Improving generalizability coefficient estimate accuracy: A way to incorporate auxiliary information

Zhehan Jiang1, Kevin Walker1, Dexin Shi2 and Jian Cao3

Abstract
Initially proposed by Marcoulides and further expanded by Raykov and Marcoulides, a structural equation modeling approach can be used in generalizability theory estimation. This article examines the utility of incorporating auxiliary variables into the structural equation modeling approach when missing data is present. In particular, the authors assert that by adapting a saturated correlates model strategy to structural equation modeling generalizability theory models, one can reduce any biased effects caused by missingness. Traditional approaches such as an analysis of variance do not possess such a feature. This article provides detailed instructions for adding auxiliary variables into a structural equation modeling generalizability theory model, demonstrates the corresponding benefits of bias reduction in generalizability coefficient estimate via simulations, and discusses issues relevant to the proposed approach.

Keywords
Auxiliary variables, factor analysis, generalizability theory, missing data, structural equation modeling

Generalizability theory
When it comes to educational and psychological measurements, the estimate of reliability—an index reflecting the precision of an assessment—provides the researcher with a critical understanding of measurement quality. For example, when attempting to determine an individual’s mathematical proficiency, it is common to use a set of relevant test items (items) to gauge proficiency. However, according to measurement theory, the measurement process itself can be rife with multiple, simultaneously occurring measurement errors. Indeed, the raw test scores within such a scenario are implicitly distinct from the phenomenon being measured and, therefore, cannot accurately reflect the true skill level of an examinee. By controlling for various measurement errors, statistical frameworks, such as the well-known classical test theory (CTT), yield more accurate estimations of proficiency or skill level. Within this context, measurement reliability is expressed as the correlation between the true skill levels of an examinee and those observed levels of skill represented by a testing score.

Within the framework of CTT, which assumes (1) each observed score has errors and (2) the errors follow a normal distribution, an observed score is equal to a true score plus the measurement error (i.e. \( X = T + e \)). Here coefficient \( \alpha \) provides the mathematical equivalent of parallel tests, allowing for a derivation of test reliability that indicates what proportion of observed test score variance is attributable to true score variance (Cronbach, 1951). However, as this approach only assumes two components for a test score (i.e. true score and measurement error), this technique promotes an overly simplistic solution within the context of most research designs. To overcome this shortcoming, CTT and Cronbach’s alpha have been further extended to generalizability theory (G theory) (Cronbach et al., 1972).

While CTT calls for the decomposition of observed variances into true score variance and a single, all-encompassing error term, G theory allows the researcher to account...
for several sources of variation, also known as facets. This provides for the disentanglement of any number of sources of measurement error originating from test items and/or graders. Such robust control of measurement error is simply not possible with CTT, and is what allows G theory to attain greater accuracy when generalizing observed scores to a broader universe.

G theory can provide the variance estimates for a wider variety of errors than CTT—errors that should not be ignored in a research design. These variance estimates allow one to calculate the level of generalizability or dependability of behavioral measurements. While research designs focused on generalizability are used to examine relative interpretations of measurement outcomes, those which aim at dependability are applied toward absolute interpretations. G theory decomposition also enables researchers to make decisions about how to increase generalizability/dependability coefficient to any specified level. For example, if one uses G theory and finds that there is large variance in grader effect (i.e. a lack of consistency among graders, which is undesirable), a follow-up G theory study (namely, a decision study, or D-study) can be used to identify the number of graders required for the measurement to reach a sufficient level of precision.

G theory provides answers to questions such as the following: What is the largest source of error? Is the study generalizable enough to provide important insights regarding latent variables of interest? Based on the current analysis, what test alterations can be made to reach a certain level of generalizability/dependability? These features provide G theory with wide applicability within the realm of psychological and educational assessments. For instance, Dunbar et al. (1991) studied the generalizability of findings across a substantial number of direct writing assessments, Gebril (2009) applied G theory in language testing, Jiang and Raymond (2018) used G theory to investigate subscore validity in score reporting, and Raymond, Clauser, and van Zanten (2009) studied spoken English proficiency scores within a G theory framework. Essentially, selecting an appropriate modeling strategy for a G theory study depends on the research designs. Nested design is a research design where levels of one facet are hierarchically subsumed under or nested within levels of another facet. Crossed design is a research design that has at least two facets that are crossed, that is, every category of one facet co-occurs in the design with every category of the other facet. Here a facet is defined by a certain set of conditions.

In a testing context, for example, test items and test administrations could be regarded as two separate facets. However, a test-taker’s ability is the object of measurement (NuiBaum, 1984). Here, an educational testing scenario, which incorporates test-takers, items, and test-raters (raters), is used for illustration. Setting rater effect aside for the moment, the mathematical expressions for a typical one-facet, fully crossed design are as follows

\[ X_{pi} = \mu + \mu_p + \mu_i + \mu_{pi} \]  
(1)

\[ \sigma(X)_{pi}^2 = \sigma_p^2 + \sigma_i^2 + \sigma_{pi}^2 + \sigma_{pi.e}^2 \]  
(2)

Equation (1) shows that an observed score, \( X \), for person \( p \) on item \( i \) is defined by the sum of the grand mean \( \mu \), person effect \( \mu_p \), item effect \( \mu_i \), and error effect \( \mu_{pi} \). Note that the fully crossed design in the current scenario means that all persons have responded to all items. Correspondingly, the relevant variance components are outlined in equation (2). A typical two-facet, fully-crossed design can be expressed as follows

\[ X_{pji} = \mu + \mu_p + \mu_i + \mu_j + \mu_{pi} + \mu_{pj} + \mu_{pji} \]  
(3)

\[ \sigma(X)_{pji}^2 = \sigma_p^2 + \sigma_i^2 + \sigma_j^2 + \sigma_{pi}^2 + \sigma_{pj}^2 + \sigma_{pji}^2 + \sigma_{pji.e}^2 \]  
(4)

In addition to person effect \( \mu_p \) and item effect \( \mu_i \), equation (3) also contains rater effect \( \mu_j \) as well as interaction terms for any two random components. Equation (4) provides variance components in accordance with equation (3). In both one- and two-facet designs, all \( \sigma \)s are dispersion parameters from independent and identically distributed (iid) normal shapes whose central locations are all equal to zero. Similarly, the fully crossed design in here means that all persons have responded to all items which are graded by all raters.

After estimating variance components, the G coefficient, a reliability-type index, can be calculated. As the aforementioned delineation between generalizability and dependability implies, the G coefficient can be either relative (norm-referenced) or absolute (criterion- or domain-referenced) (see Brennan, 1997, for details). However, here, only the norm-referenced G coefficient is addressed. The G coefficient, based on relative errors, is defined as follows

\[ E\rho_\delta^2 = \frac{\sigma_p^2}{\sigma_p^2 + \sigma_\delta^2} \]  
(5)

\( \sigma_\delta^2 \) is a relative error variance term, which can be expressed as follows

\[ \sigma_\delta^2 = \frac{\sigma_{pi}^2}{n_i} + \frac{\sigma_{pj}^2}{n_j} + \frac{\sigma_{pji.e}^2}{n_i n_j} \]  
(6)

for a design that is defined by equations (3) and (4), where \( n_i \) and \( n_j \) are the numbers of levels for \( \mu_i \) and \( \mu_j \) effects, respectively. This article focuses on the impact of missing data on G coefficient. In addition, an approach based upon
Structural equation modeling (SEM) approach in G theory

SEM is widely used in the psychological, educational, and behavioral sciences. It is a powerful technique combining multivariate observed variables with complex structures. For instance, Figure 1 shows a simplified SEM example for modeling life satisfaction, turnover intent, and organizational citizenship behavior that are linked with 12 manifest or observed variables. Here, all three latent variables are measured with statement-based, self-report instruments and a Likert-type scale (see Schreiber, 2008). Note that a manifest analysis model differs from the latent models in the variables included: unlike latent models such as SEM, the manifest analysis model only contains observed variables and does not have any latent variable.

Within a SEM framework, one can verify the relations between (1) observed variables and latent/unobservable variables and (2) any pair of latent/unobservable variables. In the case of Figure 1, SEM allows the researcher to investigate the ways in which latent variables correlate with one another. In addition to providing insights linked to causal inferences and path analyses, SEM can also be used to study those longitudinal/repeated measurements typical to growth modeling and path analysis.

Within the context outlined here, G theory estimations are based primarily upon a special case of SEM known as confirmatory factor analysis (Jöreskog, 1969), which provides several benefits not present with other methods. For instance, the SEM approach can derive model fit indices, calculate standard errors of the estimates, and handle missing data. It is likely that such benefits have led to an increased use of SEM over the past decade by researchers like Gessaroli and Folske (2002), Hagtvet (1997, 1998) Marcoulides (1996, 2000) and Raykov and Marcoulides (2006).

Figure 2. A structural equation model for estimating the variance components in a two-facet fully crossed design.

Marcoulides (1996) described how to derive estimations using G theory within the context of a covariance structure analysis, while Raykov and Marcoulides (2006) fit this methodology within the SEM framework. Here a brief introduction to these methods is provided. Consider a two-facet, fully crossed design where the number of test-takers, items, and raters are 100, 5, and 2. Following the structure and notations of Raykov and Marcoulides, a corresponding diagram is presented in Figure 2. Rectangular boxes containing the letter Y represent observed variables, whose subscripts indicate their respective groupings. For example, Y32 contains the responses for the 100 test-takers to Item 3 scored by Rater 2. Latent variables—the variance components of G theory—are represented by circles. \( \mu_p \) represents person effect. \( \mu_{pi1} \) to \( \mu_{pi5} \), constrained to be equal, represent the interactions between item and person effects. The same applies to \( \mu_{pr1} \) and \( \mu_{pr2} \) so that the interaction between rater and person effects can be derived. Each arrow is a factor loading that regresses a certain variance component on a corresponding observed variable Y. These loadings are equal to one (1). Note, not all residuals for observed variables are shown in Figure 2. Nevertheless, they are constrained to be equal when a computation is performed. Mathematical proofs of converting a G theory model to SEM can be found in Raykov and Marcoulides (2006). With the rapid development of various statistical software packages, standard errors of estimates in SEM are now computed more easily than many other solutions, such as a traditional analysis of variance (ANOVA) (Searle, 1971) or restricted maximum likelihood (REML) estimator (Shavelson and Webb, 1991).

Missing data in SEM framework

A majority of scholars outside of the field applied statistics remain blissfully unaware of the dangers of missing data.
They are unaware of techniques that could help nullify the negative effects of missing data within their research studies. Indeed, previous research has shown how cases of pairwise and case-wise deletion hurt validity—and yet the use of data flawed in this manner persists. What’s more, in the face of missing data, the use of inappropriate methods continues to inform biased estimates that lead to invalid conclusions (Enders, 2008). For these reasons, it is of the utmost importance that researchers become familiar with those statistical techniques that provide a robust solution to the problem of missing data.

There are three types of missing data. Data can be missing completely at random (MCAR), missing at random (MAR), or not missing at random (MNAR) (Little and Rubin, 2002). With data that are MCAR, missingness is defined independently of other observable and missing data. For example, a questionnaire/survey might be lost in the post, or a web browser malfunction might lead to the loss of question response records. In the case of data that are MAR, missingness is defined independently of missing data themselves, given other observed data. Furthermore, once those observed data are controlled for, missingness is explainable independently of any missing data. For example, if test data are missing due to test-taker illness (i.e. the test-taker was not present to provide data), the missingness is sufficiently explainable by the test-taker’s circumstance and has no relation to those data that would have been provided had the test-taker not fallen ill. Finally, missingness of data said to be MNAR is attributable to unobserved data (Allison, 2001; Enders, 2010). For example, a person does not complete a drug screening because they have used drugs and do not want to fail the test (Kang, 2013). Importantly, while solutions are available to account for missing data in the cases of MCAR and MAR, there are no appropriate solutions to MNAR.

To reiterate, the MAR assumption holds when the probability of missing data for a particular variable relates to other variables (informative variables) present in the model. If the informative variables are not included in the model, the analysis is performed with an MNAR mechanism, which tends to produce biased estimates. Therefore, MAR and MNAR differ in their inclusion of informative variables within a given model. In the current article, only MAR is considered because (1) true MCAR is rare in practice and (2) the solutions that work for MAR mechanism also works for MCAR situations. Particular in G theory literature, missing issues are discussed intensively within a MCAR pattern, which is essentially treated as unbalanced data situation. For example, Brennan (2001) used analogous-ANOVA estimators, also known as Henderson’s (1953) Methods 1 and 3, to adjust unequal sample sizes in different groups. In yet another example, Chiu and Wolfe (2002) proposed subdividing method by “creating data sets that exhibit structural designs that are common in generalizability analyses.”

Recent missing data studies have demonstrated advantages of an “inclusive analytic strategy” that incorporates auxiliary variables into the analysis model or into the imputation process (Enders, 2008). An auxiliary variable is defined as a variable that one would include in an analysis because it provides the information about the part of data that are missing (Collins et al., 2001; Rubin, 1976; Schafer, 1997; Schafer and Graham, 2002). As described previously, to meet the MAR assumption, the “cause” information must be considered when data are modeled. Indeed, this “cause” information defines those auxiliary variables, the primary function of which is to reduce potential bias within any explanatory model—as opposed to providing direct explanatory insights of those phenomena being studied.

Many studies have illustrated how auxiliary variables can be used to bolster explanatory power and mitigate bias by recapturing certain lost information (Collins et al., 2001; Enders, 2008). Nevertheless, such methods are not fool proof, in that they cannot guarantee unbiased estimates (Enders, 2010). Instead, success depends upon how well auxiliary variables correlate with missingness or any incomplete variables within the model under study—as signaled by the degree of bias reduction in model estimations. To illustrate, if income is an important variable within a survey analysis, a question that asks respondents to report their occupation will likely provide a good auxiliary variable, due to the correlation between the resulting data and income.

Although auxiliary variables have fewer downside effects on a model’s parameter estimates, fitting a maximum likelihood estimation with a large number of auxiliary variables can be problematic, due to complex specification (Enders, 2010: 133). That said, to successfully perform a likelihood-based estimation, one should only incorporate auxiliary variables that have more informative utility. An ideal auxiliary variable can reflect, or at least highly correlate with, the true cause of missingness within a model. Selecting useful auxiliary variables, however, is not always a straightforward process. In many cases an extensive review of the pertinent literature may be called for, coupled with good, old-fashioned guesswork. For example, knowing that family mobility can lead to school attrition, a researcher conducting survey research might ask households to report the likelihood of an upcoming move (Enders et al., 2006; Graham et al., 1997). In addition to utilizing theory to guide one’s selection of appropriate auxiliary variables, statistical evidence may also be used. Collins et al. (2001) any variable that correlating at a level of 0.4 or higher with any incomplete variable, would make for a good auxiliary variable.

Graham (2003) described two SEM-based strategies for incorporating auxiliary variables into a maximum likelihood analysis: the extra dependent variable model and the saturated correlates model. Here, the inclusion of auxiliary variables is possible without fear of altering the interpretation of substantive theoretical constructs of interest. The saturated correlates model, in particular, provides for an easier path than the extra dependent variable model, while producing similar results. Alternatively, researchers have proposed the
In this article, auxiliary variables are incorporated into the saturated correlates model, a SEM-based approach, to handle missing data problems in a G theory study. The research hypotheses are (1) MAR missing data can lead to biased G coefficient estimates to a degree in accordance with the variance size and the association with the missing cause(s), (2) using the proposed approach would yield more accurate estimates when missing data exist, and (3) it is a safe strategy to incorporate auxiliary variables into the model because it has no harmful impacts on the actual estimates even with no missingness. Figure 3 illustrates the application of G theory within a SEM framework. Aside from the inclusion of two auxiliary variables ($AV1$ and $AV2$), which correlate (1) with each other and (2) with the indicators’ error variance, this model is identical to that seen in Figure 2. In addition, bidirectional arrows represent correlations between auxiliary variables and observed variables. In the coming simulation section, synthetic data are all modeled following the same strategy.

Within this study, lavaan is used to conduct SEM analysis. As described by Rosseel (2012), lavaan is a software package for use within the R software environment. It is an open source software that provides a straightforward approach to constructing complex SEM frameworks. In the hope that it will provide a more approachable example of the concepts shared herein, Figure 4 illustrates the lavaan coding syntax used for constructing the theoretical model shown in Figure 3 (see link for downloading the script file). Within Figure 4, the first line of code displays $HS.model$, an object that contains a G theory SEM model that meets those parameters outlined in coding lines 2 through 30. Furthermore, facets such as person effect, item effect, and rater effect are represented by the coding variable labels $p_factor$, $i_factor$, and $r_factor$, respectively. Here also, $v$s represent indicators, while $aux$s represent auxiliary variables. Finally, $pvar$, $ivar$, and $rvar$ represent functions used to constrain the equivalency of estimates within the model. The right-most portion of those formulas contained in coding lines 2 through 9 are predicted by corresponding facets shown to the left of the $\sim$ symbol—where all loadings are set to 1. Coding lines 10 through 30 use $\sim$ to represent correlation (if linking different variables) and variance/residuals (if linking the same variable). For example, $aux1 \sim aux2$ specifies the correlation between the two auxiliary variables, while coding lines 11 and 12 prompt correlation comparisons between each auxiliary variable and all other observed variables. At Lines 19 and 20, the variance of $r_factor1$ and the variance of $r_factor2$ are set to equal by $rvar$. The similar idea applies to other constraint labels such as $pvar$ and $res$. Line 31 feeds the specified $HS.model$ object to the lavaan function $sem$ such that the model can be tested in most cases, particularly in cases of MAR data associated with the absence of auxiliary variables in the imputation model.

Figure 3. A saturated correlates model for estimating the variance components in a two-facet fully crossed design with missing data.
actually executed. Here also, the data set which contains missing data is named lav_dat. Furthermore, the method of deriving those missing data within this SEM model is set to full information maximum likelihood (FIML), and, finally, the exogenous latent variables are assumed to be uncorrelated by setting the orthogonal model feature to TRUE. Additional details regarding SEM model specification within the lavaan R package can be found in Rosseel (2012).

Simulation design

The process of testing the utility of a SEM framework began with a pilot study conducted using a data set that contained no missing values. The results of this pilot show that when data are complete, both SEM and REML yield extremely similar estimates. To investigate the utility of incorporating auxiliary variables into SEM G theory, a comprehensive simulation study was conducted. Instead of arbitrarily choosing variance component values, the results of the California Assessment Program (CAP) dependability analyzed by Shavelson et al. (1993) were used to serve as a baseline for data generation. For the purpose of notation consistency, the facet of measurement occasion is treated as an item facet. Here it should be noted that this type of two-item design is not uncommon within the realm of educational assessment, especially when such assessments are aimed at writing tasks.

Within the context of equations (3) and (4), outlined in an earlier section of this article, the baseline true parameters are $\sigma^2_p = 0.298$, $\sigma^2_i = 0.092$, $\sigma^2_j = 0.003$, $\sigma^2_{pi} = 0.493$, $\sigma^2_{ij} = 0.000$, $\sigma^2_{pij} = 0.002$, and $\sigma^2_{pije} = 0.148$. Correspondingly, the numbers of levels of $\mu_p$, $\mu_i$, and $\mu_j$ effects are $n_p = 50$, $n_i = 2$, and $n_j = 3$ respectively. Here a Monte Carlo sampling approach is used to simulate data responses. Furthermore, it should be noted that the CAP study was originally conducted under a two-facet, fully crossed design as exemplified in equations (3) and (4). The SEM approaches illustrated in Figures 2 and 3 formed the basis for data analysis.

In order to inject variance within the G coefficient during testing, three additional levels of person effect, $\sigma^2_p = (0.098, 0.698, 1.698)$, were used. These levels, although somewhat arbitrary, are specified here for reference purposes. That is to say, these values were chosen to help to show how results vary, depending on the degree of person effect variance. Alternatively, manipulating G coefficient

Figure 4. Lavaan specification for the model illustrated in Figure 3.
levels can be achieved by varying \( n_i \) or/and \( n_j \). However, keeping simulation parameters consistent with this study’s purpose, variations were limited to the person effect (i.e. \( \sigma^2_p \)). To summarize, there were four levels of person effect utilized in this simulation design—the original effect when no data were missing, as well as the three variable effects designated when predicting values using incomplete data sets.

After generating responses from a given true variance set, incomplete data conforming to a MAR pattern were injected into the data. Following the approach prescribed by Park and Shin (1998), a MAR pattern was injected into the data set via two auxiliary variables (AV1 and AV2) derived by simulating two columns of data correlating with \( \mu_p \). Here it should be noted that the method for deriving AV1 and AV2 falls in line with that approach taken by Wu et al. (2015). In addition, three levels of average correlation were used to generate AV1 and AV2: \( \rho_{av} = \pm 0.1, \pm 0.5, \pm 0.9 \). That is, for a specified level of \( \rho_{av} \), for example 0.9, the two actual correlations between auxiliary variables and person effect could take on the values \( \rho_{av1,p} = 0.95 \) and \( \rho_{av2,p} = 0.85 \), or \( \rho_{av1,p} = 0.90 \) and \( \rho_{av2,p} = 0.90 \).

Consistent with the modeling practice outlined in Figure 3, there were \( n_i \) * \( n_j \) indicators in the primary model structure. Missing data for these indicators were incorporated into the data set based on AV1 and AV2. To be more specific, splitting the indicators evenly into two subsets, AV1 produced missing cells for the first subset and AV2 produced missing data for the second subset. The ordered AVs were ranked from smallest to the largest, while the probability of any particular data point being marked as missing was based on that datum’s rank order. As an example, for the first indicator, the probability of having missing data was computed as 1 minus the rank order of the value of \( \text{AV1}/n_p \). This probability is then compared to a random number drawn from a uniform distribution ranging from zero (0) to one (1). Where the probability is higher than the random draw, the corresponding cell is set as missing. This procedure continues for each indicator, barring one, until a pre-defined percentage of cell is set as missing. This procedure continues for each indicator, barring one, until a pre-defined percentage of cell is set as missing. This procedure continues for each indicator, barring one, until a pre-defined percentage of cell is set as missing. This procedure continues for each indicator, barring one, until a pre-defined percentage of cell is set as missing. This procedure continues for each indicator, barring one, until a pre-defined percentage of cell is set as missing.

As simulations progress, a coding element produces a binary indicator that signals when the standard error for data estimates surpass the set confidence interval.

### Simulation results

Comprehensive results for FIML simulations can be found in Appendix 1. In addition, Figure 5 shows the G coefficient bias and MSE results when the missing percentage is set to 15%. When \( \sigma^2_p \) is smaller (i.e. a smaller, true G coefficient), the relative bias and MSE tended to be larger. To illustrate, given that the estimation was No Aux and \( \rho_{av,p} \) was equal to 0.9, the relative bias for G coefficient was -25.66%, when \( \sigma^2_p \) was equal to 0.098 which resulted in a true G coefficient equal to 0.265. The relative bias reduced to -2.20% when \( \sigma^2_p \) increased to 1.698 while holding other controlled variables constant.

For a given absolute value of \( \rho_{av,p} \), both negative and positive values tended to produce similar relative bias and MSE estimates. With Aux yielded more accurate G coefficients in 17 out of 24 conditions; the seven exceptions, however, showed that the corresponding biases produced by With Aux and No Aux were extremely close. A similar trend can be found in the MSE result panel. Therefore, it can be concluded that With Aux outperforms No Aux in terms of both estimate accuracy and efficiency. As one might expect, when the absolute value of \( \rho_{av,p} \) is small, With Aux did not improve the estimate accuracy, whereas a larger absolute value of \( \rho_{av,p} \) led to improved accuracy in the With Aux model. For
instance, when $\sigma_p^2$ was 0.698, With Aux reduces the G coefficient relative bias from $-6.67\%$ to $-0.56\%$ at the condition that $\rho_{av.p} = 0.9$, where the reduction became 1.39\% (i.e. 3.06\%–1.67\%) at $\rho_{av.p} = 0.5$ condition.

To analyze how the components of G coefficients function within this simulation study, the biases and MSEs of $\sigma_p^2$, $\sigma_{pr}^2$, $\sigma_{pe}^2$, and $\sigma_{pi,e}^2$ are examined. Figure 6 lists the relative biases and MSEs of $\sigma_p^2$ under conditions of 15\% missingness. As one can see, With Aux simulations recovered the $\sigma_p^2$ estimates with greater accuracy and efficiency than seen with No Aux simulations. For a given absolute value of $\rho_{av.p}$, being negative or positive did not change the estimates provided by both modeling strategies. When the absolute value of $\rho_{av.p}$ increased, No Aux tended toward greater bias and inefficiency. At the same time, within those same conditions, the capacity of With Aux simulations to correct for imprecise estimates strengthened. These patterns are similar to those illustrated within Figure 5.

Here it should be noted that the MSE of $\sigma_p^2$ is consistently smaller than the corresponding G coefficient. Furthermore, although $\sigma_{pi}^2$ is not the cause of the missing data, in several circumstances the estimates are noticeably biased. For example, when true $\sigma_{pi}^2$ is 0.698 and $\rho_{av.p} = 0.9$, the biases for No Aux and With Aux are 3.25\% and $-1.22\%$ respectively. Similar to the $\sigma_p^2$ outcomes, MSE results show that, across all conditions, both modeling strategies provided highly efficient estimates. The other two components ($\sigma_{pr}^2$ and $\sigma_{pi,e}^2$) were not affected by missingness. Indeed, within the context of both estimations, their biases and MSEs all fall below 0.005 (many of them were below 0.001). Note, if true $\sigma_{pr}^2$ were to be set to a non-zero number, it would have been expected to produce bias and estimates similar to that seen with $\sigma_{pr}^2$.

While the primary focus of this article is determining the accuracy of G coefficient estimates, standard errors of the estimates are also studied—since they speak to the quality of statistical inferences. As shown in Table 1, results of the true $\sigma_{pi}^2$ coverage deviate up to 0.95 when using the standard errors of the estimate. Since the direction of $\rho_{av.p}$ yielded negligible differences in the current output, they were collapsed based upon their absolute values. Here, a negative value shows that the corresponding standard error is underestimated, while a positive value indicates overestimation. In both cases, such output indicates that Type I errors are not being effectively controlled for. What’s more, as the percentage of missingness increases, the level at which standard errors deviate from our set confidence interval increases. Cases of a larger absolute value of $\rho_{av.p}$ tended to result in less accurate standard error estimates. Compared with No Aux, which produced unstable standard errors across the simulation conditions, With Aux provided more accurate estimates—as even the largest coverage rate deviation from 0.95 was as low as $-0.028$. Nevertheless, the standard errors of other variance components are fairly good, considering that all deviations fall below 0.03 for both modeling approaches.

**Discussion**

Missing data can lead to biased estimates, and thus eliminating or reducing the influence of missing data is important when G theory analysis is used for decision making, especially those that are high stakes. It can be calculated from Table 1, in the condition where $\rho_{av.p} = 0.9$ and $\sigma_{pi}^2 = 0.698$, the true G coefficient is 0.720 and the estimate can be as low as 0.676 if the auxiliary variables are not taken into account. If 0.7 is used as a criterion for an assessment quality evaluation, the biased G coefficient estimate, due to missingness, would lead to a conclusion that the assessment is not reliable enough. This finding indicates that when missingness occurs in a G theory study, one can find variables related to the application of the proposed method, a recent study using “symptom-check-list-27-plus” (SCL-27-plus) along with the scale of the quality of life (QOL) is delineated here. Hardt et al. (2012) investigate symptoms of depression, agoraphobia, social anxiety, pain, and vegetative symptoms via the SCL-27-plus where the database involved missing cells that are susceptible yielding an inaccurate analysis. In order to reduce the estimate biases, Hardt, Herke, and Leohart use
QOL responses to serve as auxiliary variables by matching the identifications of respondents and the results show that the estimates are less biased when the QOL responses are aligned with the SCL-27-plus. Despite the fact Hardt, Herke, and Leohart’s study is not based upon G-theory, the idea of applying auxiliary variables to trade more accurate estimates is identical to that of the present article: when conduct a G-theory study, it is preferable for researchers to collect information from multiple resources, even some of the information is not the focus of research interest, such that the negative consequences of missing data can be minimized.

To keep simulation design manageable, the missing data mechanism was based upon person effect only. Although theoretically, it could relate to item effect and/or rater effect, in practice it is more viable to assume participants themselves are the missingness reason. Besides, in

Figure 6. Sample average relative biases and mean square errors for $\sigma^2$ estimates.
some situations, it would be redundant to assume that the
causes of missing data stem from various effects simulta-
neously. To illustrate, participants from a high-income
family are likely to miss a survey question about free and
reduced lunch. Assuming that the missingness takes place
due to the item being too difficult for persons from a high-
income family, is in fact, commensurate with the assump-
tion on persons’ socioeconomic statuses. This practice,
correlating auxiliary variables with person effect only,
would be attenuated when missing cells are removed from
the full data set. That said, the actual correlation between
missingness in an observed data set and an auxiliary vari-
able is lower than what it would be at the person effect
level.

Providing model fit is another important feature of SEM.
In particular, if a G theory framework fits data poorly, fur-
ther analyses based on the estimation will be untrustworthy.
Essentially, G theory is a modeling framework like any
other statistical models. Traditional estimations do not pro-
vide fit information about a certain G theory framework
used in research; if a data set that was planned in a two-facet
design but is inappropriately/wrongfully fitted into a one-
facet design, the model fit in SEM is expected to capture the
misfit. In addition, SEM has assembled with a FIML estimator,
which makes the modeling easier for handling missing
data problems without complicated corrective procedures
(Allison, 2000). Recent model fit interpretation can be found
in Kline (2015), Shi et al. (2017) and Shi et al. (2018).

Theoretically, SEM approach can estimate other main
effects such as item effect and rater effect; it needs to handle
the data format differently. For example, if one is interested
in estimating the item effect, each row of the input data
should be responses of an item (instead of a person).
Practically, however, this practice often is not permissible as
the levels of other main effects are not sufficiently large,
while the person effect can result in many indicators. To
solve the problem, Marcoulides (2000) derived an approach
of estimating other main effects by analyzing the matrix of
correlations. In this article, if a measurement is domain-re-
ferred, SEM approach used here is not appropriate as it
provides insufficient information. If absolute errors are
needed and current existing software packages are consid-
ered, the saturated correlates model can be used to cross-
validate the feasibility of the traditional approaches such as
REML. If the common variance components from both esti-
mations are close to a certain acceptable level, the effects
that the saturated correlates model does not estimate can be
obtained from REML. Alternatively, one can use a Bayesian
framework to analyze a G theory study (Little and Rubin,
2014). During the Bayesian estimation process, missing data
imputation can be accommodated simultaneously (Jiang and
Skorupski, 2017; Qin, 2018).

Declaration of conflicting interests
The author(s) declared no potential conflicts of interest with respect
to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, author-
ship, and/or publication of this article.

ORCID iD
Zhehan Jiang https://orcid.org/0000-0002-1376-9439

References
Allison PD (2000) Multiple imputation for missing data: A caution-
ary tale. Sociological Methods & Research 28(3): 301–309.
Allison P (2001) *Missing Data*. Thousand Oaks, CA: SAGE.
Brennan RL (1997) A perspective on the history of generalizability theory. *Educational Measurement: Issues and Practice* 16(4): 14–20.
Brennan RL (2001) *Generalizability Theory*. New York: Springer.
Chiu CW and Wolfe EW (2002) A method for analyzing sparse data matrices in the generalizability theory framework. *Applied Psychological Measurement* 26(3): 321–338.
Collins LM, Schafer JL and Kam C-H (2001) A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods* 6: 330–351.
Cronbach LJ (1951) Coefficient alpha and the internal structure of tests. *Psychometrika* 16: 297–334.
Cronbach LJ, Gleser GC, Nanda H, et al. (1972) *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. New York: Wiley.
Dunbar SB, Koretz DM and Hoover HD (1991) Quality control in the development and use of performance assessments. *Applied Measurement in Education* 4(4): 289–303.
Enders CK (2008) A note on the use of missing auxiliary variables in full information maximum likelihood-based structural equation models. *Structural Equation Modeling* 15(3): 434–448.
Enders CK (2010) *Applied Missing Data Analysis*. New York: Guilford Press.
Enders CK, Dietz S, Montague M, et al. (2006) Modern alternatives for dealing with missing data in special education research. In: Scruggs TE and Mastropieri MA (eds) *Applications of Research Methodology*. Bingley: Emerald Group Publishing, pp. 101–129.
Feinberg RA and Rubright JD (2016) Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice* 35(2): 36–49.
Gebril A (2009) Score generalizability of academic writing tasks: Does one test method fit it all? *Language Testing* 26(4): 507–531.
Gessaroli ME and Folske JC (2002) Generalizing the reliability of tests comprised of testlets. *International Journal of Testing* 2: 277–295.
Graham JW (2003) Adding missing-data-relevant variables to FIML-based structural equation models. *Structural Equation Modeling* 10(1): 80–100.
Graham JW, Hofer SM, Donaldson SI, et al. (1997) Analysis with missing data in prevention research. *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research* 1: 325–366.
Hagtvet KA (1997) The function of indicators and errors in construct measures: An application of generalizability theory. *Journal of Vocational Education Research* 22(4): 247–266.
Hagtvet KA (1998) Assessment of latent constructs: A joint application of generalizability theory and covariance structure modelling with an emphasis on inference and structure. *Scandinavian Journal of Educational Research* 42: 41–63.
Hardt J, Herke M and Leonhart R (2012) Auxiliary variables in multiple imputation in regression with missing X: A warning against including too many in small sample research. *BMC Medical Research Methodology* 12(1): 184.
Henderson CR (1953) Estimation of variance and covariance components. *Biometrics* 9(2): 226–252.
Jiang Z and Raymond M (2018) The use of multivariate generalizability theory to evaluate the quality of subscores. *Applied Psychological Measurement*. Epub ahead of print 3 April. DOI: 10.1177/0146621618758698.
Jiang Z and Skorupski W (2017) A Bayesian approach to estimating variance components within a multivariate generalizability theory framework. *Behavior Research Methods*. Epub ahead of print 12 December. DOI: 10.3758/s13428-017-0986-3.
Jöreskog KG (1969) Efficient estimation in image factor analysis. *Psychometrika* 34(1): 51–75.
Kang H (2013) The prevention and handling of the missing data. *Korean Journal of Anesthesiology* 64(5): 402–406.
Kline RB (2015) *Principles and Practice of Structural Equation Modeling*. New York: Guilford Press.
Little RJ and Rubin DB (2002) Bayes and multiple imputation. *Statistical Analysis with Missing Data*. New York: Wiley.
Marcoulides GA (1996) Estimating variance components in generalizability theory: The covariance structure analysis approach. *Structural Equation Modeling* 3: 290–299.
Marcoulides GA (2000) Generalizability theory: Advancements and implementations. In: *Proceedings of the 22nd language testing research colloquium*, Vancouver, BC, Canada, 9 March.
Nußbaum A (1984) Multivariate generalizability theory in educational measurement: An empirical study. *Applied Psychological Measurement* 8: 219–230.
Park GC and Shin DW (1998) An algorithm for generating correlated random variables in a class of infinitely divisible distributions. *Journal of Statistical Computation and Simulation* 61(1–2): 127–139.
Qin L (2018) Estimating nonlinear indirect effects in Bayesian semiparametric structural equation model. *Multivariate Behavioral Research* 53(1): 130–131.
Raykov T and Marcoulides GA (2006) Estimation of generalizability coefficients via a structural equation modeling approach to scale reliability evaluation. *International Journal of Testing* 6: 81–95.
Raymond MR, Clauer BE and 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.
Rosseel Y (2012) lavaan: An R package for structural equation modeling. *Journal of Statistical Software* 48(2): 1–36.
Rubin DB (1976) Inference and missing data. *Biometrika* 63: 581–592.
Savalei V and Bentler PM (2007) Structural Equation Modeling. In: Grover R and Vriens M (eds) *The Handbook of Marketing Research: Uses, Misuses, and Future Advances*. Thousand Oaks, CA: Sage Publications, pp. 330–364.
Schafer JL (1997) *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall.
Schafer JL and Graham JW (2002) Missing data: Our view of the state of the art. *Psychological Methods* 7: 147–177.
Schreiber JB (2008) Core reporting practices in structural equation modeling. *Research in Social and Administrative Pharmacy* 4(2): 83–97.
Searle SR (1971) *Linear Models*. New York: John Wiley & Sons.
Shavelson RJ and Webb NM (1991) *Generalizability Theory: A Primer*. Newbury Park, CA: SAGE.
Shavelson RJ, Baxter GP and Gao X (1993) Sampling variability of performance assessments. *Journal of Educational Measurement* 30(3): 215–232.

Shi D, DiStefano C, McDaniel HL, et al. (2018) Examining chi-square test statistics under conditions of large model size and ordinal data. *Structural Equation Modeling*. Epub ahead of print 30 March. DOI: 10.1080/10705511.2018.1449653.

Shi D, Song H, Liao X, et al. (2017) Bayesian SEM for specification search problems in testing factorial invariance. *Multivariate Behavioral Research* 52(4): 430–444.

Wu W, Jia F and Enders C (2015) A comparison of imputation strategies for ordinal missing data on Likert scale variables. *Multivariate Behavioral Research* 50(5): 484–503.

Yoo JE (2009) The effect of auxiliary variables and multiple imputation on parameter estimation in confirmatory factor analysis. *Educational and Psychological Measurement* 69(6): 929–947.

Yuan KH and Bentler PM (2000) Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology* 30(1): 165–200.

**Author biographies**

Zhehan Jiang is an assistant professor at the University of Alabama, where his research interest is about psychometrics modeling and computations.

Kevin Walker is an associate professor at the University of Alabama, where he directs the assessment department at the libraries.

Dexin Shi is an assistant professor at University of South Carolina, where his works lies at structural equation modelings and model fits.

Jian Cao is a research assistant of High Performance Computing Center at Weifang University, where he is also a visiting scholar at the University of Queensland.

---

**Appendix 1**

**Result 1**

The G coefficient relative bias and MSE results at the conditions with 15% of missingness.

| Estimation | \( \sigma^2_p = 0.098 \) | \( \sigma^2_p = 0.298 \) | \( \sigma^2_p = 0.698 \) | \( \sigma^2_p = 1.698 \) |
|------------|-------------------|-------------------|-------------------|-------------------|
|            | No Aux | With Aux | No Aux | With Aux | No Aux | With Aux | No Aux | With Aux | No Aux | With Aux |
| **BIAS (rel)** |       |          |        |          |        |          |        |          |        |          |
| \( \rho_{np} = 0.1 \) | –1.51% | –2.26% | –2.10% | –2.48% | –0.28% | –0.42% | –0.23% | –0.35% |
| \( \rho_{np} = 0.5 \) | –15.09% | –8.68% | –5.73% | –2.48% | –3.06% | –1.67% | –0.58% | –0.23% |
| \( \rho_{np} = 0.9 \) | –25.66% | –4.53% | –16.03% | –3.05% | –6.67% | –0.56% | –2.09% | –0.46% |
| \( \rho_{np} = -0.1 \) | –4.53% | –5.28% | –1.72% | –2.10% | –1.25% | –1.39% | –0.23% | –0.23% |
| \( \rho_{np} = -0.5 \) | –5.28% | 0.75% | –4.77% | –1.72% | –1.94% | –0.69% | –0.81% | –0.46% |
| \( \rho_{np} = -0.9 \) | –25.66% | –6.42% | –16.41% | –3.24% | –5.69% | 0.00% | –2.20% | –0.46% |
| **MSE** |       |          |        |          |        |          |        |          |        |          |
| \( \rho_{np} = 0.1 \) | 0.027  | 0.027  | 0.011  | 0.012  | 0.003  | 0.004  | 0.001  | 0.001  |
| \( \rho_{np} = 0.5 \) | 0.031  | 0.028  | 0.014  | 0.012  | 0.005  | 0.004  | 0.001  | 0.001  |
| \( \rho_{np} = 0.9 \) | 0.037  | 0.027  | 0.024  | 0.012  | 0.009  | 0.004  | 0.002  | 0.001  |
| \( \rho_{np} = -0.1 \) | 0.029  | 0.029  | 0.011  | 0.011  | 0.004  | 0.004  | 0.001  | 0.001  |
| \( \rho_{np} = -0.5 \) | 0.028  | 0.026  | 0.014  | 0.012  | 0.005  | 0.005  | 0.001  | 0.001  |
| \( \rho_{np} = -0.9 \) | 0.037  | 0.028  | 0.026  | 0.013  | 0.008  | 0.004  | 0.002  | 0.001  |

Note: The smaller values within a comparison between No Aux and With Aux were bolded.
### Result 2

The $\sigma_p^2$ relative bias and MSE results at the conditions with 15% of missingness.

| Estimation | $\sigma_p^2 = 0.098$ | $\sigma_p^2 = 0.298$ | $\sigma_p^2 = 0.698$ | $\sigma_p^2 = 1.698$ |
|------------|----------------------|----------------------|----------------------|----------------------|
|            | No Aux | With Aux | No Aux | With Aux | No Aux | With Aux | No Aux | With Aux |
| BIAS (rel) |         |         |         |         |         |         |         |         |
| $\rho_{\omega p} = 0.1$ | 4.08% | 4.08% | -1.34% | -1.34% | -0.29% | -0.43% | -0.41% | -0.47% |
| $\rho_{\omega p} = 0.5$ | -11.22% | -3.06% | -7.72% | -2.01% | -6.73% | -2.87% | -2.41% | -0.18% |
| $\rho_{\omega p} = 0.9$ | -23.47% | 2.04% | -22.48% | -2.35% | -14.18% | 0.43% | -9.31% | -1.59% |
| $\rho_{\omega p} = -0.5$ | 1.02% | 1.02% | -0.34% | -0.67% | -1.15% | -1.15% | 0.12% | 0.06% |
| $\rho_{\omega p} = -0.5$ | -6.88% | -1.02% | -5.70% | -0.34% | -4.30% | -0.57% | -3.00% | -0.82% |
| MSE     |         |         |         |         |         |         |         |         |
| $\rho_{\omega p} = 0.1$ | 0.005 | 0.005 | 0.008 | 0.009 | 0.021 | 0.022 | 0.076 | 0.076 |
| $\rho_{\omega p} = 0.5$ | 0.005 | 0.005 | 0.009 | 0.009 | 0.023 | 0.022 | 0.084 | 0.084 |
| $\rho_{\omega p} = 0.9$ | 0.005 | 0.005 | 0.013 | 0.009 | 0.035 | 0.025 | 0.114 | 0.090 |
| $\rho_{\omega p} = -0.1$ | 0.005 | 0.005 | 0.009 | 0.009 | 0.023 | 0.023 | 0.077 | 0.077 |
| $\rho_{\omega p} = -0.5$ | 0.005 | 0.005 | 0.009 | 0.009 | 0.025 | 0.025 | 0.082 | 0.080 |
| $\rho_{\omega p} = -0.9$ | 0.005 | 0.005 | 0.014 | 0.009 | 0.033 | 0.026 | 0.110 | 0.088 |

Note: The smaller values within a comparison between No Aux and With Aux were bolded.

### Result 3

The $\sigma_p^2$ relative bias and MSE results at the conditions with 15% of missingness.

| Estimation | $\sigma_p^2 = 0.098$ | $\sigma_p^2 = 0.298$ | $\sigma_p^2 = 0.698$ | $\sigma_p^2 = 1.698$ |
|------------|----------------------|----------------------|----------------------|----------------------|
|            | No Aux | With Aux | No Aux | With Aux | No Aux | With Aux | No Aux | With Aux |
| BIAS (rel) |         |         |         |         |         |         |         |         |
| $\rho_{\omega p} = 0.1$ | -2.64% | -2.23% | -0.81% | -0.41% | -2.23% | -1.62% | -1.01% | -0.20% |
| $\rho_{\omega p} = 0.5$ | -0.20% | -0.61% | -0.61% | -1.22% | 0.41% | -0.20% | -0.61% | -1.01% |
| $\rho_{\omega p} = 0.9$ | 0.20% | -1.42% | 2.23% | -1.22% | 3.25% | -1.22% | 1.83% | -1.42% |
| $\rho_{\omega p} = -0.1$ | -1.83% | -1.42% | -0.81% | -0.20% | 0.00% | 0.61% | -1.01% | -0.41% |
| $\rho_{\omega p} = -0.5$ | -1.83% | -2.03% | -0.61% | -1.22% | -1.01% | -1.83% | 0.41% | 0.00% |
| $\rho_{\omega p} = -0.9$ | -0.61% | -2.03% | 3.65% | -0.20% | 1.83% | -2.23% | 2.64% | -0.61% |

MSE

| $\rho_{\omega p} = 0.1$ | 0.006 | 0.006 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 |
| $\rho_{\omega p} = 0.5$ | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 | 0.007 |
| $\rho_{\omega p} = 0.9$ | 0.007 | 0.006 | 0.007 | 0.006 | 0.009 | 0.007 | 0.008 | 0.007 |
| $\rho_{\omega p} = -0.1$ | 0.007 | 0.007 | 0.006 | 0.006 | 0.007 | 0.007 | 0.008 | 0.008 |
| $\rho_{\omega p} = -0.5$ | 0.007 | 0.007 | 0.007 | 0.007 | 0.008 | 0.008 | 0.007 | 0.007 |
| $\rho_{\omega p} = -0.9$ | 0.008 | 0.007 | 0.009 | 0.007 | 0.007 | 0.006 | 0.007 | 0.006 |

Note: The smaller values within a comparison between No Aux and With Aux were bolded.
Result 4
The G coefficient bias comparison between conditions with different missing percentage levels.

| Estimation | $\sigma_p^2 = 0.098$ | $\sigma_p^2 = 0.298$ | $\sigma_p^2 = 0.698$ | $\sigma_p^2 = 1.698$ |
|------------|-----------------------|-----------------------|-----------------------|-----------------------|
|            | No Aux | With Aux | No Aux | With Aux | No Aux | With Aux | No Aux | With Aux |
| 15% vs 45% | $\rho_{mp} = 0.1$ | 9.81% | 10.57% | 3.24% | 3.44% | 0.97% | 1.39% | -0.12% | -0.12% |
|            | $\rho_{mp} = 0.5$ | 12.83% | 11.32% | -0.95% | -0.38% | -0.14% | 0.14% | 0.35% | 0.35% |
|            | $\rho_{mp} = 0.9$ | 6.04% | 4.91% | -0.76% | -0.38% | -0.69% | -0.28% | -0.58% | -0.46% |
|            | $\rho_{mp} = -0.1$ | 4.91% | 4.91% | 2.86% | 3.24% | -0.14% | 0.28% | -0.23% | 0.00% |
|            | $\rho_{mp} = -0.5$ | 9.43% | 8.68% | 2.67% | 3.44% | -1.11% | -0.14% | 0.23% | 0.35% |
|            | $\rho_{mp} = -0.9$ | 8.68% | 4.53% | -0.38% | -0.57% | 1.39% | 1.39% | -0.12% | 0.12% |
| 15% vs 75% | $\rho_{mp} = 0.1$ | 11.70% | 12.08% | 1.72% | 1.72% | 1.25% | 1.53% | -0.12% | 0.00% |
|            | $\rho_{mp} = 0.5$ | -1.13% | -0.38% | 2.67% | 3.63% | -0.28% | 0.14% | -0.12% | 0.00% |
|            | $\rho_{mp} = 0.9$ | 7.55% | 10.19% | -3.82% | -1.53% | -0.56% | 0.56% | -0.93% | -0.46% |
|            | $\rho_{mp} = -0.1$ | 0.00% | 0.38% | 2.29% | 2.10% | -0.14% | 0.14% | 0.00% | 0.12% |
|            | $\rho_{mp} = -0.5$ | 12.83% | 10.57% | 2.48% | 2.67% | 0.56% | 0.97% | -0.35% | -0.23% |
|            | $\rho_{mp} = -0.9$ | 8.30% | 7.17% | -1.53% | 0.00% | 0.83% | 1.53% | -0.81% | -0.35% |