

Rating Scale	1	2	3	4	5	6	7	8
1	**0.01**							
2	-0.01	**0.02**						
3	-0.02	0.03	**0.04**					
4	0.01	-0.01	-0.02	**0.01**				
5	0.01	-0.02	-0.04	0.02	**0.03**			
6	0.00	0.01	0.01	0.00	-0.01	**0.00**		
7	-0.01	0.01	0.02	-0.01	-0.02	0.00	**0.01**	
8	-0.04	0.06	0.09	-0.05	-0.08	0.02	0.04	**0.21**

Table 6 displays the variances (diagonal), covariances (lower triangle), and correlations (upper triangle) for the errors, or residuals, which are the random errors plus the interactions between students and raters. For a given rating scale, the residual variance in bold was generally larger than the rater variance and smaller than the person variance. The residual variances ranged from $\hat{\sigma}_6^2(r) = 0.16$ to $\hat{\sigma}_2^2(r) = 0.21$.

Table 6. Variance/covariance/correlation matrix for residuals. Variances are in bold font along the diagonal, covariance terms are in the lower triangle, and correlations are italicized in the upper triangle.

Rating Scale	1	2	3	4	5	6	7	8
1	**0.19**	*0.30*	*0.22*	*0.32*	*0.25*	*0.11*	*0.27*	*0.16*
2	0.06	**0.21**	*0.26*	*0.15*	*0.13*	*0.09*	*0.21*	*0.21*
3	0.04	0.05	**0.17**	*0.14*	*0.03*	*0.11*	*0.19*	*0.12*
4	0.06	0.03	0.03	**0.19**	*0.24*	*0.07*	*0.25*	*0.03*
5	0.05	0.03	0.01	0.05	**0.21**	*0.02*	*0.23*	*0.13*
6	0.02	0.02	0.02	0.01	0.00	**0.16**	*0.22*	*0.11*
7	0.05	0.04	0.03	0.05	0.04	0.04	**0.18**	*0.32*
8	0.03	0.04	0.02	0.00	0.03	0.02	0.06	**0.19**

Discussion

Brennan (2010) states “any univariate mixed model can always be reformulated as a multivariate model”. There are several advantages to using a MG-theory design rather than its univariate counterpart. These advantages include, but are not limited to, the following:

- For multivariate G studies, each level of the fixed facet receives its own set of variance components as well as covariances between the variance components.

- Similarly, multivariate D studies yield absolute error variances for each level of the fixed facet as well as covariances and correlations between error variances for different levels.
- Multivariate D studies allow for the computation of composite reliability indices, which in the current context is the combined reliability over all eight $p \times r$ studies. Moreover, different rating scales can be given different weights in the computation of the composite reliability, which is not available in the univariate counterpart.

MG-theory is a powerful, flexible modeling framework, but these advantages come at the cost of complex computations and interpretations (Brennan, 2001a).

G-theory has been applied to numerous fields, including educational psychology, medicine, and business. However, random G-theory models generally receive more attention than mixed models, and thus there is a lack of practical guides for performing G and D studies for univariate mixed models and well as for multivariate models. This tutorial has analyzed a real data set under both random and mixed G-theory models and has aimed to shed light on conceptual difference between the two. From a substantive standpoint, we have shown how treating rating scales as fixed versus random facets may affect the reliability of the writing scores. Treating the ratings scales as a fixed factor rather than random, i.e., assuming we have exhausted all possible rating dimensions, resulted in an increase in reliability from 0.89 to 0.95. In other words, the error for the fixed case is smaller because the variation due to rating scales and the pS interaction do not count toward error. We emphasize that this is didactic in nature—we are not suggesting that the facets should be considered as fixed solely to increase reliability. It is hoped that this paper and the accompanying computational resources will help current practitioners of G-theory as well as facilitate its use among those new to the field.

Received: 9/19/2024. **Accepted:** 8/4/2025. **Published:** 8/14/2025.

Citation: Huebner, A. Skar, G. B., & Huang, M. (2025). Mixed Model Generalizability Theory: A Case Study and Tutorial. *Practical Assessment, Research, & Evaluation*, 30(7). Available online: https://doi.org/10.7275/pare.2376

Corresponding Author: Alan Huebner, University of Notre Dame. Email: Alan.Huebner.10@nd.edu

References

Brennan, R. L. (2001a). *Generalizability theory*. New York, NY: Springer.
Brennan, R. L. (2001b). Manual for mGENOVA. CASMA, University of Iowa.
Brennan, R. L. (2010) Generalizability Theory and Classical Test Theory, *Applied Measurement in Education*, 24:1, 1-21, DOI: 10.1080/08957347.2011.532417
Gleser, G. C., Cronbach, L. J., & Rajaratnam, N., (1965). Generalizability of scores influenced by multiple sources of variance. Psychometrika, 30, 395-418.
Huebner, A. & Lucht, M. (2019) "Generalizability Theory in R," *Practical Assessment, Research, and Evaluation. 24*(5). DOI: https://doi.org/10.7275/5065-gc10 Available at: https://scholarworks.umass.edu/pare/vol24/iss1/5
In'nami, Y., & Koizumi, R. (2016). Task and rater effects in L2 speaking and writing: A synthesis of generalizability studies. *Language Testing*, 33(3), 341–366. https://doi.org/10.1177/0265532215587390
Institute for Objective Measurement. (2024, March 13). *Disattenuating Correlation Coefficients*. https://www.rasch.org/rmt/rmt101g.htm

Jiang, Z. (2018). Using the linear mixed-effect model framework to estimate generalizability variance components in R: A lme4 Package Application *Methodology*, *14*, 133-142.

Kane, M. T. (2002). Inferences about Variance Components and Reliability-Generalizability Coefficients in the Absence of Random Sampling. *Journal of Educational Measurement*, *39*(2), 165-181. https://www.jstor.org/stable/1435254

Kane, M. T., Crooks, T. J., & Cohen, A. S. (1999). Validating Measures of Performance. *Educational Measurement: Issues and Practice*, *18*(2), 5-17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x

Keller, L., Clauser, B., & Swanson D (2010). Using multivariate generalizability theory to assess the effect of content stratification on the reliability of a performance assessment. *Advances in Health Sciences Education*, 15:717-733

McNamara, T. F. (1996). *Measuring second language performance*. Longman.

Moore, C. T. (2016). gtheory: Apply Generalizability Theory with R. R package version 0.1.2. Retrieved from https://CRAN.R-project.org/package=gtheory.

Ohta, R., Plakans, L. M., & Gebril, A. (2018). Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. Assessing Writing, 38, 21–36. https://doi.org/10.1016/j.asw.2018.08.001

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Schoonen, R. (2005). Generalizability of writing scores: An application of structural equation modeling. Language Testing, 22(1), 1–30. https://doi.org/10.1191/0265532205lt295oa

Shavelson, R. J., & Webb, N.M. (1991). *Generalizability theory: A primer*. Newbury Park, CA: Sage Publications.

Skar, G. B., Aasen, A. J., & Jølle, L. (2020). Functional writing in the primary years: protocol for a mixed-methods writing intervention study. *Nordic Journal of Literacy Research*, *6*(1), 201-216. https://doi.org/10.23865/njlr.v6.2040

Skar, G. B., Jølle, L., & Aasen, A. J. (2020). Establishing scales to assess writing proficiency development in young learners. *Acta Didactica Norge*, *14*(1), 1-30. https://doi.org/10.5617/adno.7909

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), 1-26.

Webb, N. M., Shavelson, R. J., & Haertel, E. H. (2006). Reliability coefficients and Generalizability Theory. In C.R. Rao, & S. Sinharay (Eds.), *Handbook of Statistics Vol 26* (pp. 81-124). ScienceDirect.

Appendix: Nested Design

This manuscript dealt with data resulting from a study in which the facets are fully crossed. However, G-theory is capable of handling a wide variety of designs, and nested designs are common in practice. Specifically, a plausible variation of the crossed design presented in the manuscript is that all students are rated on all subscales, but the raters are nested within students. The univariate G-theory version of this model would be notated as $(r{:}\,p) \times s$, while the multivariate G-theory version is notated as $r^{\bullet}{:}\,p^{\bullet}$ (see table 9.2 in Brennan [2001a]). To perform a similar analysis with that design, the univariate G-study can be conducted using the example code in Table 11 of Huebner and Lucht (2019), and the multivariate G-study can be conducted using the freely available mGENOVA software (Brennan, 2001b). Note, the algorithm executed by the R code in the GitHub link provided above does not estimate the $r^{\bullet}{:}\,p^{\bullet}$ model.