

Article

Not peer-reviewed version

---

# Structural Equation Modeling Techniques for Estimating Score Dependability within Generalizability Theory-Based Univariate, Multivariate, and Bifactor Designs

---

[Walter P. Vispoel](#)\*, [Hyeryung Lee](#), [Tingting Chen](#)

Posted Date: 10 January 2025

doi: 10.20944/preprints202501.0761.v1

Keywords: structural equation models; generalizability theory; multivariate analysis; bifactor models; R programming; reliability; validity; subscale viability; Music Self-Perception Inventory; mGENOVA



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

## Article

# Structural Equation Modeling Techniques for Estimating Score Dependability within Generalizability Theory-Based Univariate, Multivariate, and Bifactor Designs

Walter P. Vispoel \*, Hyeryung Lee, Tingting Chen

\* Correspondence: walter-vispoel@uiowa.edu

**Abstract:** Generalizability theory (GT) provides an all-encompassing framework for estimating accuracy of scores and effects of multiple sources of measurement error when using measures intended for either norm or criterion referencing purposes. Structural equation models (SEMs) can reproduce results obtained from GT-based analysis-of-variance (ANOVA) procedures while further extending those procedures to correct for scale coarseness, derive Monte Carlo based confidence intervals for key parameters, separate universe score variance into general and group factor effects, and determine subscale score viability. We demonstrate how to apply these techniques in R to univariate, multivariate, and bifactor designs using a novel indicator-mean approach to estimating absolute error. When representing responses to items from the Music Self-Perception Inventory (MUSPI-S) using two-, four-, and eight-point response metrics over two occasions, SEMs accurately reproduced results from the ANOVA-based *mGENOVA* package for univariate and multivariate designs and yielded score accuracy and subscale viability indices within bifactor designs comparable to those from corresponding multivariate designs. Corrections for scale coarseness improved score accuracy on all response metrics but to a greater extent with dichotomously scored items. Despite the dominance of general-factor effects, subscale viability was supported in all instances with transient measurement error leading to the greatest reductions in score accuracy.

**Keywords:** structural equation models; generalizability theory; multivariate analysis; bifactor models; R programming; reliability; validity; subscale viability; Music Self-Perception Inventory; *mGENOVA*

## 1. Introduction

Although originally developed during the 1960s [1-3], generalizability theory (GT) continues to be used across numerous disciplines, in large part, because it can be applied to both objectively and subjectively scored measures, quantify effects of multiple sources of measurement error, and produce a wide variety of coefficients to assess the accuracy of observed scores for both norm- and criterion-referencing purposes. Introductions to doing traditional analysis of variance (ANOVA)-based GT analyses can be found in full-length books devoted exclusively to the topic [4-9]; chapters within measurement textbooks [10-12], research handbooks [13-16], encyclopedias [17-23], and edited volumes [25-39]; as well as articles or tutorials within professional journals [1-3, 40-64]. Examples of content areas to which such analyses have been conducted over the last five years alone include medical education and training [65-71], radiology [72], rehabilitation [73], nursing, [74] pharmacology [75], K-12 writing skills [76-79], second language education [80-82], speech and hearing research [83-84], thinking skills and creativity [85], psychiatry and psychology [86-90], sports [91], among others.

Univariate ANOVA-based GT analyses can be run using variance component programs within comprehensive statistical packages such as *SPSS*, *SAS*, *R*, *STATA*, *MATLAB*, and *Minitab* (see, e.g.,

[92]) or standalone programs devoted exclusively to those purposes, which include the *GENOVA* suite (*GENOVA* [93], *urGENOVA* [94], and *mGENOVA* [95]), *G\_string IV* (manual available at <https://www.papaworx.com/HOP/Manual.pdf>), *EduG* [9], and the *gtheory* package in R [96-97; also see 62]). Most recently, applications of GT have been enhanced by conducting such analyses within structural equation model (SEM) frameworks (see, e.g., [59-60, 62, 64, 98-111]). GT-based SEMs can be analyzed using numerous readily accessible statistical packages and provide effective methods for incorporating univariate, multivariate, and bifactor model designs, deriving confidence intervals for key parameters, adjusting for scale coarseness effects common when using binary or ordinal data, assessing scale viability, and handling missing data.

The most common indices of score accuracy reported in GT analyses include generalizability coefficients for norm-referencing purposes (e.g., rank ordering scores) and global and cut-score-specific dependability coefficients for criterion-referencing purposes (e.g., using absolute levels of scores for decision making). Initial applications of SEMs were limited to derivation of generalizability (G) coefficients within univariate designs [112-113] but later modified to allow for computation of dependability (D) coefficients [64, 98, 103, 114-115] and analysis of multivariate [64, 108-111] and bifactor designs [64, 101-104, 108-109, 111]. Whenever composite scores are reported in addition to subscale scores in practice, multivariate and bifactor GT designs produce more appropriate indices of score accuracy than would a univariate design for the composite because they take subscale score representation and interrelationships into account [2, 4, 8, 64, 102, 111].

In a recent study by Lee and Vispoel [107]), an indicator mean-based procedure within SEMs was introduced to derive absolute error indices needed to estimate D coefficients for univariate designs that produced comparable or better results than previous methods for deriving such indices. Our primary goals in the study reported here are to replicate previous uses of SEMs within GT univariate designs using a new instrument and to expand upon such research by applying the indicator-mean method to derivation of absolute error indices for multivariate and bifactor designs based on original observed scores and ones corrected for scale coarseness effects. In doing so, the methods discussed here serve as guides to conducting complete GT analyses for common multivariate and bifactor designs with univariate analyses for individual subscales subsumed within those frameworks. We further provide computer code in R to enable readers to apply all illustrated techniques to their own data.

## 2. Background

### 2.1. GT designs

We will apply SEM procedures for analyzing *persons*  $\times$  *items* (*pi*) single-facet and *persons*  $\times$  *items*  $\times$  *occasions* (*pio*) two-facet GT designs based on responses to a self-report measure (i.e., the shortened form of the Music Self-Perception Inventory (MUSPI-S) [116-119]). Each facet of measurement (items and occasions here) represents a domain to which results are generalized. Partitioning of observed score variance at the individual score level for these designs is described in Equations (1) and (2).

$$p \times i \text{ (pi) design: } \sigma_{Y_{pi}}^2 = \sigma_p^2 + \sigma_{pi,e}^2 + \sigma_i^2. \quad (1)$$

$$p \times i \times o \text{ (pio) design: } \sigma_{Y_{pio}}^2 = \sigma_p^2 + \sigma_{pi}^2 + \sigma_{po}^2 + \sigma_{pio,e}^2 + \sigma_i^2 + \sigma_o^2 + \sigma_{io}^2. \quad (2)$$

In the *pi* design, variance across all observed scores is partitioned into three components that represent persons ( $\sigma_p^2$ ), items ( $\sigma_i^2$ ), and the interaction between persons and items ( $\sigma_{pi,e}^2$ ). In the *pio* design, observed score variance is partitioned into seven components that represent persons ( $\sigma_p^2$ ), items ( $\sigma_i^2$ ), occasions ( $\sigma_o^2$ ) and all possible interactions among persons, items, and occasions ( $\sigma_{pi}^2, \sigma_{po}^2, \sigma_{pio,e}^2, \sigma_{io}^2$ ). The variance component for persons ( $\sigma_p^2$ ) in both designs represents *universe score* variance that parallels true score variance in classical test theory and communality in factor analysis. Interactions between persons and the measurement facets items and/or occasions represent sources of relative measurement error. The *pi* design includes only a single source of relative

measurement error ( $\sigma_{pi,e}^2$ ) in contrast to the *pio* design that includes multiple sources of such error ( $\sigma_{pi}^2, \sigma_{po}^2, \sigma_{pio,e}^2$ ). The subscript “*e*” within a variance component term indicates that it also includes any remaining relative residual error not captured by other terms in the given design. Variance components involving persons are used to derive G coefficients for norm-referencing purposes, whereas those not involving persons reflect differences in mean scores for facet conditions within the design relevant for criterion-referencing purposes in which absolute score values are used for decision making. These components are combined with those for relative error to reflect absolute or total overall error when deriving D coefficients.

As already noted, GT-based analyses produce three primary indices of score accuracy that are represented in Equations (3)-(5): generalizability (G or  $E_p^2$ ), global dependability (D or  $\phi$ ), and cut-score-specific dependability. We provide more detailed formulas to estimate these coefficients and related variance components within tables presented in later sections.

$$\text{G coefficient} = \frac{\text{Universe score variance}}{\text{Universe score variance} + \text{Relative error variance}} \quad (3)$$

$$\text{Global D coefficient} = \frac{\text{Universe score variance}}{\text{Universe score variance} + \text{Absolute error variance}} \quad (4)$$

$$\text{Cut-score-specific D coefficient} = \frac{\text{Universe score variance} + (\mu_Y - \text{cut score})^2}{\text{Universe score variance} + (\mu_Y - \text{cut score})^2 + \text{Absolute error variance}} \quad (5)$$

G coefficients are similar and sometimes identical to conventional alpha, split-half, parallel form, and test-retest reliability estimates in that they reflect relative differences in scores across persons (see [59] for further details) but are interpreted in relation to all possible facet conditions (items, occasion, raters, etc.) within the targeted assessment domains of interest. Equation (3) for a G coefficient would represent universe score (or person) variance divided by the sum of person variance and all error variance components involving persons (i. e.,  $\sigma_p^2$  and  $\sigma_{pi,e}^2$  in the *pi* design and  $\sigma_p^2, \sigma_{pi}^2, \sigma_{po}^2$ , and  $\sigma_{pio,e}^2$  in the *pio* design). Equation (4) for global dependability resembles Equation (3), but with all variance components for facets and their interactions (i. e.,  $\sigma_i^2$  in the *pi* design and  $\sigma_i^2, \sigma_o^2$ , and  $\sigma_{io}^2$  in the *pio* design) combined with relative error variance components to represent absolute error in the denominator of the equation. Consequently, global D coefficients can be no larger than G coefficients and equal to them only when all facet condition means are identical.

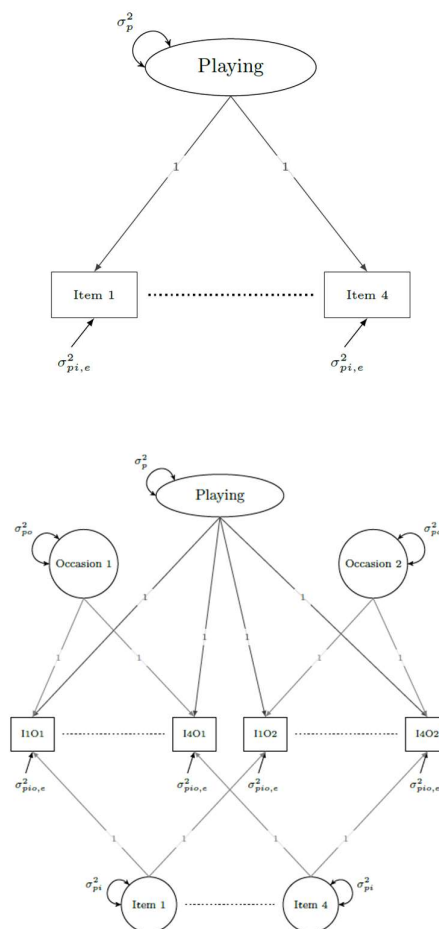
Global D coefficients broaden the conceptualization of measurement error to include mean differences in scores and thereby reflect the contribution of the assessment procedure to overall dependability when making criterion-referenced interpretations of scores [25, 28, 114-115]. Finally, cut-score-specific dependability represented in Equation (5) parallels Equation (4), but with the squared difference between the grand score mean and cut score added to both the numerator and denominator of the equation. Accordingly, the value of this coefficient can change depending on the value of the cut score and represents dependability specific to that cut score. Conceptually, cut-score-specific D coefficients reflect the contribution of the assessment procedure to the decision made from the cut score over what would be expected by chance agreement [25, 28, 114-115]. These coefficients are especially useful for gauging accuracy in determining whether an individual's standing truly falls above or below the targeted cut score. Like conventional reliability estimates, G and D coefficients can vary from 0 to 1 with higher values representing greater accuracy in scores for their intended purposes.

## 2.2. Representing univariate, multivariate, and bifactor GT designs within SEMs

In the data analyses presented here, we focus on univariate, multivariate, and bifactor *pi* designs based on a single occasion of administration in which items are the sole measurement facet of interest, and univariate, multivariate, and bifactor *pio* designs based on two occasions of administration with both items and occasions serving as measurement facets. Our illustrations represent data collected using the Instrument Playing, Reading Music, Listening, and Composing subscales from the shortened form of the Music Self-Perception Inventory (MUSPI-S; [116-119]). Each subscale consists of four items with item scores summed across all subscales to create a composite score that is used here to represent overall perceptions of music proficiency within the multivariate and bifactor designs.

## 2.3. Univariate GT designs

**GT *persons* × *items* (*pi*) single-facet univariate designs.** In Figure 1, we depict SEM diagrams for univariate *pi* and *pio* designs. The top diagram in Figure 1 represents a *pi* SEM for the Instrument Playing subscale. The SEM has a single factor for the construct of interest that is linked to all items measuring that construct. Item loadings are set equal to one, and all item uniquenesses are set equal. In total, two parameters are estimated that represent the variance component for person or universe scores ( $\sigma_p^2$ ) and the variance component for relative measurement error across items ( $\sigma_{pi,e}^2$ ).



**Figure 1.** Structural equation models for univariate G-theory *pi* and *pio* designs for the Instrument Playing subscale. *Note.* *p* = Person; *S* = Subscale; *I* = Item; *O* = Occasion;  $\sigma_p^2$  = person, universe score, or trait variance;  $\sigma_{pi}^2$  = specific-factor error variance;  $\sigma_{po}^2$  = transient error variance; and  $\sigma_{pio,e}^2$  = random-response error variance. Numbers within lines represent loadings for items and occasions.



When using the indicator-mean method [107], the remaining variance component for items ( $\sigma_i^2$ ) is estimated using intercepts for each item that are equivalent to their corresponding means. Specifically, the squared differences between each item mean and grand mean across items are summed and divided by the number of items minus one as shown in Table 1. Once the three variance components of interest ( $\sigma_p^2$ ,  $\sigma_{pi,e}^2$ , and  $\sigma_i^2$ ) are estimated. They can be placed in the general equations shown in Table 2 to estimate G, global D, and cut-score-specific D coefficients for the *pi* design.

**Table 1.** Formulas for variance components within GT univariate *pi* and *pio* designs.

| Design/VC                | Formula                                                                                                                                                                                                                                                                                                                                                                                                  |
|--------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <b><i>pi</i> design</b>  |                                                                                                                                                                                                                                                                                                                                                                                                          |
| <i>p</i>                 | $\hat{\sigma}_p^2$                                                                                                                                                                                                                                                                                                                                                                                       |
| <i>pi,e</i>              | $\hat{\sigma}_{pi,e}^2$                                                                                                                                                                                                                                                                                                                                                                                  |
| <i>i</i>                 | $\hat{\sigma}_i^2 = \frac{\sum_{i=1}^{n_I} (\hat{\beta}_i - grand \hat{\mu})^2}{n_I - 1},$ <p>where <math>n_I</math>= the number of items, <math>\hat{\beta}_i</math>= intercept of <math>i_{th}</math> item, and <math>grand \hat{\mu} = \frac{\sum_{i=1}^{n_I} \hat{\beta}_i}{n_I}</math>.</p>                                                                                                         |
| <b><i>pio</i> design</b> |                                                                                                                                                                                                                                                                                                                                                                                                          |
| <i>p</i>                 | $\hat{\sigma}_p^2$                                                                                                                                                                                                                                                                                                                                                                                       |
| <i>pi,e</i>              | $\hat{\sigma}_{pi}^2$                                                                                                                                                                                                                                                                                                                                                                                    |
| <i>po</i>                | $\hat{\sigma}_{po}^2$                                                                                                                                                                                                                                                                                                                                                                                    |
| <i>pio,e</i>             | $\hat{\sigma}_{pio,e}^2$                                                                                                                                                                                                                                                                                                                                                                                 |
| <i>i</i>                 | $\hat{\sigma}_i^2 = \frac{\sum_{i=1}^{n_I} (\hat{\mu}_i - grand \hat{\mu})^2}{n_I - 1},$ <p>where <math>\hat{\beta}_{io}</math>= intercept of <math>i_{th}</math> item on the <math>o_{th}</math> occasion, <math>grand \hat{\mu} = \frac{\sum_{i=1}^{n_I, o=1}^{n_I, o=n_o} \hat{\beta}_{io}}{n_I \times n_o}</math>, and <math>\hat{\mu}_i = \frac{\sum_{o=1}^{n_o} \hat{\beta}_{io}}{n_o}</math>.</p> |
| <i>o</i>                 | $\hat{\sigma}_o^2 = \frac{\sum_{o=1}^{n_o} (\hat{\mu}_o - grand \hat{\mu})^2}{n_o - 1},$ <p>where <math>n_o</math>= the number of occasions, and <math>\hat{\mu}_o = \frac{\sum_{i=1}^{n_I} \hat{\beta}_{io}}{n_I}</math>.</p>                                                                                                                                                                           |
| <i>io</i>                | $\hat{\sigma}_{io}^2 = \frac{\sum_{i=1, o=1}^{n_I, o=n_o} (\hat{\beta}_{io} - \hat{\mu}_i - \hat{\mu}_o + grand \hat{\mu})^2}{(n_I \times n_o) - 1}.$                                                                                                                                                                                                                                                    |

*Note.* GT= generalizability theory. Variance components without formulas come directly from structural equation model computer output (see Figure 1 and supplementary materials).

**Table 2.** Formulas for estimating GT G, global D, and cut-score-specific D coefficients for all designs and models.

| Design/Index                | Formula                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                      |
|-----------------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>pi design</i>            |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| G                           | $\frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi,e}^2}{n'_i}}$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| $\hat{\omega}_{HComposite}$ | $\frac{\hat{\sigma}_{gen}^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi,e}^2}{n'_i}}$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| $\hat{\omega}_{HSubscale}$  | $\frac{\hat{\sigma}_{grp}^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi,e}^2}{n'_i}}$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| G-D                         | $\frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi,e}^2 + \hat{\sigma}_i^2}{n'_i}}$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                          |
| Cut-score-specific D        | $\frac{\hat{\sigma}_p^2 + (\mu_Y - Cut\ score)^2 - \hat{\sigma}_Y^2}{\hat{\sigma}_p^2 + (\mu_Y - Cut\ score)^2 - \hat{\sigma}_Y^2 + \frac{\hat{\sigma}_{pi,e}^2 + \hat{\sigma}_i^2}{n'_i}}, \text{ where } \hat{\sigma}_Y^2$ $= \frac{\hat{\sigma}_p^2}{n'_p} + \frac{\hat{\sigma}_{pi,e}^2}{n'_p n'_i} + \frac{\hat{\sigma}_i^2}{n'_i} \text{ and corrects for bias.}$                                                                                                                                                                                                                                                                                      |
| Total relative error        | $1 - \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi,e}^2}{n'_i}}$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                         |
| <i>pio design</i>           |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                              |
| G                           | $\frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi}^2}{n'_i} + \frac{\hat{\sigma}_{po}^2}{n'_o} + \frac{\hat{\sigma}_{pio,e}^2}{n'_i n'_o}}$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |
| $\hat{\omega}_{HComposite}$ | $\frac{\hat{\sigma}_{gen}^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi}^2}{n'_i} + \frac{\hat{\sigma}_{po}^2}{n'_o} + \frac{\hat{\sigma}_{pio,e}^2}{n'_i n'_o}}$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| $\hat{\omega}_{HSubscale}$  | $\frac{\hat{\sigma}_{grp}^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi}^2}{n'_i} + \frac{\hat{\sigma}_{po}^2}{n'_o} + \frac{\hat{\sigma}_{pio,e}^2}{n'_i n'_o}}$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                             |
| G-D                         | $\frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi}^2 + \hat{\sigma}_i^2}{n'_i} + \frac{\hat{\sigma}_{po}^2 + \hat{\sigma}_o^2}{n'_o} + \frac{\hat{\sigma}_{pio,e}^2 + \hat{\sigma}_{io}^2}{n'_i n'_o}}$                                                                                                                                                                                                                                                                                                                                                                                                                                     |
| Cut-score-specific D        | $\frac{\hat{\sigma}_p^2 + (\mu_Y - Cut\ score)^2 - \hat{\sigma}_Y^2}{\hat{\sigma}_p^2 + (\mu_Y - Cut\ score)^2 - \hat{\sigma}_Y^2 + \frac{\hat{\sigma}_{pi}^2 + \hat{\sigma}_i^2}{n'_i} + \frac{\hat{\sigma}_{po}^2 + \hat{\sigma}_o^2}{n'_o} + \frac{\hat{\sigma}_{pio,e}^2 + \hat{\sigma}_{io}^2}{n'_i n'_o}},$ $\text{ where } \hat{\sigma}_Y^2 = \frac{\hat{\sigma}_p^2}{n'_p} + \frac{\hat{\sigma}_{pi}^2}{n'_p n'_i} + \frac{\hat{\sigma}_{po}^2}{n'_p n'_o} + \frac{\hat{\sigma}_{pio,e}^2}{n'_p n'_i n'_o} + \frac{\hat{\sigma}_i^2}{n'_i} + \frac{\hat{\sigma}_o^2}{n'_o}$ $+ \frac{\hat{\sigma}_{io}^2}{n'_i n'_o} \text{ and corrects for bias.}$ |
| Specific-factor error       | $\frac{\frac{\hat{\sigma}_{pi}^2}{n'_i}}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi}^2}{n'_i} + \frac{\hat{\sigma}_{po}^2}{n'_o} + \frac{\hat{\sigma}_{pio,e}^2}{n'_i n'_o}}$                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 |

|                       |                                                                                                                                                                                      |
|-----------------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Transient error       | $\frac{\frac{\hat{\sigma}_{po}^2}{n'_o}}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi}^2}{n'_i} + \frac{\hat{\sigma}_{po}^2}{n'_o} + \frac{\hat{\sigma}_{pio,e}^2}{n'_i n'_o}}$         |
| Random-response error | $\frac{\frac{\hat{\sigma}_{pio,e}^2}{n'_i n'_o}}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi}^2}{n'_i} + \frac{\hat{\sigma}_{po}^2}{n'_o} + \frac{\hat{\sigma}_{pio,e}^2}{n'_i n'_o}}$ |
| Total relative error  | $1 - \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi}^2}{n'_i} + \frac{\hat{\sigma}_{po}^2}{n'_o} + \frac{\hat{\sigma}_{pio,e}^2}{n'_i n'_o}}$                     |

Note.  $pi$  = persons  $\times$  items design,  $pio$  = persons  $\times$  items  $\times$  occasions design, GT= generalizability theory,  $G$  = generalizability coefficient,  $D$  = dependability coefficient,  $\hat{\omega}_{H_{Composite}}$  = Omega hierarchical composite coefficient,  $\hat{\omega}_{H_{Subscale}}$  = Omega hierarchical subscale coefficient. Generalizability coefficients for the bifactor model are equivalent to omega total coefficients.

**GT persons  $\times$  items  $\times$  occasions ( $pio$ ) two-facet univariate designs.** The bottom diagram in Figure 1 represents a SEM for the MUSPI-S's Instrument Playing subscale for a  $pio$  design with four items administered on two occasions. When measuring psychological traits, the  $pio$  design is generally preferred over the  $pi$  design because it allows for separation of three key sources of measurement error and reduces confounding of trait and measurement error variance typically found in the  $pi$  design. Within the  $pio$  SEM shown in Figure 1, the factor for the construct of interest is linked to each item on each occasion. Separate factors are included for each item across all occasions and for each occasion across all items. All factor loadings are set equal to one, with item variances set equal, occasion variances set equal, and uniquenesses set equal. Collectively, these constraints result in estimation of four parameters to represent the variance of person or universe scores for the construct of interest ( $\sigma_p^2$ ) and three sources of relative measurement error ( $\sigma_{pi}^2$ ,  $\sigma_{po}^2$ , and  $\sigma_{pio,e}^2$ ) common when using objectively scored measures such as Likert-style questionnaires or multiple-choice tests in which all scorers would obtain the same results.

Within the psychological research literature,  $\sigma_{pi}^2$ ,  $\sigma_{po}^2$ , and  $\sigma_{pio,e}^2$  are often respectively referred as specific-factor error (or method effects), transient error (or state effects), and random-response error (or within-occasion “noise”; [120-122], also see [123]). Specific-factor error represents enduring person-specific effects on scores that are unrelated to the construct of interest such as understandings of words within items and response options. Transient error represents consistent effects on scores within a given occasion that do not generalize across occasions such as a respondent’s disposition, mindset, and physiological condition at that time as well as his or her reactions to environmental and administration conditions. Random-response error refers to fleeting motion-to-moment effects within an occasion such as momentary lapses of attention, distractions, and other effects that follow no systematic pattern. Within  $pi$  designs, universe score and transient error are confounded within the person variance component ( $\sigma_p^2$ ), as are specific-factor and random-response error within the relative measurement error component ( $\sigma_{pi,e}^2$ ). An important advantage of the  $pio$  design is that these sources of variance can be separated to provide more appropriate estimates of score accuracy and overall measurement error. Variance component formulas for each source of measurement error are provided in Table 1.

When using the indicator-mean method to derive the remaining variance components ( $\sigma_i^2$ ,  $\sigma_o^2$ , and  $\sigma_{io}^2$ ) within the  $pio$  design, intercepts will represent means for all combinations of items and occasions. Once estimated, these means can be integrated into the formulas shown in Table 1 to obtain the remaining variance components and insert them into the formulas shown in Table 2 to estimate related indices of score dependability and proportions of measurement error (see [107] for further details).



### 2.3. Multivariate GT designs

**GT persons  $\times$  items (pi) single-facet multivariate design.** Multivariate GT designs best represent indices of score accuracy when both subscale and composite scores are reported in practice. Such designs also can produce correlation coefficients corrected for all sources of measurement error included within a design to provide further insights into subscale score dimensionality, overlap, interrelationships, and validity. Embedded within the overall multivariate design are the same univariate analyses already described for each individual subscale. Variance components for the composite score, in contrast, entail formulas based on the variance components for each subscale, the covariances between each pair of subscale scores, and eventual weighting of each subscale when forming the composite (see Table 3 and [8, 64, 109]).

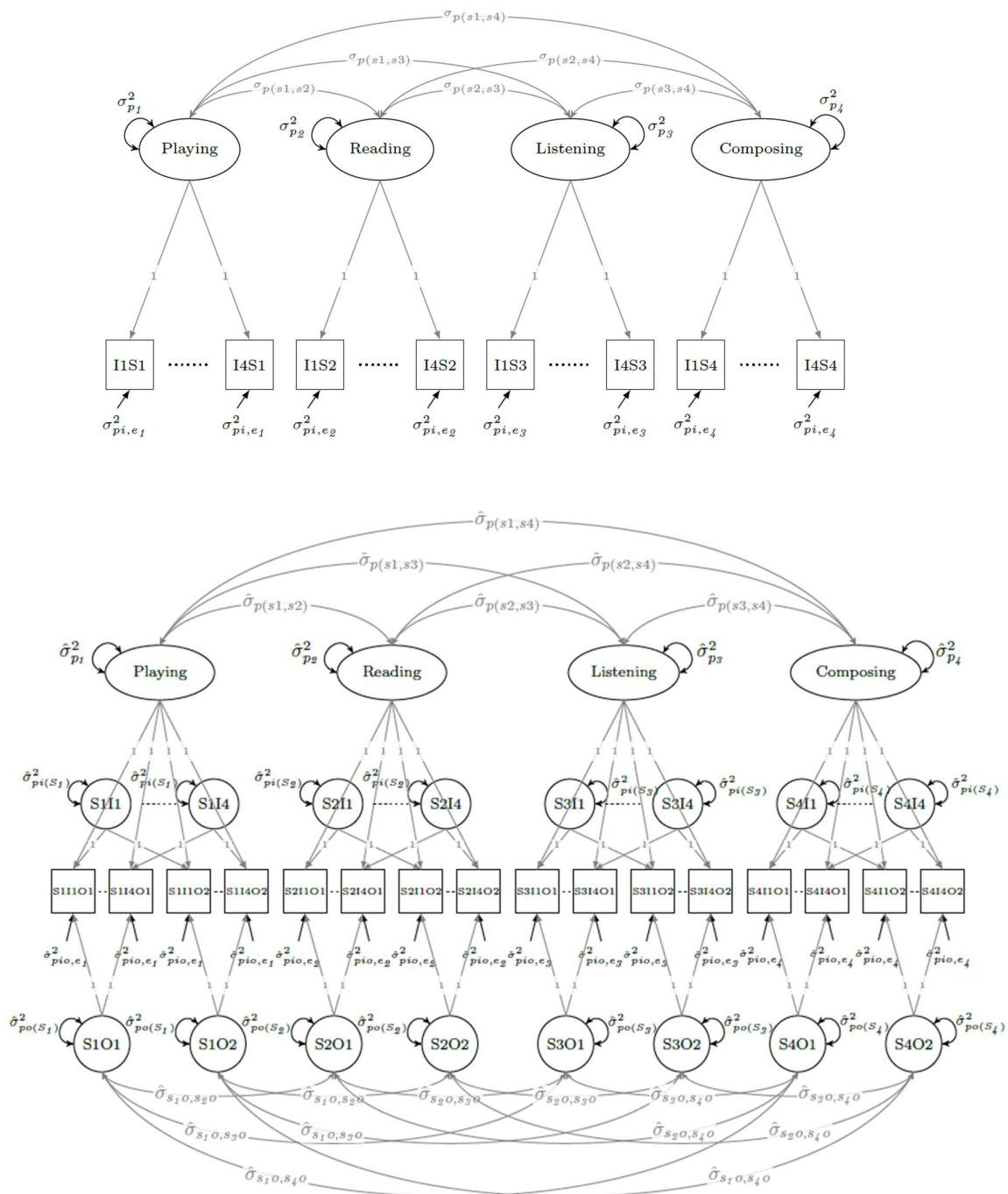
**Table 3.** Formulas for variance components within GT *persons  $\times$  items (pi)* and *persons  $\times$  items  $\times$  occasions (pio)* multivariate designs.

| Design/<br>VC     | Formula                                                                                                                                                                                                                                                                                                                                  |                                                                                                       |
|-------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|
|                   | Composite                                                                                                                                                                                                                                                                                                                                | Subscale                                                                                              |
| <b>pi design</b>  |                                                                                                                                                                                                                                                                                                                                          |                                                                                                       |
| <i>p</i>          | $\hat{\sigma}_{pc}^2 = \sum_{s=1}^{n_s} \hat{\sigma}_{ps}^2 + \sum_{s1=1}^{n_{s1}} \sum_{s2 \neq s1}^{n_{s2}} \hat{\sigma}_{p(s1,s2)},$<br>where $n_s$ = the number of subscales                                                                                                                                                         | $\hat{\sigma}_{ps}^2$                                                                                 |
| <i>pi,e</i>       | $\hat{\sigma}_{pi,e}^2 = \sum_{s=1}^{n_s} \hat{\sigma}_{pi,es}^2$                                                                                                                                                                                                                                                                        | $\hat{\sigma}_{pi,es}^2$                                                                              |
| <i>i</i>          | $\hat{\sigma}_{ic}^2 = \sum_{s=1}^{n_s} \left[ \frac{\sum_{i=1}^{n_{is}} (\hat{\beta}_i - grand \hat{\mu}_s)^2}{n_{is} - 1} \right],$<br>where $n_{is}$ = the number of items in subscale S,<br>$\hat{\beta}_i$ = intercept of $i_{th}$ item, and $grand \hat{\mu}_s = \frac{\sum_{i=1}^{n_{is}} \hat{\beta}_i}{n_{is}}$ .               | $\hat{\sigma}_{is}^2 = \frac{\sum_{i=1}^{n_{is}} (\hat{\beta}_i - grand \hat{\mu}_s)^2}{n_{is} - 1}$  |
| <b>pio design</b> |                                                                                                                                                                                                                                                                                                                                          |                                                                                                       |
| <i>p</i>          | $\hat{\sigma}_{pc}^2 = \sum_{s=1}^{n_s} \hat{\sigma}_{ps}^2 + \sum_{s1=1}^{n_{s1}} \sum_{s2 \neq s1}^{n_{s2}} \hat{\sigma}_{p(s1,s2)}$                                                                                                                                                                                                   | $\hat{\sigma}_{ps}^2$                                                                                 |
| <i>pi,e</i>       | $\hat{\sigma}_{pic}^2 = \sum_{s=1}^{n_s} \hat{\sigma}_{pis}^2$                                                                                                                                                                                                                                                                           | $\hat{\sigma}_{pis}^2$                                                                                |
| <i>po</i>         | $\hat{\sigma}_{poc}^2 = \sum_{s=1}^{n_s} \hat{\sigma}_{pos}^2 + \sum_{s1=1}^{n_{s1}} \sum_{s2 \neq s1}^{n_{s2}} \hat{\sigma}_{po(s1,s2)}$                                                                                                                                                                                                | $\hat{\sigma}_{pos}^2$                                                                                |
| <i>pio,e</i>      | $\hat{\sigma}_{pio,e}^2 = \sum_{s=1}^{n_s} \hat{\sigma}_{pio,es}^2$                                                                                                                                                                                                                                                                      | $\hat{\sigma}_{pio,es}^2$                                                                             |
| <i>i</i>          | $\hat{\sigma}_{ic}^2 = \sum_{s=1}^{n_s} \left[ \frac{\sum_{i=1}^{n_{is}} (\hat{\mu}_{si} - grand \hat{\mu}_s)^2}{n_{is} - 1} \right],$<br>where $grand \hat{\mu}_s = \frac{\sum_{i=1, o=1}^{i=n_{is}, o=n_{os}} \hat{\beta}_{io}}{n_{is} \times n_{os}}$<br>and $\hat{\mu}_{si} = \frac{\sum_{o=1}^{n_{os}} \hat{\beta}_{io}}{n_{os}}$ . | $\hat{\sigma}_{is}^2 = \frac{\sum_{i=1}^{n_{is}} (\hat{\mu}_{si} - grand \hat{\mu}_s)^2}{n_{is} - 1}$ |

|    |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                  |                                                                                                        |
|----|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|
| o  | $\hat{\sigma}_{o_c}^2 = \sum_{s=1}^{n_s} \left[ \frac{\sum_{o=1}^{n_{so}} (\hat{\mu}_{so} - grand \hat{\mu}_s)^2}{n_{so} - 1} \right]$ $+ \sum_{s1=1}^{n_{s1}} \sum_{s2 \neq s1}^{n_{s2}} \left[ \frac{\sum_{o=1}^{n_{so}} (\hat{\mu}_{s1o} - grand \hat{\mu}_s)(\hat{\mu}_{s2o} - grand \hat{\mu}_s)}{n_{so} - 1} \right]$ <p>where <math>n_{so}</math> = the number of occasions in subscale S,<br/>and <math>\hat{\mu}_{so} = \frac{\sum_{i=1}^{n_{is}} \hat{\beta}_{io}}{n_{is}}</math>.</p> | $\hat{\sigma}_{o_s}^2 = \frac{\sum_{o=1}^{n_{so}} (\hat{\mu}_{so} - grand \hat{\mu}_s)^2}{n_{so} - 1}$ |
| io | $\hat{\sigma}_{io_c}^2 = \sum_{s=1}^{n_s} \left[ \frac{\sum_{i=1, o=1}^{i=n_{si}, o=n_{so}} (\hat{\beta}_{io} - \hat{\mu}_{si} - \hat{\mu}_{so} + grand \hat{\mu}_s)^2}{(n_{si} \times n_{so}) - 1} \right]$ $\hat{\sigma}_{io_s}^2 = \frac{\sum_{i=1, o=1}^{i=n_{si}, o=n_{so}} (\hat{\beta}_{io} - \hat{\mu}_{si} - \hat{\mu}_{so} + grand \hat{\mu}_s)^2}{(n_{si} \times n_{so}) - 1}$                                                                                                        |                                                                                                        |

Note. GT= generalizability theory. VC = variance component. Variance components without formulas above come directly from structural equation model computer output (see Figure 2 and supplementary materials).

Figure 2 includes SEM diagrams for multivariate *pi* and *pio* designs that can be used to derive variance and covariance components for computing G coefficients for subscale and composite scores. The diagrams respectively represent the 4-item subscales from the MUSPI-S (Instrument Playing, Reading Music, Listening Skill, and Composing Ability) mentioned earlier administered on one or two occasions. Scores for each individual subscale are modeled and constrained in the same way as in a univariate analysis but allowed to covary/correlate with each other. Within the *pi* design, eight variance components ( $\sigma_p^2$  and  $\sigma_{pi,e}^2$  for each subscale) and six covariance components (one for each possible pair of subscale scores) are estimated. The variance component for items ( $\sigma_i^2$ ) within each subscale is computed in the same way as described for the univariate design.  $\sigma_p^2$ ,  $\sigma_{pi,e}^2$ , and  $\sigma_i^2$  indices for the composite score can be estimated using the formulas shown in Table 3 and corresponding generalizability and dependability coefficients using the formulas shown in Table 2.



**Figure 2.** Structural equation models for multivariate G-theory  $\pi_i$  and  $\pi_{io}$  designs for MUSPI-S subscale scores. Note.  $p$  = Person;  $S$  = Subscale;  $I$  = Item;  $O$  = Occasion;  $\sigma_p^2$  = person, universe score, or trait variance;  $\sigma_{pi}^2$  = specific-factor error variance;  $\sigma_{po}^2$  = transient error variance; and  $\sigma_{pio,e}^2$  = random-response error variance. Symbols linking subscales at the top of the models and linking occasions at the bottom of the model for  $\pi_{io}$  design represent covariances. Numbers within lines represent loadings for items and occasions.

**GT persons  $\times$  items  $\times$  occasions ( $\pi_{io}$ ) two-facet multivariate design.** The  $\pi_{io}$  univariate design for each embedded subscale within the multivariate design has additional factors for each item across occasions and for each occasion across items. Scores for each pair of subscales again are allowed to

covary/correlate but to the same degree across occasions. Transient error ( $\sigma_{po,}^2$ ) indices also are allowed to covary/correlate when all measures are administered together within a common occasion. In all, sixteen variance components ( $\sigma_p^2, \sigma_{pi}^2, \sigma_{po}^2$ , and  $\sigma_{pio,e}^2$  for each subscale), six covariance components for person subscale scores (one for each possible pair of subscale scores), and six covariances for transient errors ( $\sigma_{pos_1}$ ) are estimated. The  $\sigma_i^2, \sigma_o^2$ , and  $\sigma_{io}^2$  variance components for each subscale can be derived in the same ways described in the univariate designs, and those for the composite score using the formulas provided in Table 3. Formulas for deriving relevant generalizability and dependability coefficients for both subscale and composites scores are given in Table 2.

**Correcting correlation coefficients for measurement error.** In addition to providing appropriate indices of accuracy for composite and subscale scores, multivariate designs can yield correlations between all pairs of subscale scores corrected for the sources of measurement error estimated within the design. Corrected correlations can be conceptualized in relation to the formula first proposed by Spearman ([124], also see [125]), shown in Equation (6), in which the correlation coefficient between observed scores for the pair of measures of interest is divided by the square root of the product of their corresponding reliability coefficients to estimate the correlation between true scores for the targeted measures that is free of measurement error. In applications of GT, G coefficients would be substituted for conventional reliability coefficients, and universe scores for true scores.

$$\hat{\rho}_{T_xT_y} = \frac{r_{xy}}{\sqrt{r_{xx'} * r_{yy'}}}, \tag{6}$$

where  $\hat{\rho}_{T_xT_y}$  = estimated correlation between true scores for measures  $X$  and  $Y$ ,

$r_{xy}$  = observed correlation coefficient between measures  $X$  and  $Y$ ,  
 $r_{xx'}$  = reliability coefficient for measure  $X$ ,  
and  $r_{yy'}$  = reliability coefficient for measure  $Y$ .

2.4. Bifactor GT designs

Bifactor and multivariate GT designs both can be used to simultaneously partition score variance at subscale and composite levels and distinguish multiple sources of measurement error within *pio* designs. However, bifactor designs further allow for partitioning of universe score variance into general and group factor effects to produce indices reflecting just general factor effects, just group factor effects, or both effects combined. General factor effects reflect common explained variance shared across all indicators, whereas group factor effects reflect unique explained variance, unrelated to general factor variance, that is shared by all indicators representing a given subscale.

Bifactor models produce four key coefficients in addition ones already discussed: omega total composite, omega total subscale, omega hierarchical composite, and omega hierarchical subscale [11, 126-130]. *Omega total* coefficients for composite and subscale scores represent proportions of variance accounted for by both general and group factor effects. They parallel overall G coefficients for the *pi* and *pio* univariate and multivariate GT designs, except that universe score variance represents the sum of general and group factor variances (see Table 2). *Omega hierarchical composite* score coefficients represent the proportion of variance accounted for by the *general factor* alone, whereas *omega hierarchical subscale* coefficients represent the proportion of variance accounted for by the *group factor* alone. We provide formulas for estimating variance components in Table 4 that can be inserted into formulas shown in Table 2 to derive G, global D, cut-score-specific D, and omega coefficients for *pi* and *pio* bifactor designs.

**Table 4.** Formulas for variance components within GT *persons × items* (*pi*) and *persons × items × occasions* (*pio*) bifactor designs.

| Design/VC         | Formula                                                                                                                                                                                                                                                                                                                                                                   |                                                                                                       |
|-------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------|
|                   | Composite                                                                                                                                                                                                                                                                                                                                                                 | Subscale                                                                                              |
| <i>pi</i> design  |                                                                                                                                                                                                                                                                                                                                                                           |                                                                                                       |
| <i>General</i>    | $\hat{\sigma}_{gen_c}^2 = \sum_{s=1}^{n_s} \hat{\lambda}_s^2$                                                                                                                                                                                                                                                                                                             | $\hat{\sigma}_{gen_s}^2 = \hat{\lambda}_s^2$                                                          |
| <i>Group</i>      | $\hat{\sigma}_{grp_c}^2 = \sum_{s=1}^{n_s} \hat{\sigma}_{grp_s}^2$                                                                                                                                                                                                                                                                                                        | $\hat{\sigma}_{grp_s}^2$                                                                              |
| <i>p</i>          | $\hat{\sigma}_{p_c}^2 = \sum_{s=1}^{n_s} \left[ \hat{\lambda}_s^2 + \hat{\sigma}_{grp_s}^2 \right]$                                                                                                                                                                                                                                                                       | $\hat{\sigma}_{p_s}^2 = \hat{\lambda}_s^2 + \hat{\sigma}_{grp_s}^2$                                   |
| <i>pi,e</i>       | $\hat{\sigma}_{pi,e_c}^2 = \sum_{s=1}^{n_s} \hat{\sigma}_{pi,e_s}^2$                                                                                                                                                                                                                                                                                                      | $\hat{\sigma}_{pi,e_s}^2$                                                                             |
| <i>i</i>          | $\hat{\sigma}_{i_c}^2 = \sum_{s=1}^{n_s} \left[ \frac{\sum_{i=1}^{n_{IS}} (\hat{\beta}_i - grand \hat{\mu}_s)^2}{n_{IS} - 1} \right],$ <p>where <math>n_{IS}</math>= the number of items in subscale S,<br/><math>\hat{\beta}_i</math>= intercept of <math>i_{th}</math> item, and <math>grand \hat{\mu}_s = \frac{\sum_{i=1}^{n_{IS}} \hat{\beta}_i}{n_{IS}}</math>.</p> | $\hat{\sigma}_{i_c}^2 = \frac{\sum_{i=1}^{n_{IS}} (\hat{\beta}_i - grand \hat{\mu}_s)^2}{n_{IS} - 1}$ |
| <i>pio</i> design |                                                                                                                                                                                                                                                                                                                                                                           |                                                                                                       |
| <i>General</i>    | $\hat{\sigma}_{gen_c}^2 = \sum_{s=1}^{n_s} \hat{\lambda}_s^2$                                                                                                                                                                                                                                                                                                             | $\hat{\sigma}_{gen_s}^2 = \hat{\lambda}_s^2$                                                          |

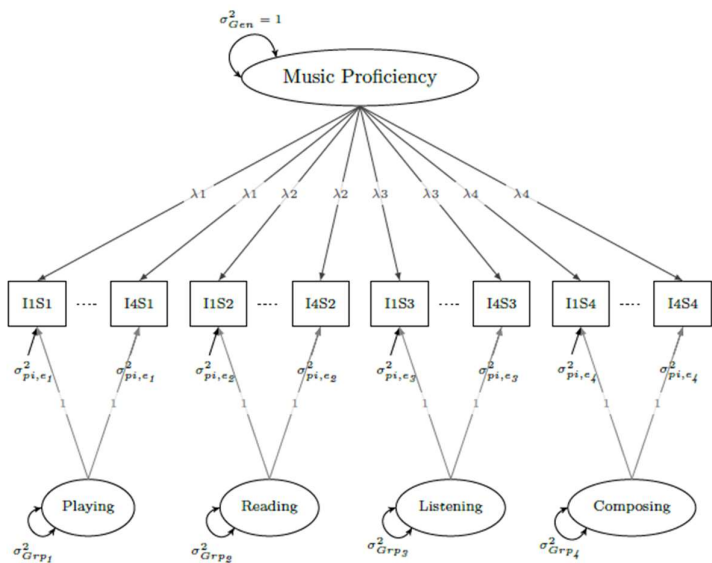
|       |                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                       |                                                                                                        |
|-------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------|
| Group | $\hat{\sigma}_{grp_c}^2 = \sum_{s=1}^{n_s} \hat{\sigma}_{Grp_s}^2$                                                                                                                                                                                                                                                                                                                                                                                                                                                    | $\hat{\sigma}_{Grp_s}^2$                                                                               |
| p     | $\hat{\sigma}_{p_c}^2 = \sum_{s=1}^{n_s} \left[ \hat{\lambda}_s^2 + \hat{\sigma}_{Grp_s}^2 \right]$                                                                                                                                                                                                                                                                                                                                                                                                                   | $\hat{\sigma}_{p_s}^2 = \hat{\lambda}_s^2 + \hat{\sigma}_{Grp_s}^2$                                    |
| pi    | $\hat{\sigma}_{pi_c}^2 = \sum_{s=1}^{n_s} \hat{\sigma}_{pi_s}^2$                                                                                                                                                                                                                                                                                                                                                                                                                                                      | $\hat{\sigma}_{pi_s}^2$                                                                                |
| po    | $\hat{\sigma}_{po_c}^2 = \sum_{s=1}^{n_s} \hat{\sigma}_{po_s}^2 + \sum_{s1=1}^{n_{s1}} \sum_{s2 \neq s1}^{n_{s2}} \hat{\sigma}_{po(s1,s2)}$                                                                                                                                                                                                                                                                                                                                                                           | $\hat{\sigma}_{po_s}^2$                                                                                |
| pio,e | $\hat{\sigma}_{pio,e_c}^2 = \sum_{s=1}^{n_s} \hat{\sigma}_{pio,e_s}^2$                                                                                                                                                                                                                                                                                                                                                                                                                                                | $\hat{\sigma}_{pio,e_s}^2$                                                                             |
| i     | $\hat{\sigma}_{i_c}^2 = \sum_{s=1}^{n_s} \left[ \frac{\sum_{i=1}^{n_{IS}} (\hat{\mu}_{si} - grand \hat{\mu}_s)^2}{n_{IS} - 1} \right],$ <p>where <math>grand \hat{\mu}_s = \frac{\sum_{i=1, o=1}^{i=n_{IS}, o=n_{OS}} \hat{\beta}_{io}}{n_{IS} \times n_{OS}}</math></p> <p>and <math>\hat{\mu}_{si} = \frac{\sum_{o=1}^{n_{OS}} \hat{\beta}_{io}}{n_{OS}}</math>.</p>                                                                                                                                                | $\hat{\sigma}_{i_s}^2 = \frac{\sum_{i=1}^{n_{IS}} (\hat{\mu}_{si} - grand \hat{\mu}_s)^2}{n_{IS} - 1}$ |
| o     | $\hat{\sigma}_{o_c}^2 = \sum_{s=1}^{n_s} \left[ \frac{\sum_{o=1}^{n_{SO}} (\hat{\mu}_{so} - grand \hat{\mu}_s)^2}{n_{SO} - 1} \right]$ <p>+ <math>\sum_{s1=1}^{n_{s1}} \sum_{s2 \neq s1}^{n_{s2}} \left[ \frac{\sum_{o=1}^{n_{SO}} (\hat{\mu}_{s1o} - grand \hat{\mu}_s)(\hat{\mu}_{s2o} - grand \hat{\mu}_s)}{n_{SO} - 1} \right],</math></p> <p>where <math>n_{SO}</math>= the number of occasions in subscale S,</p> <p>and <math>\hat{\mu}_{so} = \frac{\sum_{i=1}^{n_{IS}} \hat{\beta}_{io}}{n_{IS}}</math>.</p> | $\hat{\sigma}_{o_s}^2 = \frac{\sum_{o=1}^{n_{SO}} (\hat{\mu}_{so} - grand \hat{\mu}_s)^2}{n_{SO} - 1}$ |

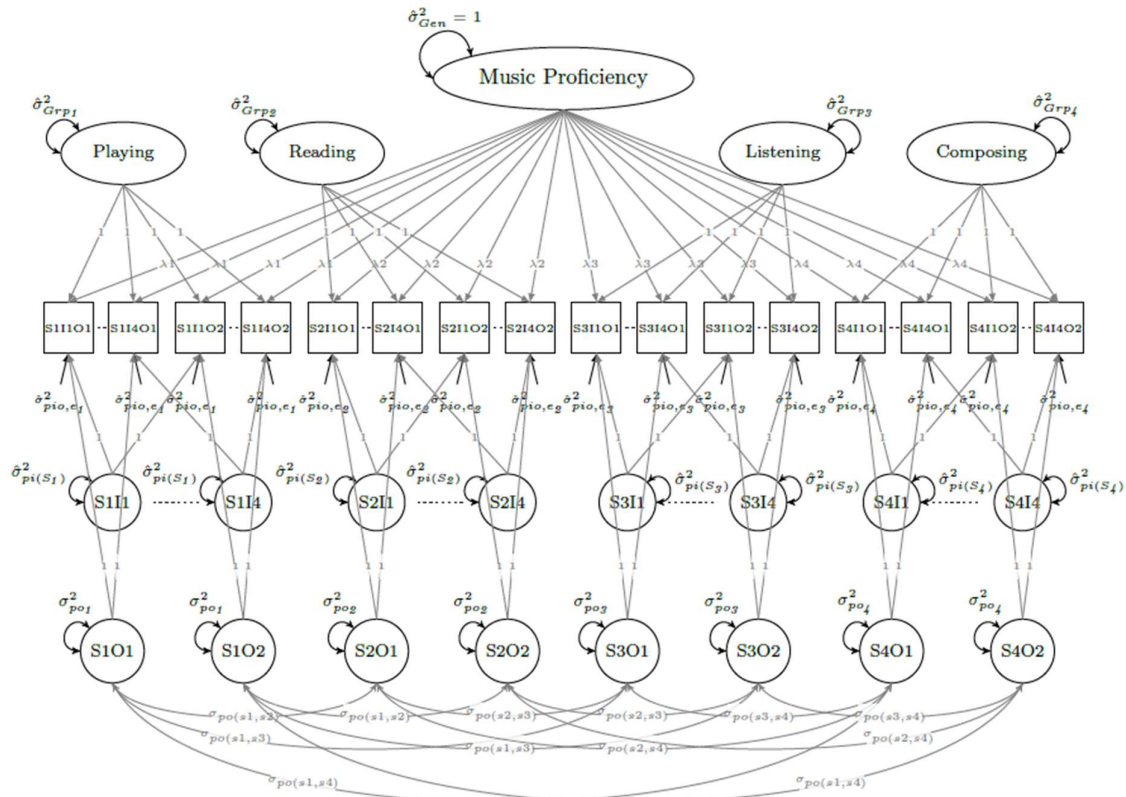


|           |                                                                                                                                                                                                              |                                                                                                                                                                              |
|-----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <i>io</i> | $\hat{\sigma}_{io_c}^2 = \sum_{s=1}^{n_s} \left[ \frac{\sum_{i=1, o=1}^{i=n_{SI}, o=n_{SO}} (\hat{\beta}_{io} - \hat{\mu}_{Si} - \hat{\mu}_{So} + grand \hat{\mu}_s)^2}{(n_{SI} \times n_{SO}) - 1} \right]$ | $\hat{\sigma}_{io_s}^2 = \frac{\sum_{i=1, o=1}^{i=n_{SI}, o=n_{SO}} (\hat{\beta}_{io} - \hat{\mu}_{Si} - \hat{\mu}_{So} + grand \hat{\mu}_s)^2}{(n_{SI} \times n_{SO}) - 1}$ |
|-----------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

Note. GT= generalizability theory, VC = variance component. Variance components without formulas come directly from structural equation model computer output (see Figure 3 and supplementary materials).

**GT *persons* × *items* (*pi*) single-facet bifactor design.** The *pi* bifactor SEM representing MUSPIS scores is shown in the top diagram within Figure 3. The general factor is linked to all items, with independent group factors linked only to items included within each subscale. To allow for differential general factor effects across subscales, model identification constraints differ somewhat from those in the previous designs. Specifically, variances for the general factor and loadings for the group factors are set equal to one, and general factor loadings, group factor variances, and uniquenesses are estimated but set equal within but not across subscales. In all, twelve parameters are estimated ( $\lambda$ ,  $\sigma_{grp}^2$  &  $\sigma_{pi,e}^2$  for each subscale).





**Figure 3.** Structural equation models for bifactor G-theory  $pi$  and  $pio$  designs for MUSPI-S scores. Note.  $p$  = Person;  $S$  = Subscale;  $I$  = Item;  $O$  = Occasion;  $\sigma_p^2$  = person, universe score, or trait variance;  $\sigma_{pi}^2$  = specific-factor error variance;  $\sigma_{po}^2$  = transient error variance; and  $\sigma_{pio,e}^2$  = random-response error variance. Symbols linking occasions at the bottom of the model for the  $pio$  design represent covariances. Symbols or numbers within lines represent loadings for items or occasions.

**GT persons  $\times$  items  $\times$  occasions ( $pio$ ) two-facet bifactor design.** The bottom diagram in Figure 3 represents the  $pio$  bifactor design. It has the same constraints as the  $pi$  design, but with additional factors included for occasions and items. Item and occasion factor loadings are set equal to one, with item factor variances, occasion factor variances, and uniquenesses estimated. As in the multivariate  $pio$  design, separate occasion factors are included for each subscale that are allowed to covary/correlate with each other but to the same degree across occasions. Five parameters are estimated for each subscale ( $\lambda$ ,  $\sigma_{grp}^2$ ,  $\sigma_{pi}^2$ ,  $\sigma_{po}^2$ , &  $\sigma_{pio,e}^2$ ) as well as six additional covariances to model possible correlated within-occasion transient error effects.

## 2.5. Using GT multivariate and bifactor designs to evaluate subscale viability.

An important question to consider whenever using measures that produce both composite and subscale scores in practice is the extent to which subscale scores yield useful information beyond the composite score. To address this question, Haberman ([131], also see [132-135]) devised a classical test theory-based procedure to determine whether a subscale's true scores are better estimated using subscale or composite observed scores. Vispoel and colleagues [64, 103-104, 110-111] later adapted this procedure to single- and multi-facet GT multivariate and bifactor designs by replacing true score with universe score estimation.

Haberman's method is based on comparison of indices for subscale and composite scores reflecting proportional reduction in mean-squared error (PRMSE). The PRMSE for the subscale is

equivalent to its conventional reliability or GT-based generalizability coefficient, whereas the PRMSE(C) for the composite can be derived using Equation 7.

$$PRMSE(C) = r_{T_{S_j}, T_C}^2 * r_{X_C, X_C'} = \frac{\hat{\sigma}_{T_{S_j}, T_C}^2}{\hat{\sigma}_{T_{S_j}}^2 * \hat{\sigma}_{T_C}^2} * \frac{\hat{\sigma}_{X_C}^2}{\hat{\sigma}_{X_C}^2} = \frac{(\hat{\sigma}_{T_{S_j}}^2 + \sum_{j \neq k} \hat{\sigma}_{T_{S_j}, T_{S_k}})^2}{\hat{\sigma}_{T_{S_j}}^2 * \hat{\sigma}_{X_C}^2}, \quad (7)$$

where  $T$  = true score,  $X$  = observed score,  $S$  = subscale,  $C$  = composite score, and  $r_{X_C, X_C'}$  = composite reliability.

Conceptually, a PRMSE index represents an estimate of the proportion of true or universe score variance in the present context accounted for by the targeted observed scores (subscale or composite). Once PRMSEs are obtained for a subscale and its associated composite, they can be inserted into Equation 8 to form a value-added ratio (VAR; see [136]). Subscale viability is increasingly supported as VARs deviate upwardly from 1.00.

$$\text{Value-Added Ratio (VAR)} = \frac{PRMSE(\text{Subscale})}{PRMSE(\text{Composite})} \quad (8)$$

## 2.6. Comparing GT univariate, multivariate, and bifactor designs.

As previously noted, univariate and multivariate GT designs will produce the same results for individual subscales because univariate designs for subscales are embedded within the overall multivariate design. However, two additional benefits of multivariate over univariate designs described earlier are that they produce correlation coefficients between all pairs of subscale scores corrected for the sources of measurement error estimated within the design and yield more appropriate indices of generalizability and dependability for composite scores by taking subscale representation and interrelationships into account. Common findings across recent studies include stronger relationships between subscale scores when corrected for measurement error, and G and D coefficients for composite scores within multivariate designs that generally exceed those derived strictly from univariate designs in which subscale representation and interrelationships are ignored [64, 110-111].

Either GT multivariate or bifactor designs can produce appropriate G and D coefficients at both composite and subscale levels as well as VARs for all subscale scores. In recent studies of personality constructs, GT multivariate and bifactor designs have produced highly comparable G coefficients, D coefficients, and subscale VARs [64, 102, 111] but with subscale and composite scores partitioned into general and group factor effects within the bifactor designs to provide additional insights into score dimensionality and overlap among constructs. In the vast majority of bifactor model studies, proportions of general factor exceed proportions of group factor variance at both composite and subscale levels (see, e.g., [101-102, 111, 128]).

## 2.7. Further advantages of using SEMs to perform GT analyses

Two additional benefits of conducting GT analyses using SEMs to be demonstrated here are to derive Monte Carlo-based confidence intervals for G, D, and omega coefficients and to use estimation procedures that correct for scale coarseness effects commonly encountered when analyzing dichotomous or ordinal-level data. When doing SEM analyses using the *lavaan* package in R [137-138], Monte Carlo-based confidence intervals [139] can be derived for nearly any parameter of interest through linkages with the *semTools* package [140], and dichotomous and ordinal data can be transformed to continuous latent variable metrics using diagonally weighted least squares (WLSMV in R) or other relevant estimation procedures (see e.g., [62, 107]). In general, differences in G and D coefficients between observed score and continuous latent variable metrics diminish as numbers of scale points increase with the largest differences observed when items have only two response options [62, 141]. Although the data analyzed here had no missing values, the *auxiliary*, *BootMiss-class*, and *bsMissBoot* routines within the *semTools* package can be linked to *lavaan* to handle missing

data using auxiliary information, multiple imputation, bootstrapping, and related procedures (see [140] for further details).

### 3. This Investigation

Our main purpose within the present study is to demonstrate how indicator-mean and related procedures can be integrated into SEMs to allow for complete analyses of GT-based univariate, multivariate, and bifactor designs on both observed score and continuous latent response variable metrics. To evaluate the congruence of observed score results between SEM and ANOVA-based procedures, we compare G coefficients, D coefficients, and variance components obtained from the univariate and multivariate GT SEM designs to those obtained from the conventional package *mGENOVA* [95], which is often considered the gold standard when analyzing multivariate designs. Further comparisons of results involving the GT SEMs are made for composite and subscale scores across the multivariate and bifactor designs, across observed and continuous latent response variable score metrics, and across numbers of item scale points (2, 4, & 8).

In relation to the previous research studies cited here, we anticipate the following results:

1. G coefficients, D coefficients, and variance components obtained from the GT-based univariate and multivariate SEMs will be highly congruent with those obtained from *mGENOVA*.
2. Multivariate and bifactor GT SEMs will yield comparable G and D coefficients for subscale and composite scores.
3. G and D coefficients for the *pi* designs will exceed those for the *pio* designs due to control of fewer sources of measurement error.
4. Across all multivariate designs, correlation coefficients between scale scores will be higher after correcting for measurement error, but the difference between corrected and uncorrected coefficients will be greater in *pio* than in *pi* designs.
5. General factor effects will exceed group factor effects at both subscale and composite levels within the bifactor designs.
6. Similar patterns of VARs for subscales will be found across multivariate and bifactor designs.
7. Composite and subscale scores will be affected by specific-factor (method), transient (state), and response-random (within-occasion noise) measurement error within the *pio* designs, but those effects be greater overall at the subscale than composite level due to inclusion of fewer item scores.
8. Differences in G and D coefficients for two, four, and eight item scale points will be greater on observed score than on continuous latent response variable metrics.
9. G and D coefficients will be greater on continuous latent response variable than on observed score metrics but to diminishing degrees with increases in numbers of item scale points.

### 4. Methods

#### 4.1. Participants, Measures, and Procedure

We used the same dataset from Lee and Vispoel [107] in which 511 college students from educational psychology and statistics courses within a large Midwestern university (77.50% female, 82.00 Caucasian, mean age = 21.16) completed the full form of the adult level of the Music Self-Perception Inventory (MUSPI) [116, 142-144] on two occasions, a week apart. However, for sake of efficiency, variety, and comparison, we analyzed responses to the same subscales (Instrument Playing, Reading Music, Listening, and Composing) using items from the shortened form of the MUSPI (MUSPI-S), all of which are included in the full form.

Each subscale within the MUSPI-S includes four positively phrased items answered along an 8-point item response metric with the following options: (1) *Definitely False*, (2) *Mostly False*, (3) *Moderately False*, (4) *More False Than True*, (5) *More True Than False*, (6), *Moderately True*, (7) *Mostly True*, (8) *Definitely True*. We computed subscale scores by adding responses to all items within the given subscale, and composite scores by summing all subscale scores to represent Overall Music Proficiency. Psychometric evidence supporting the use of the MUSPI-S scores includes alpha

reliability coefficients for subscale scores no lower than 0.91, confirmatory factor analyses of responses yielding excellent fits to the data, and verification of expected relationships of MUSPI-S subscale scores with each other and with a wide variety of external criterion variables (see, e.g., [116-119]). To evaluate effects of number of item scale points across analyses, we recoded original scores of 1-2, 3-4, 5-6, and 7-8, respectively, to 1, 2, 3, 4 to reduce responses to four scale points, and recoded original scores of 1-4 and 5-8, respectively, to 1 and 2 to reduce responses to two scale points.

#### 4.2. Analyses

Initial analyses included estimation of means, standard deviations, alpha reliability coefficients, and test-retest reliability coefficients for MUSPI-S subscale and composite scores. Subsequent analyses included derivation of variance components, G coefficients, D coefficients, correlation coefficients (corrected and not corrected for measurement error), and VARs for relevant MUSPI-S scales across the *pi* and *pio* univariate, multivariate, and bifactor designs. The *pi* designs include data collected on the first measurement occasion only, and the *pio* designs include data collected on both occasions. Within the multivariate and bifactor *pi* and *pio* designs, items are nested within subscales, and occasions are crossed with subscales in the *pio* designs.

All SEM-based indices were estimated using procedures within the computer package R. For sake of comparison, variance components, G coefficients, and D coefficients for observed scores also were derived for univariate and multivariate designs using *mGENOVA* [95]. To parallel conventional ANOVA-based procedures for observed scores, SEM-based analyses were based on unweighted least squares (ULS) estimation. To convert observed score results to continuous latent response variable metrics within the SEM-based analyses, we used WLSMV estimation within the *lavaan* package [137-138], which is described by its authors as a diagonally weighted least squares procedure with robust standard errors and a mean and variance adjusted test statistic. We also derived 95% Monte Carlo based confidence intervals [139] using the *semTools* package [140] to gauge precision in estimating G, D, and omega coefficients. More detail and computer code for deriving all key indices are provided in our Supplementary Materials.

## 5. Results

### 5.1. Means, Standard Deviations, and Conventional Reliability Estimates for MUSPI-S Scores

Table 5 includes means, standard deviations, and alpha reliability estimates for all MUSPI-S scales and response metrics within each occasion, as well as test-retest reliability estimates across occasions. In relation to individual item scale metrics, means for all scales fall near or below their respective scale midpoint values of 1.5, 2.5, and 4.5 for the two-, four-, and eight-point metrics, with the Composing subscale always having the lowest mean. Given the possible range of scores on each metric (1-2, 1-4, and 1-8), standard deviations for each scale reflect a high degree of variability, respectively, ranging from 0.38-0.46, 0.93-1.13, and 1.96-2.37. These results make sense given the likely heterogeneity of music-related skills within this college student sample. Across scales and response metrics, alpha coefficients within occasions are uniformly high, ranging from 0.91 to 0.98. Test-retest coefficients range from 0.80 to 0.93 and are lower than corresponding alpha coefficients on all instances, thereby reflecting lower occasion-to-occasion than item-to-item consistency for all MUSPI-S scales represented here.

Table 5. Descriptive statistics and conventional reliability coefficients for MUSPI-S scores.

| Metric/ subscale   | Occasion/index          |                       |          |                         |                       |          |             |
|--------------------|-------------------------|-----------------------|----------|-------------------------|-----------------------|----------|-------------|
|                    | Time 1                  |                       |          | Time 2                  |                       |          | Test-Retest |
|                    | Mean<br>Scale<br>(Item) | SD<br>Scale<br>(Item) | $\alpha$ | Mean<br>Scale<br>(Item) | SD<br>Scale<br>(Item) | $\alpha$ |             |
| <b>2-Point</b>     |                         |                       |          |                         |                       |          |             |
| Composite          | 22.90<br>(1.43)         | 6.06 (0.38)           | 0.954    | 23.12<br>(1.45)         | 6.15 (0.38)           | 0.957    | 0.906       |
| Instrument playing | 5.78 (1.44)             | 1.83 (0.46)           | 0.940    | 5.81 (1.45)             | 1.85 (0.46)           | 0.946    | 0.885       |
| Reading music      | 5.77 (1.44)             | 1.85 (0.46)           | 0.948    | 5.83 (1.46)             | 1.84 (0.46)           | 0.942    | 0.908       |
| Listening          | 5.78 (1.44)             | 1.83 (0.46)           | 0.935    | 6.12 (1.53)             | 1.84 (0.46)           | 0.940    | 0.799       |
| Composing          | 5.32 (1.33)             | 1.67 (0.42)           | 0.911    | 5.37 (1.34)             | 1.73 (0.43)           | 0.934    | 0.826       |
| <b>4-Point</b>     |                         |                       |          |                         |                       |          |             |
| Composite          | 35.83<br>(2.24)         | 14.98<br>(0.94)       | 0.967    | 36.48<br>(2.28)         | 14.85<br>(0.93)       | 0.971    | 0.936       |
| Instrument playing | 9.08 (2.27)             | 4.46 (1.11)           | 0.961    | 9.20 (2.30)             | 4.41 (1.10)           | 0.966    | 0.921       |
| Reading music      | 9.03 (2.26)             | 4.52 (1.13)           | 0.964    | 9.14 (2.29)             | 4.38 (1.09)           | 0.971    | 0.930       |
| Listening          | 9.08 (2.27)             | 4.46 (1.11)           | 0.958    | 9.88 (2.47)             | 4.26 (1.07)           | 0.959    | 0.871       |
| Composing          | 8.06 (2.02)             | 3.79 (0.95)           | 0.932    | 8.26 (2.06)             | 3.80 (0.95)           | 0.950    | 0.870       |
| <b>8-Point</b>     |                         |                       |          |                         |                       |          |             |
| Composite          | 63.06<br>(3.94)         | 31.64<br>(1.98)       | 0.971    | 64.54<br>(4.03)         | 31.32<br>(1.96)       | 0.975    | 0.944       |
| Instrument playing | 15.97<br>(3.99)         | 9.42 (2.36)           | 0.967    | 16.29<br>(4.07)         | 9.24 (2.31)           | 0.973    | 0.929       |
| Reading music      | 15.91<br>(3.98)         | 9.49 (2.37)           | 0.968    | 16.12<br>(4.03)         | 9.17 (2.29)           | 0.978    | 0.937       |
| Listening          | 15.97<br>(3.99)         | 9.42 (2.36)           | 0.966    | 17.72<br>(4.43)         | 8.79 (2.20)           | 0.967    | 0.890       |
| Composing          | 14.01<br>(3.50)         | 7.89 (1.97)           | 0.937    | 14.41<br>(3.60)         | 7.91 (1.98)           | 0.958    | 0.891       |

Note. MUSPI-S = shortened form of the Music Self-Perception Inventory.

5.2. Key indices for GT pi Designs

**Univariate and multivariate analyses using SEMs and mGENOVA.** Results for G coefficients, global D coefficients and variance components for observed composite scores within the SEM-ULS multivariate designs and their embedded univariate designs for subscales shown in Table 6 are highly consistent with those obtained from the mGENOVA package. Between the two approaches, G coefficients are identical, global D coefficients differ by no more than 0.001, and variance components differ by no more than 0.002.

Table 6. MUSPI-S G, coefficients, global D coefficients, and variance components for persons × items multivariate designs.

| Scale | Metric/ Procedure |         |         |
|-------|-------------------|---------|---------|
|       | 2-Point           | 4-Point | 8-Point |



|                         | <i>m</i> GENOVA | ULS                      | WLSMV                   | <i>m</i> GENOVA | ULS                     | WLSMV                   | <i>m</i> GENOVA | ULS                     | WLSMV                   |
|-------------------------|-----------------|--------------------------|-------------------------|-----------------|-------------------------|-------------------------|-----------------|-------------------------|-------------------------|
| Composite               |                 |                          |                         |                 |                         |                         |                 |                         |                         |
| G                       | .977            | .977<br>(.964,<br>.989)  | .996<br>(.995,<br>.997) | .985            | .985<br>(.983,<br>.987) | .993<br>(.992,<br>.994) | .988            | .988<br>(.987,<br>.988) | .993<br>(.992,<br>.994) |
| TRE                     | .023            | .023                     | .004                    | .015            | .015                    | .007                    | .012            | .012                    | .007                    |
| G-D                     | .977            | .977<br>(.964,<br>.988)  | .996<br>(.995,<br>.997) | .985            | .985<br>(.983,<br>.987) | .993<br>(.992,<br>.994) | .987            | .987<br>(.987,<br>.988) | .993<br>(.991,<br>.994) |
| $\hat{\sigma}_p^2$      | .140            | .140                     | 1.262                   | .864            | .864                    | 1.093                   | 3.861           | 3.861                   | 1.233                   |
| $\hat{\sigma}_{pi,e}^2$ | .013            | .013                     | .020                    | .052            | .052                    | .031                    | .194            | .194                    | .036                    |
| $\hat{\sigma}_i^2$      | .000            | .000                     | .000                    | .000            | .000                    | .000                    | .002            | .003                    | .001                    |
| Instrument playing      |                 |                          |                         |                 |                         |                         |                 |                         |                         |
| G                       | .940            | .940<br>(.866,<br>1.000) | .989<br>(.984,<br>.993) | .961            | .961<br>(.950,<br>.973) | .980<br>(.976,<br>.984) | .967            | .967<br>(.965,<br>.970) | .982<br>(.979,<br>.985) |
| TRE                     | .060            | .060                     | .011                    | .039            | .039                    | .020                    | .033            | .033                    | .018                    |
| G-D                     | .940            | .940<br>(.863,<br>1.000) | .988<br>(.984,<br>.993) | .961            | .961<br>(.949,<br>.972) | .980<br>(.976,<br>.984) | .967            | .967<br>(.965,<br>.970) | .982<br>(.978,<br>.985) |
| $\hat{\sigma}_p^2$      | .197            | .197                     | 1.995                   | 1.195           | 1.195                   | 1.869                   | 5.367           | 5.367                   | 1.771                   |
| $\hat{\sigma}_{pi,e}^2$ | .050            | .050                     | .092                    | .192            | .192                    | .150                    | .722            | .722                    | .131                    |
| $\hat{\sigma}_i^2$      | .000            | .000                     | .001                    | .000            | .001                    | .001                    | .002            | .004                    | .001                    |
| Reading music           |                 |                          |                         |                 |                         |                         |                 |                         |                         |
| G                       | .948            | .948<br>(.876,<br>1.000) | .991<br>(.987,<br>.994) | .964            | .964<br>(.953,<br>.976) | .982<br>(.979,<br>.986) | .968            | .968<br>(.966,<br>.971) | .984<br>(.982,<br>.987) |
| TRE                     | .052            | .052                     | .009                    | .036            | .036                    | .018                    | .032            | .032                    | .016                    |
| G-D                     | .948            | .947<br>(.873,<br>1.000) | .990<br>(.986,<br>.994) | .964            | .964<br>(.952,<br>.975) | .982<br>(.978,<br>.985) | .967            | .967<br>(.965,<br>.970) | .983<br>(.980,<br>.986) |
| $\hat{\sigma}_p^2$      | .202            | .202                     | 1.556                   | 1.230           | 1.230                   | 1.364                   | 5.445           | 5.445                   | 1.768                   |
| $\hat{\sigma}_{pi,e}^2$ | .045            | .045                     | .059                    | .181            | .181                    | .098                    | .715            | .715                    | .112                    |
| $\hat{\sigma}_i^2$      | .000            | .000                     | .003                    | .003            | .003                    | .004                    | .021            | .022                    | .008                    |
| Listening               |                 |                          |                         |                 |                         |                         |                 |                         |                         |
| G                       | .935            | .935<br>(.860,<br>1.000) | .986<br>(.981,<br>.991) | .958            | .958<br>(.946,<br>.969) | .978<br>(.973,<br>.982) | .966            | .966<br>(.963,<br>.969) | .981<br>(.977,<br>.983) |
| TRE                     | .065            | .065                     | .014                    | .042            | .042                    | .022                    | .034            | .034                    | .019                    |

|                         |      |                 |                 |       |                 |                 |       |                 |                 |
|-------------------------|------|-----------------|-----------------|-------|-----------------|-----------------|-------|-----------------|-----------------|
|                         |      | .935            | .986            |       | .958            | .978            |       | .966            | .980            |
| G-D                     | .935 | (.857,<br>.998) | (.981,<br>.991) | .958  | (.945,<br>.969) | (.973,<br>.982) | .966  | (.963,<br>.969) | (.977,<br>.983) |
| $\hat{\sigma}_p^2$      | .196 | .196            | 1.546           | 1.158 | 1.158           | 1.626           | 4.927 | 4.927           | 1.413           |
| $\hat{\sigma}_{pi,e}^2$ | .055 | .055            | .088            | .205  | .205            | .149            | .692  | .692            | .112            |
| $\hat{\sigma}_i^2$      | .000 | .000            | .000            | .000  | .000            | .000            | .000  | .000            | .000            |
| Composing               |      |                 |                 |       |                 |                 |       |                 |                 |
|                         |      | .911            | .979            |       | .932            | .962            |       | .937            | .954            |
| G                       | .911 | (.818,<br>.993) | (.972,<br>.986) | .932  | (.915,<br>.948) | (.954,<br>.968) | .937  | (.934,<br>.941) | (.946,<br>.960) |
| TRE                     | .089 | .089            | .021            | .068  | .068            | .038            | .063  | .063            | .046            |
|                         |      | .911            | .978            |       | .931            | .961            |       | .936            | .953            |
| G-D                     | .911 | (.815,<br>.990) | (.971,<br>.986) | .931  | (.914,<br>.947) | (.953,<br>.967) | .937  | (.932,<br>.940) | (.944,<br>.959) |
| $\hat{\sigma}_p^2$      | .159 | .159            | 1.026           | .837  | .837            | .595            | 3.649 | 3.649           | 1.123           |
| $\hat{\sigma}_{pi,e}^2$ | .062 | .062            | .088            | .246  | .246            | .095            | .974  | .974            | .218            |
| $\hat{\sigma}_i^2$      | .000 | .000            | .003            | .003  | .003            | .002            | .016  | .018            | .006            |

*Note.* MUSPI-S = shortened form of the Music Self-Perception Inventory, G = generalizability coefficient, G-D = global dependability coefficient, TRE = total relative error, ULS = unweighted least squares estimation, WLSMV = diagonally weighted least squares estimation with robust standard errors and a mean and variance adjusted test statistic. Values within parentheses represent 95% confidence interval limits.

**G and global D coefficients within the SEMs.** Across scales and numbers of item scale points, G and global D coefficients for the SEMs are uniformly higher for WLSMV than for ULS estimation, but these differences are noticeably greater for subscale than for composite scores. This indicates that scale coarseness effects can be especially pronounced when using a small number of items. These differences are further exacerbated, but to a lesser degree here, when using more limited numbers of scale points.

In relation to widths of confidence intervals (i.e., differences between upper and lower limits) shown in Table 6, precision in estimating G and global D coefficients is weakest with two item scale points and improves with increases in numbers of item scale points for ULS but not always for WLSMV estimates. With eight-point items, widths of confidence intervals are uniformly narrower for ULS than for WLSMV estimation, thereby indicating greater relative increases in precision with increases in item scale points on the observed score metric than on the continuous latent response variable metric.

**Cut-score-specific D coefficients.** For purposes of comparison, cut-score-specific D coefficients for composite and subscale scores using both ULS and WLSMV estimation are depicted on Z-score metrics ( $M = 0$ ,  $SD = 1$ ) in Figure 4. Across scales and estimation procedures, dependability is lowest at the scale mean and progressively increases as scores deviate farther and farther away from the mean. Consistent with the results for G and global D coefficients, composite cut-score-specific D coefficients based on ULS estimates always exceed those for subscales at common standard deviation distances away from the scale mean. As expected, differences in G coefficients, global D coefficients (Table 6), and cut-score-specific D coefficients (Figure 4) for observed scores based on ULS estimation are typically greater between two and four scale points than between four and eight scale points. However, this pattern does not hold for the continuous latent variable response metric based on

WLSMV estimation in which corrections for coarseness are greater for two-point than for four- or eight-point scales. However, as scores become increasingly extreme for all scales and estimation procedures, cut-score-specific coefficients across numbers of item scale points begin to coincide, thereby illustrating that numbers of scale point effects on score dependability are noticeably greater within the middle of score distributions than at the extremes.

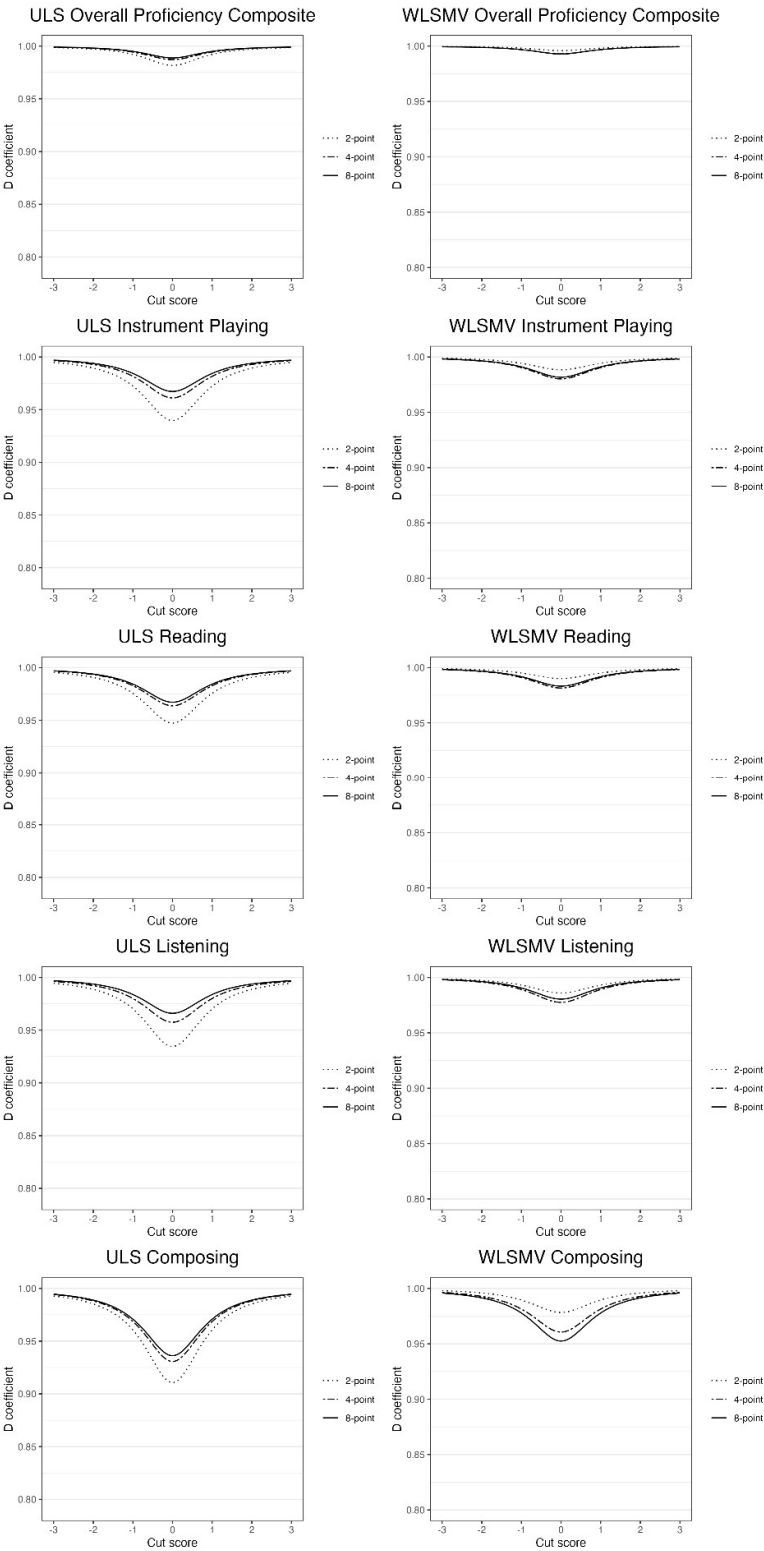


Figure 4. Cut-score-specific dependability coefficients for *pi* multivariate designs. *Note.* *pi* = *persons* × *items* design. Within the present multivariate designs, persons are crossed with subscales, and items are nested within subscales. Results for each individual subscale are the same as those that would be obtained from a separate univariate analysis. Although not pictured here, cut-score-specific dependability coefficients for corresponding *pi* bifactor designs are virtually identical to those shown here.

**Subscale intercorrelations.** As noted earlier, an important advantage of GT multivariate analyses is to produce correlation coefficients between subscale scores corrected for measurement error. As expected, the corrected and uncorrected correlation coefficients between subscales in Table 7 reveal that the relationship between each pair for measured constructs is greater than would otherwise be inferred (ULS: uncorrected  $\bar{r}$  = 0.664, mean corrected  $\bar{r}$  = 0.699; WLSMV: uncorrected  $\bar{r}$  = 0.748, mean corrected  $\bar{r}$  = 0.764). The modest average differences between corrected and uncorrected coefficients observed here result from the generally high G coefficients for the subscales using either estimation method. Such differences are expected to increase when G coefficients are corrected for additional sources of measurement error (see Equation (6) and results of the *pio* designs presented in subsequent sections).

**Table 7.** MUSPI-S subscale correlation coefficients for GT *persons* × *items* multivariate designs.

| Metric/Estimator       | Correlation Coefficient |               |           |           |
|------------------------|-------------------------|---------------|-----------|-----------|
| <b>2-Point / ULS</b>   | Instrument playing      | Reading music | Listening | Composing |
| Instrument playing     | -----                   | 0.844         | 0.609     | 0.626     |
| Reading music          | 0.797                   | -----         | 0.657     | 0.558     |
| Listening              | 0.571                   | 0.618         | -----     | 0.646     |
| Composing              | 0.579                   | 0.518         | 0.596     | -----     |
| <b>2-Point / WLSMV</b> | Instrument playing      | Reading music | Listening | Composing |
| Instrument playing     | -----                   | 0.920         | 0.731     | 0.749     |
| Reading music          | 0.910                   | -----         | 0.778     | 0.685     |
| Listening              | 0.721                   | 0.769         | -----     | 0.784     |
| Composing              | 0.737                   | 0.675         | 0.770     | -----     |
| <b>4-Point / ULS</b>   | Instrument playing      | Reading music | Listening | Composing |
| Instrument playing     | -----                   | 0.878         | 0.678     | 0.681     |
| Reading music          | 0.845                   | -----         | 0.716     | 0.617     |
| Listening              | 0.650                   | 0.689         | -----     | 0.689     |
| Composing              | 0.645                   | 0.585         | 0.651     | -----     |
| <b>4-Point / WLSMV</b> | Instrument playing      | Reading music | Listening | Composing |
| Instrument playing     | -----                   | 0.915         | 0.736     | 0.736     |
| Reading music          | 0.898                   | -----         | 0.776     | 0.679     |
| Listening              | 0.721                   | 0.760         | -----     | 0.745     |
| Composing              | 0.715                   | 0.660         | 0.722     | -----     |
| <b>8-Point / ULS</b>   | Instrument playing      | Reading music | Listening | Composing |
| Instrument playing     | -----                   | 0.884         | 0.696     | 0.704     |
| Reading                | 0.855                   | -----         | 0.727     | 0.652     |
| Listening              | 0.673                   | 0.703         | -----     | 0.719     |
| Composing              | 0.671                   | 0.621         | 0.684     | -----     |

| 8-Point / WLSMV    | Instrument playing | Reading music | Listening | Composing |
|--------------------|--------------------|---------------|-----------|-----------|
| Instrument playing | -----              | 0.885         | 0.725     | 0.732     |
| Reading music      | 0.871              | -----         | 0.744     | 0.685     |
| Listening          | 0.711              | 0.731         | -----     | 0.748     |
| Composing          | 0.708              | 0.664         | 0.724     | -----     |

*Note.* Observed score correlation coefficients are in the lower triangle of the matrices, and disattenuated correlations are in the upper triangle. MUSPI-S = shortened form of the Music Self-Perception Inventory, GT = generalizability theory, ULS = unweighted least squares estimation, WLSMV = diagonally weighted least squares estimation with robust standard errors and a mean and variance adjusted test statistic.

**GT bifactor analyses for subscale and composite scores.** GT bifactor analyses provide an additional perspective on results by generally yielding G and D coefficients comparable to those obtained from parallel multivariate analyses but further subdividing universe score variance into general and group factor effects. Comparing G and global D coefficients for the bifactor designs in Table 8 to those for the multivariate designs in Table 6 verifies that this congruence holds, with coefficients being identical with ULS estimation and varying by no more than 0.001 with WLSMV estimation. Accordingly, differences in results between ULS and WLSMV estimation and across numbers of scale points previously discussed for multivariate designs hold here, as would the differences between cut-score specific D coefficients shown in Figure 4.

**Table 8.** MUSPI-S G coefficients, global D coefficients, and variance components for GT *persons × items* bifactor designs.

| Scale/ Index                | Metric/ Procedure  |                   |                   |                   |                   |                   |
|-----------------------------|--------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
|                             | 2-Point            |                   | 4-Point           |                   | 8-Point           |                   |
|                             | ULS                | WLSMV             | ULS               | WLSMV             | ULS               | WLSMV             |
| Composite                   |                    |                   |                   |                   |                   |                   |
| G                           | .977 (.964, .989)  | .996 (.995, .997) | .985 (.983, .987) | .993 (.991, .994) | .988 (.987, .988) | .992 (.991, .993) |
| $\hat{\omega}_{HComposite}$ | .869 (.837, .899)  | .938 (.925, .950) | .900 (.895, .905) | .930 (.918, .940) | .909 (.908, .910) | .924 (.913, .934) |
| $\hat{\omega}_{HSubscale}$  | .108 (.073, .143)  | .058 (.046, .071) | .085 (.080, .091) | .063 (.053, .073) | .078 (.077, .079) | .068 (.060, .079) |
| TRE                         | .023               | .004              | .015              | .007              | .012              | .008              |
| G-D                         | .977 (.963, .988)  | .996 (.995, .997) | .985 (.983, .987) | .992 (.991, .993) | .987 (.987, .988) | .992 (.991, .993) |
| $\hat{\sigma}_p^2$          | .140               | 4.299             | .863              | 2.573             | 3.859             | 1.982             |
| $\hat{\sigma}_{general}^2$  | .125               | 4.050             | .789              | 2.411             | 3.553             | 1.846             |
| $\hat{\sigma}_{group}^2$    | .015               | .249              | .075              | .162              | .305              | .137              |
| $\hat{\sigma}_{pi,e}^2$     | .013               | .071              | .052              | .078              | .194              | .060              |
| $\hat{\sigma}_i^2$          | .000               | .002              | .000              | .001              | .003              | .002              |
| Instrument playing          |                    |                   |                   |                   |                   |                   |
| G                           | .940 (.865, 1.000) | .989 (.984, .993) | .961 (.950, .973) | .980 (.976, .984) | .967 (.965, .970) | .982 (.979, .985) |
| $\hat{\omega}_{HComposite}$ | .744 (.568, .953)  | .888 (.830, .948) | .806 (.774, .840) | .882 (.843, .920) | .819 (.811, .826) | .860 (.826, .890) |
| $\hat{\omega}_{HSubscale}$  | .196 (.000, .394)  | .101 (.042, .158) | .155 (.118, .191) | .098 (.062, .135) | .149 (.140, .157) | .122 (.094, .153) |
| TRE                         | .060               | .011              | .039              | .020              | .033              | .018              |

|                             |                    |                   |                   |                   |                   |                   |
|-----------------------------|--------------------|-------------------|-------------------|-------------------|-------------------|-------------------|
| G-D                         | .940 (.863, 1.000) | .988 (.984, .993) | .961 (.949, .972) | .980 (.976, .984) | .967 (.965, .970) | .982 (.978, .984) |
| $\hat{\sigma}_p^2$          | .197               | 5.114             | 1.195             | 4.327             | 5.367             | 2.621             |
| $\hat{\sigma}_{general}^2$  | .156               | 4.593             | 1.002             | 3.895             | 4.541             | 2.294             |
| $\hat{\sigma}_{group}^2$    | .041               | .521              | .193              | .432              | .825              | .326              |
| $\hat{\sigma}_{pl,e}^2$     | .050               | .237              | .192              | .348              | .722              | .194              |
| $\hat{\sigma}_i^2$          | .000               | .004              | .001              | .002              | .004              | .002              |
| Reading music               |                    |                   |                   |                   |                   |                   |
| G                           | .948 (.876, 1.000) | .991 (.987, .994) | .964 (.953, .976) | .982 (.979, .986) | .968 (.966, .971) | .984 (.982, .987) |
| $\hat{\omega}_{HComposite}$ | .741 (.568, .941)  | .889 (.830, .947) | .800 (.768, .833) | .881 (.842, .920) | .810 (.803, .818) | .844 (.809, .876) |
| $\hat{\omega}_{HSubscale}$  | .206 (.000, .401)  | .102 (.044, .159) | .165 (.128, .200) | .101 (.064, .139) | .158 (.150, .166) | .140 (.110, .174) |
| TRE                         | .052               | .009              | .036              | .018              | .032              | .016              |
| G-D                         | .947 (.873, 1.000) | .990 (.986, .994) | .964 (.952, .975) | .982 (.978, .985) | .967 (.965, .970) | .983 (.980, .986) |
| $\hat{\sigma}_p^2$          | .202               | 6.891             | 1.230             | 2.628             | 5.445             | 2.529             |
| $\hat{\sigma}_{general}^2$  | .158               | 6.182             | 1.020             | 2.358             | 4.557             | 2.169             |
| $\hat{\sigma}_{group}^2$    | .044               | .710              | .210              | .270              | .888              | .360              |
| $\hat{\sigma}_{pl,e}^2$     | .045               | .263              | .181              | .189              | .715              | .161              |
| $\hat{\sigma}_i^2$          | .000               | .013              | .003              | .008              | .022              | .011              |
| Listening                   |                    |                   |                   |                   |                   |                   |
| G                           | .935 (.861, 1.000) | .986 (.981, .991) | .958 (.945, .969) | .978 (.973, .982) | .966 (.963, .969) | .981 (.977, .983) |
| $\hat{\omega}_{HComposite}$ | .529 (.399, .684)  | .695 (.613, .776) | .602 (.576, .628) | .674 (.614, .730) | .626 (.619, .632) | .663 (.612, .709) |
| $\hat{\omega}_{HSubscale}$  | .405 (.217, .559)  | .291 (.212, .373) | .356 (.325, .386) | .304 (.249, .362) | .341 (.333, .348) | .318 (.272, .367) |
| TRE                         | .065               | .014              | .042              | .022              | .034              | .019              |
| G-D                         | .935 (.859, .999)  | .986 (.981, .991) | .958 (.945, .969) | .978 (.973, .981) | .966 (.963, .969) | .980 (.977, .983) |
| $\hat{\sigma}_p^2$          | .196               | 3.577             | 1.158             | 2.671             | 4.927             | 2.305             |
| $\hat{\sigma}_{general}^2$  | .111               | 2.522             | .727              | 1.842             | 3.190             | 1.558             |
| $\hat{\sigma}_{group}^2$    | .085               | 1.055             | .430              | .830              | 1.737             | .747              |
| $\hat{\sigma}_{pl,e}^2$     | .055               | .204              | .205              | .245              | .692              | .183              |
| $\hat{\sigma}_i^2$          | .000               | .000              | .000              | .000              | .000              | .000              |
| Composing                   |                    |                   |                   |                   |                   |                   |
| G                           | .911 (.819, .992)  | .979 (.971, .986) | .932 (.915, .948) | .962 (.954, .968) | .937 (.934, .941) | .954 (.946, .960) |
| $\hat{\omega}_{HComposite}$ | .468 (.342, .626)  | .651 (.562, .735) | .527 (.501, .556) | .606 (.541, .667) | .569 (.563, .576) | .626 (.574, .675) |
| $\hat{\omega}_{HSubscale}$  | .443 (.237, .606)  | .328 (.245, .415) | .404 (.368, .438) | .356 (.297, .417) | .368 (.360, .377) | .327 (.282, .375) |
| TRE                         | .089               | .021              | .068              | .038              | .063              | .046              |
| G-D                         | .911 (.816, .989)  | .978 (.970, .985) | .931 (.914, .947) | .961 (.953, .967) | .936 (.932, .940) | .953 (.944, .959) |

|                            |      |       |      |       |       |       |
|----------------------------|------|-------|------|-------|-------|-------|
| $\hat{\sigma}_p^2$         | .159 | 5.052 | .837 | 2.869 | 3.649 | 2.187 |
| $\hat{\sigma}_{general}^2$ | .082 | 3.358 | .474 | 1.808 | 2.216 | 1.436 |
| $\hat{\sigma}_{group}^2$   | .077 | 1.694 | .363 | 1.062 | 1.433 | .751  |
| $\hat{\sigma}_{pi,e}^2$    | .062 | .434  | .246 | .459  | .974  | .424  |
| $\hat{\sigma}_i^2$         | .000 | .013  | .003 | .011  | .018  | .011  |

*Note.* MUSPI-S = shortened form of the Music Self-Perception Inventory, GT = generalizability theory, G = generalizability coefficient, G-D = global dependability coefficient,  $\hat{\omega}_{HComposite}$  = Omega hierarchical composite coefficient,  $\hat{\omega}_{HSubscale}$  = Omega hierarchical subscale coefficient, TRE = total relative error, ULS = unweighted least squares estimation, WLSMV = diagonally weighted least squares estimation with robust standard errors and a mean and variance adjusted test statistic. Generalizability coefficients for bifactor model are equivalent to omega total coefficients. Values within parentheses represent 95% confidence interval limits.

Relative general and group factor effects on composite and subscale score variance can be examined using omega hierarchical composite ( $\hat{\omega}_{HComposite}$ ) and omega hierarchical subscale ( $\hat{\omega}_{HSubscale}$ ) coefficients that respectively represent proportions of explained general and group factor effects (see Table 8). Across numbers of scale points, these coefficients reveal that the strongest general and weakest group factor effects are found for the Overall Music Proficiency composite ( $\hat{\omega}_{HComposite}$ : 0.869-0.938;  $\hat{\omega}_{HSubscale}$ : 0.058-0.108), Instrument Playing subscale ( $\hat{\omega}_{HComposite}$ : 0.744-0.888;  $\hat{\omega}_{HSubscale}$ : 0.098-0.196), and Reading Music subscale ( $\hat{\omega}_{HComposite}$ : 0.741-0.889;  $\hat{\omega}_{HSubscale}$ : 0.101-0.206), whereas the weakest general and strongest group effects are found for the Composing ( $\hat{\omega}_{HComposite}$ : 0.468-0.651;  $\hat{\omega}_{HSubscale}$ : 0.327-0.443) and Listening ( $\hat{\omega}_{HComposite}$ : 0.529-0.695;  $\hat{\omega}_{HSubscale}$ : 0.291-0.405) subscales.

Precision in estimating G, global D, and omega coefficients in relation to the widths of confidence intervals within the bifactor designs display patterns of G and global D coefficients parallel to those within the corresponding multivariate designs. Widths are generally narrower for composite than for subscale scores, they progressively narrow with increasing numbers of item scale points with ULS estimates, and they narrow when moving from two to either four or eight item scale points with WLSMV estimation, but not necessarily when moving between four to eight item scale points. This same relative pattern of width differences holds for omega hierarchical coefficients.

**Subscale VARs within GT multivariate and bifactor designs.** To use both subscale and composite scores in practice, subscale scores should provide unique information beyond that represented within the composite that subsumes the subscale scores. The VARs for MUSPI-S subscale scores in Table 9 provide a useful mechanism to verify such expectations within both GT multivariate and bifactor designs. Despite the relatively high correlations between many pairs of subscale scores shown in Table 7, the VARs for all MUSPI-S subscales examined here exceed 1.00, thereby supporting their added value beyond the composite. Consistent with results for omega hierarchical coefficients, VARs are higher for the Composing and Listening subscales than for the Instrument Playing and Reading Music subscales.

**Table 9.** MUSPI-S subscale value-added ratios for GT *persons* × *items* multivariate and bifactor designs.

| Design/ subscale           | Metric/Estimator |       |         |       |         |       |
|----------------------------|------------------|-------|---------|-------|---------|-------|
|                            | 2-Point          |       | 4-Point |       | 8-Point |       |
|                            | ULS              | WLSMV | ULS     | WLSMV | ULS     | WLSMV |
| <b>Multivariate design</b> |                  |       |         |       |         |       |
| Instrument playing         | 1.196            | 1.122 | 1.154   | 1.111 | 1.145   | 1.136 |



|                    |       |       |       |       |       |       |
|--------------------|-------|-------|-------|-------|-------|-------|
| Reading            | 1.216 | 1.134 | 1.168 | 1.120 | 1.155 | 1.154 |
| Listening          | 1.338 | 1.225 | 1.278 | 1.217 | 1.268 | 1.259 |
| Composing          | 1.421 | 1.305 | 1.373 | 1.367 | 1.322 | 1.289 |
| Bifactor design    |       |       |       |       |       |       |
| Instrument playing | 1.185 | 1.105 | 1.146 | 1.093 | 1.137 | 1.123 |
| Reading            | 1.199 | 1.098 | 1.153 | 1.113 | 1.143 | 1.137 |
| Listening          | 1.363 | 1.272 | 1.299 | 1.264 | 1.287 | 1.274 |
| Composing          | 1.438 | 1.263 | 1.386 | 1.292 | 1.335 | 1.263 |

*Note.* MUSPI-S = shortened form of the Music Self-Perception Inventory, ULS = unweighted least squares estimation, WLSMV = diagonally weighted least squares estimation with robust standard errors and a mean and variance adjusted test statistic.

5.2. Key indices for GT *pio* Designs

**Univariate and multivariate analyses using SEMs and *mGENOVA*.** As previously noted, *pio* designs generally provide more appropriate indices of score accuracy for trait-based measures because they allow for estimation of multiple sources of measurement error (specific-factor, transient, and random-response) likely to affect scores. In Table 10, we provide G coefficients; global D coefficients; proportions of specific-factor, transient, and random-response measurement error; and variance components for observed composite scores within the ULS-SEM multivariate designs and for observed subscale scores within the embedded univariate designs. Results in the table for *mGENOVA* and the ULS-SEMs are essentially equivalent, with reported indices being identical in nearly all instances and differing by no than 0.002 in other instances.

**Table 10.** MUSPI-S G coefficients, global D coefficients, and variance components for GT *persons × items × occasions* multivariate designs.

| Scale<br>/<br>Inde<br>x | Metric/ Procedure  |                         |                         |                    |                             |                         |                    |                             |                         |
|-------------------------|--------------------|-------------------------|-------------------------|--------------------|-----------------------------|-------------------------|--------------------|-----------------------------|-------------------------|
|                         | 2-Point            |                         |                         | 4-Point            |                             |                         | 8-Point            |                             |                         |
|                         | <i>mGENOV</i><br>A | ULS                     | WLSM<br>V               | <i>mGENOV</i><br>A | ULS                         | WLSM<br>V               | <i>mGENOV</i><br>A | ULS                         | WLSM<br>V               |
| Composite               |                    |                         |                         |                    |                             |                         |                    |                             |                         |
| G                       | .904               | .904<br>(.853,<br>.956) | .946<br>(.925,<br>.967) | .935               | .935<br>(.926<br>,<br>.943) | .954<br>(.944,<br>.964) | .943               | .943<br>(.941<br>,<br>.945) | .948<br>(.939,<br>.955) |
| SFE                     | .002               | .002                    | .001                    | .001               | .001                        | .001                    | .001               | .001                        | .001                    |
| TE                      | .075               | .075                    | .050                    | .052               | .052                        | .039                    | .046               | .046                        | .046                    |
| RRE                     | .021               | .021                    | .003                    | .012               | .012                        | .006                    | .010               | .010                        | .005                    |
| TRE                     | .097               | .097                    | .054                    | .065               | .065                        | .046                    | .057               | .057                        | .052                    |
| G-D                     | .904               | .904                    | .945                    | .934               | .934                        | .953                    | .942               | .942                        | .947                    |

|                          |      |                              |                         |       |                             |                         |       |                             |                         |
|--------------------------|------|------------------------------|-------------------------|-------|-----------------------------|-------------------------|-------|-----------------------------|-------------------------|
|                          |      | (.851,<br>.954)              | (.923,<br>.966)         |       | (.925<br>,<br>.942)         | (.942,<br>.963)         |       | (.939<br>,<br>.944)         | (.937,<br>.954)         |
| $\hat{\sigma}_p^2$       | .132 | .132                         | 1.766                   | .812  | .812                        | 1.282                   | 3.649 | 3.64<br>9                   | 1.088                   |
| $\hat{\sigma}_{pi}^2$    | .001 | .001                         | .004                    | .004  | .004                        | .003                    | .017  | .017                        | .003                    |
| $\hat{\sigma}_{po}^2$    | .011 | .011                         | .094                    | .045  | .045                        | .053                    | .179  | .179                        | .053                    |
| $\hat{\sigma}_{pio,e}^2$ | .012 | .012                         | .024                    | .043  | .043                        | .032                    | .153  | .153                        | .023                    |
| $\hat{\sigma}_i^2$       | .000 | .000                         | .001                    | .000  | .000                        | .000                    | .001  | .002                        | .001                    |
| $\hat{\sigma}_o^2$       | .000 | .000                         | .002                    | .001  | .001                        | .001                    | .004  | .004                        | .001                    |
| $\hat{\sigma}_{io}^2$    | .000 | .000                         | .000                    | .000  | .000                        | .000                    | .000  | .000                        | .000                    |
| Instrument playing       |      |                              |                         |       |                             |                         |       |                             |                         |
| G                        | .880 | .879<br>(.740,<br>1.000<br>) | .959<br>(.943,<br>.976) | .917  | .917<br>(.892<br>,<br>.943) | .953<br>(.943,<br>.962) | .926  | .926<br>(.920<br>,<br>.931) | .944<br>(.934,<br>.952) |
| SFE                      | .006 | .006                         | .002                    | .003  | .003                        | .002                    | .003  | .003                        | .000                    |
| TE                       | .063 | .063                         | .030                    | .046  | .046                        | .029                    | .045  | .045                        | .039                    |
| RRE                      | .051 | .051                         | .009                    | .033  | .033                        | .017                    | .027  | .027                        | .017                    |
| TRE                      | .120 | .121                         | .041                    | .083  | .083                        | .047                    | .074  | .074                        | .057                    |
| G-D                      | .879 | .879<br>(.737,<br>1.000<br>) | .959<br>(.942,<br>.975) | .917  | .917<br>(.891<br>,<br>.942) | .952<br>(.942,<br>.962) | .925  | .925<br>(.919<br>,<br>.931) | .943<br>(.933,<br>.951) |
| $\hat{\sigma}_p^2$       | .186 | .186                         | 1.878                   | 1.128 | 1.12<br>8                   | 2.057                   | 5.038 | 5.03<br>8                   | 1.490                   |
| $\hat{\sigma}_{pi}^2$    | .005 | .005                         | .017                    | .016  | .016                        | .013                    | .064  | .064                        | .003                    |
| $\hat{\sigma}_{po}^2$    | .013 | .013                         | .059                    | .057  | .057                        | .062                    | .243  | .243                        | .062                    |
| $\hat{\sigma}_{pio,e}^2$ | .043 | .043                         | .067                    | .164  | .164                        | .147                    | .580  | .580                        | .106                    |
| $\hat{\sigma}_i^2$       | .000 | .000                         | .002                    | .001  | .001                        | .001                    | .003  | .004                        | .001                    |
| $\hat{\sigma}_o^2$       | .000 | .000                         | .000                    | .000  | .000                        | .001                    | .003  | .003                        | .001                    |
| $\hat{\sigma}_{io}^2$    | .000 | .000                         | .000                    | .000  | .000                        | .000                    | .000  | .000                        | .000                    |
| Reading music            |      |                              |                         |       |                             |                         |       |                             |                         |
| G                        | .902 | .902<br>(.761,<br>1.000<br>) | .972<br>(.959,<br>.985) | .926  | .926<br>(.901<br>,<br>.952) | .960<br>(.951,<br>.968) | .934  | .934<br>(.928<br>,<br>.939) | .946<br>(.937,<br>.953) |
| SFE                      | .006 | .006                         | .002                    | .003  | .003                        | .002                    | .003  | .003                        | .004                    |
| TE                       | .042 | .042                         | .018                    | .041  | .041                        | .024                    | .039  | .039                        | .041                    |
| RRE                      | .049 | .049                         | .008                    | .029  | .029                        | .014                    | .024  | .024                        | .009                    |

|                          |      |               |              |       |              |              |       |              |              |
|--------------------------|------|---------------|--------------|-------|--------------|--------------|-------|--------------|--------------|
| TRE                      | .098 | .098          | .028         | .074  | .074         | .040         | .066  | .066         | .054         |
| G-D                      | .902 | .902          | .971         | .926  | .925         | .959         | .933  | .933         | .945         |
|                          |      | (.757, 1.000) | (.957, .984) |       | (.899, .951) | (.950, .967) |       | (.927, .938) | (.936, .953) |
| $\hat{\sigma}_p^2$       | .192 | .192          | 2.661        | 1.145 | 1.145        | 2.028        | 5.077 | 5.077        | 1.397        |
| $\hat{\sigma}_{pi}^2$    | .005 | .005          | .025         | .015  | .015         | .015         | .059  | .059         | .024         |
| $\hat{\sigma}_{po}^2$    | .009 | .009          | .049         | .051  | .051         | .051         | .214  | .214         | .060         |
| $\hat{\sigma}_{pio,e}^2$ | .042 | .042          | .089         | .145  | .145         | .121         | .530  | .530         | .054         |
| $\hat{\sigma}_i^2$       | .000 | .000          | .003         | .001  | .002         | .003         | .009  | .011         | .003         |
| $\hat{\sigma}_o^2$       | .000 | .000          | .001         | .000  | .000         | .001         | .000  | .001         | .000         |
| $\hat{\sigma}_{io}^2$    | .000 | .000          | .001         | .000  | .000         | .001         | .003  | .002         | .001         |
| Listening                |      |               |              |       |              |              |       |              |              |
| G                        | .799 | .800          | .909         | .869  | .869         | .912         | .888  | .888         | .893         |
|                          |      | (.664, .949)  | (.879, .940) |       | (.843, .895) | (.895, .928) |       | (.882, .894) | (.876, .907) |
| SFE                      | .000 | .000          | .000         | .002  | .002         | .001         | .001  | .001         | .001         |
| TE                       | .138 | .138          | .078         | .089  | .089         | .064         | .078  | .078         | .090         |
| RRE                      | .063 | .063          | .013         | .040  | .040         | .022         | .032  | .032         | .016         |
| TRE                      | .201 | .201          | .092         | .131  | .131         | .088         | .112  | .112         | .107         |
| G-D                      | .799 | .799          | .907         | .868  | .868         | .911         | .886  | .886         | .891         |
|                          |      | (.661, .944)  | (.876, .938) |       | (.841, .893) | (.893, .927) |       | (.880, .893) | (.873, .905) |
| $\hat{\sigma}_p^2$       | .168 | .168          | 2.240        | 1.019 | 1.019        | 1.316        | 4.407 | 4.407        | 1.299        |
| $\hat{\sigma}_{pi}^2$    | .000 | .000          | .000         | .008  | .008         | .006         | .029  | .029         | .007         |
| $\hat{\sigma}_{po}^2$    | .029 | .029          | .193         | .105  | .105         | .093         | .389  | .389         | .131         |
| $\hat{\sigma}_{pio,e}^2$ | .053 | .053          | .133         | .189  | .189         | .128         | .638  | .638         | .091         |
| $\hat{\sigma}_i^2$       | .000 | .000          | .000         | .000  | .000         | .000         | .000  | .000         | .000         |
| $\hat{\sigma}_o^2$       | .000 | .000          | .004         | .001  | .002         | .002         | .009  | .010         | .003         |
| $\hat{\sigma}_{io}^2$    | .000 | .000          | .000         | .000  | .000         | .000         | .000  | .001         | .000         |
| Composing                |      |               |              |       |              |              |       |              |              |
| G                        | .818 | .818          | .927         | .864  | .864         | .910         | .883  | .883         | .889         |
|                          |      | (.661, .998)  | (.902, .952) |       | (.830, .898) | (.893, .924) |       | (.875, .891) | (.871, .904) |
| SFE                      | .007 | .007          | .003         | .007  | .007         | .003         | .008  | .008         | .003         |
| TE                       | .105 | .105          | .058         | .077  | .077         | .058         | .065  | .065         | .081         |

|                          |      |                         |                         |      |                             |                         |       |                             |                         |
|--------------------------|------|-------------------------|-------------------------|------|-----------------------------|-------------------------|-------|-----------------------------|-------------------------|
| RRE                      | .070 | .070                    | .013                    | .053 | .053                        | .029                    | .044  | .044                        | .026                    |
| TRE                      | .182 | .182                    | .074                    | .136 | .136                        | .090                    | .117  | .117                        | .110                    |
| G-D                      | .817 | .817<br>(.657,<br>.991) | .925<br>(.900,<br>.951) | .862 | .862<br>(.827<br>,<br>.896) | .908<br>(.890,<br>.923) | .881  | .881<br>(.873<br>,<br>.889) | .888<br>(.869,<br>.902) |
| $\hat{\sigma}_p^2$       | .148 | .148                    | 1.753                   | .777 | .777                        | .865                    | 3.445 | 3.44<br>5                   | 1.051                   |
| $\hat{\sigma}_{pi}^2$    | .005 | .005                    | .025                    | .024 | .024                        | .012                    | .122  | .122                        | .015                    |
| $\hat{\sigma}_{po}^2$    | .019 | .019                    | .109                    | .069 | .069                        | .055                    | .253  | .253                        | .096                    |
| $\hat{\sigma}_{pio,e}^2$ | .051 | .051                    | .096                    | .189 | .189                        | .110                    | .693  | .693                        | .123                    |
| $\hat{\sigma}_i^2$       | .000 | .000                    | .006                    | .002 | .002                        | .003                    | .011  | .012                        | .004                    |
| $\hat{\sigma}_o^2$       | .000 | .000                    | .001                    | .001 | .001                        | .001                    | .004  | .005                        | .001                    |
| $\hat{\sigma}_{io}^2$    | .000 | .000                    | .000                    | .000 | .000                        | .000                    | .000  | .001                        | .000                    |

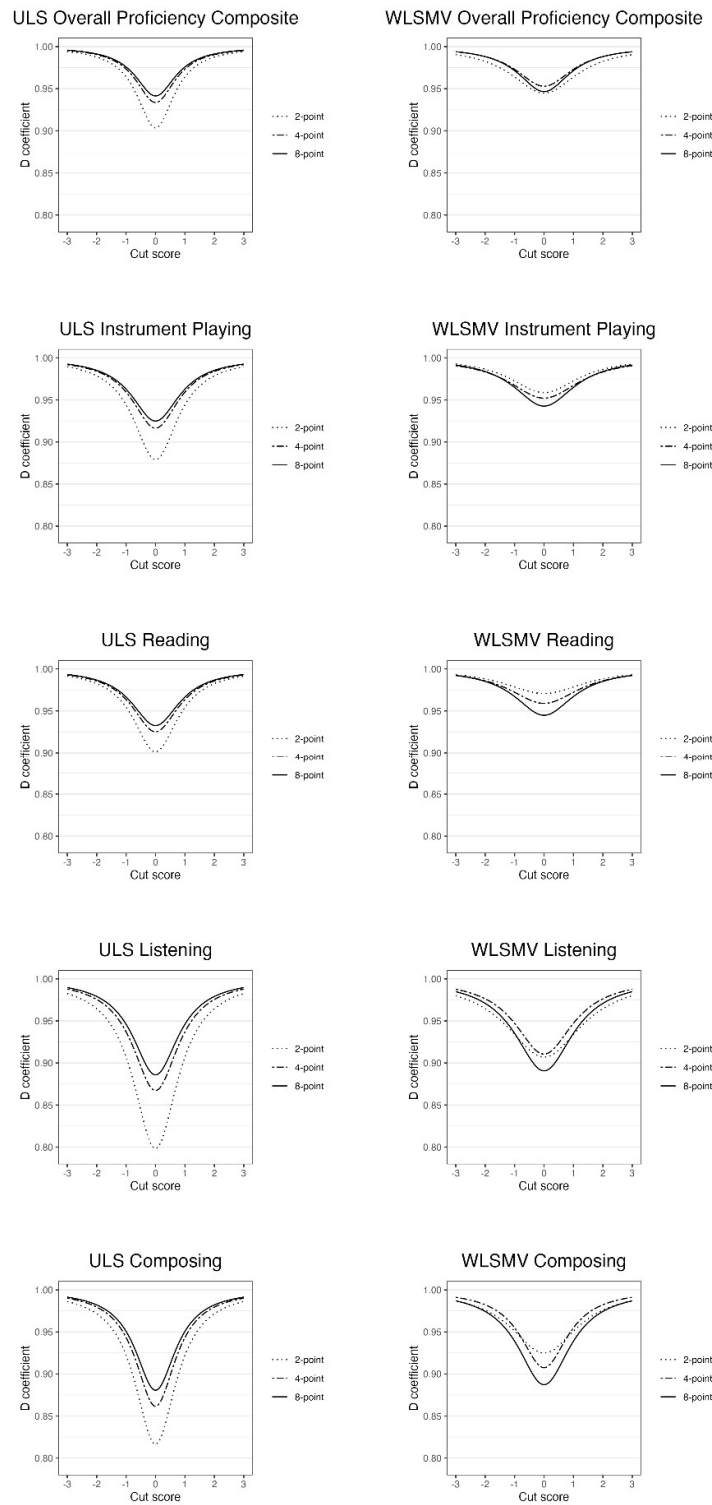
Note. MUSPI-S = shortened form of the Music Self-Perception Inventory, GT = generalizability theory, G = generalizability coefficient, SFE = specific-factor error, TE = transient error; RRE = random-response error, TRE = total relative error, G-D = global dependability coefficient, ULS = unweighted least squares estimation, WLSMV = diagonally weighted least squares estimation with robust standard errors and a mean and variance adjusted test statistic. Values within parentheses represent 95% confidence interval limits.

**G and global D coefficients within the SEMs.** Consistent with results for the *pi* designs, G and global D coefficients within the *pio* designs shown in Table 10 are uniformly greater for WLSMV than for ULS estimation, again implying that score accuracy is reduced due to the dichotomous or ordinal nature of the item response metrics. G and global D coefficients based on ULS estimation are more affected by changes in numbers of item scale points than those based on WLSMV estimation, with dichotomous item scales always producing the lowest observed score coefficients. In most cases, corrections for scale coarseness using WLSMV estimates and increasing numbers of item scale points using ULS estimates reduces all sources of measurement error, and this is especially true with two and four item scale points. In line with the conventional alpha and test-retest coefficients presented earlier, transient error (i.e., occasion) effects are greater than specific-factor error (i.e., item) effects for all scales, and this holds true across both estimation procedures.

As was the case for the *pi* designs, confidence intervals for G and global D coefficients shown in Table 10 within the *pio* designs reveal different patterns in precision for ULS and WLSMV estimates as numbers of item scale points increase. Widths of the confidence intervals for these coefficients progressively narrow with increases in numbers of item scale points for ULS estimates, but this is not always the case for WLSMV estimates. With WLSMV estimates, confidence interval widths are generally wider with two item scale points and similar with four and eight item scale points. That is, precision in estimating G, global D, and omega coefficients consistently improves with increases in numbers of item scale points on the observed score metrics but not necessarily on continuous latent response variable metrics.

**Cut-score-specific D coefficients.** Cut-score-specific D coefficients for all scales and both estimation procedures within the *pio* designs are plotted in Figure 5. Overall, trends observed for the *pio* designs parallel those for the *pi* designs but with coefficients being lower on average because additional sources of measurement error are represented. As before, cut-score-specific D coefficients across scales steadily increase as scores move farther and farther away from the scale mean, are higher for composite than for subscale scores, are higher on continuous latent response variable than on

observed score metrics, and vary less with changes in numbers of item scale points on continuous latent response variable than on observed score metrics.



**Figure 5.** Cut-score-specific dependability coefficients for *pio* multivariate designs. *Note.* *pio* = persons  $\times$  items  $\times$  occasions design. Within the present multivariate designs, persons and occasions are crossed with subscales, and items are nested within subscales. Results for each individual subscale are the same as those that would be obtained from a separate univariate analysis. Although not pictured here, cut-score-specific dependability coefficients for corresponding *pio* bifactor designs are virtually identical to those shown here.

**Subscale intercorrelations.** Table 11 includes correlation coefficients corrected and uncorrected for the three sources of measurement error estimated within the *pio* multivariate designs. As would be expected, corrected correlation coefficients again always exceed corresponding uncorrected coefficients (ULS: uncorrected  $\bar{r}$  = 0.638, corrected  $\bar{r}$  = 0.721; WLSMV: uncorrected  $\bar{r}$  = 0.723, corrected  $\bar{r}$  = 0.776). However, due to inclusion of additional sources of measurement error, corrected coefficients on average exceed those from the *pi* designs (ULS:  $\bar{r}$  = 0.721 versus 0.699; WLSMV:  $\bar{r}$  = 0.776 versus 0.764) as do the average differences between corrected and uncorrected coefficients across those designs (ULS: mean *pio* difference = 0.083, mean *pi* difference = 0.035; WLSMV: mean *pio* difference = 0.053, mean *pi* difference = 0.016).

**Table 11.** Correlation coefficients for GT *persons* × *items* × *occasions* multivariate designs.

| Metric/Estimator       |       | Correlation        |               |           |           |
|------------------------|-------|--------------------|---------------|-----------|-----------|
| <b>2-Point / ULS</b>   |       | Instrument playing | Reading music | Listening | Composing |
| Instrument playing     | ----- |                    | 0.848         | 0.660     | 0.629     |
| Reading music          | 0.756 | -----              |               | 0.689     | 0.571     |
| Listening              | 0.553 | 0.585              | -----         |           | 0.666     |
| Composing              | 0.534 | 0.491              | 0.539         | -----     |           |
| <b>2-Point / WLSMV</b> |       | Instrument playing | Reading music | Listening | Composing |
| Instrument playing     | ----- |                    | 0.915         | 0.756     | 0.735     |
| Reading music          | 0.883 | -----              |               | 0.785     | 0.682     |
| Listening              | 0.706 | 0.738              | -----         |           | 0.779     |
| Composing              | 0.693 | 0.647              | 0.715         | -----     |           |
| <b>4-Point / ULS</b>   |       | Instrument playing | Reading music | Listening | Composing |
| Instrument playing     | ----- |                    | 0.880         | 0.701     | 0.700     |
| Reading music          | 0.811 | -----              |               | 0.741     | 0.650     |
| Listening              | 0.626 | 0.665              | -----         |           | 0.723     |
| Composing              | 0.623 | 0.581              | 0.627         | -----     |           |
| <b>4-Point / WLSMV</b> |       | Instrument playing | Reading music | Listening | Composing |
| Instrument playing     | ----- |                    | 0.911         | 0.749     | 0.744     |
| Reading music          | 0.871 | -----              |               | 0.789     | 0.701     |
| Listening              | 0.698 | 0.738              | -----         |           | 0.767     |
| Composing              | 0.692 | 0.655              | 0.699         | -----     |           |
| <b>8-Point / ULS</b>   |       | Instrument playing | Reading music | Listening | Composing |
| Instrument playing     | ----- |                    | 0.889         | 0.715     | 0.723     |
| Reading music          | 0.826 | -----              |               | 0.754     | 0.680     |
| Listening              | 0.649 | 0.687              | -----         |           | 0.749     |
| Composing              | 0.654 | 0.617              | 0.663         | -----     |           |
| <b>8-Point / WLSMV</b> |       | Instrument playing | Reading music | Listening | Composing |
| Instrument playing     | ----- |                    | 0.899         | 0.747     | 0.750     |
| Reading music          | 0.849 | -----              |               | 0.777     | 0.707     |
| Listening              | 0.686 | 0.714              | -----         |           | 0.775     |
| Composing              | 0.687 | 0.649              | 0.691         | -----     |           |

*Note.* Observed score correlation coefficients are in the lower triangle of the matrices, and disattenuated correlations are in the upper triangle. MUSPI-S = shortened form of the Music Self-Perception Inventory, GT = generalizability theory, ULS = unweighted least squares estimation, WLSMV = diagonally weighted least squares estimation with robust standard errors and a mean and variance adjusted test statistic.

**GT bifactor analyses for subscale and composite scores.** As previously noted, GT *pio* bifactor designs allow for partitioning of measurement error variance into three sources (specific-factor, transient, and random-response) and universe score variance into two sources (general factor and group factor). This partitioning is reflected in G coefficients, global D coefficients, proportions of measurement error, and variance components for the bifactor *pio* designs that appear in Table 12. G coefficients, global D coefficients, and proportions of measurement error for bifactor and corresponding multivariate designs are identical for subscale scores and differ by no more than 0.001 for composite scores when using ULS estimation. Slightly greater differences occur between the bifactor and multivariate designs when using WLSMV estimation with the maximum difference between composite score G or global D coefficients equaling 0.008. Although not depicted here, cut-score-specific D coefficients are virtually identical to those for the corresponding multivariate design shown in Figure 5.

**Table 12.** MUSPI-S G coefficients, global D coefficients, and variance components for GT *persons × items × occasions* bifactor designs.

| Scale/<br>Index                | Metric/ Procedure       |                         |                         |                         |                         |                      |
|--------------------------------|-------------------------|-------------------------|-------------------------|-------------------------|-------------------------|----------------------|
|                                | 2-Point                 |                         | 4-Point                 |                         | 8-Point                 |                      |
|                                | ULS                     | WLSMV                   | ULS                     | WLSMV                   | ULS                     | WLSMV                |
| Composite                      |                         |                         |                         |                         |                         |                      |
| G                              | .904<br>(.854,<br>.955) | .949<br>(.927,<br>.969) | .934<br>(.926,<br>.943) | .962<br>(.952,<br>.971) | .942<br>(.940,<br>.944) | .952 (.943,<br>.959) |
| $\hat{\omega}_{H_{Composite}}$ | .813<br>(.761,<br>.864) | .882<br>(.852,<br>.910) | .861<br>(.852,<br>.870) | .909<br>(.893,<br>.924) | .875<br>(.873,<br>.877) | .890 (.873,<br>.904) |
| $\hat{\omega}_{H_{Subscale}}$  | .091<br>(.070,<br>.119) | .066<br>(.054,<br>.082) | .073<br>(.069,<br>.077) | .052<br>(.045,<br>.062) | .067<br>(.067,<br>.068) | .062 (.054,<br>.072) |
| SFE                            | .002                    | .001                    | .001                    | .001                    | .001                    | .001                 |
| TE                             | .075                    | .047                    | .052                    | .032                    | .047                    | .043                 |
| RRE                            | .021                    | .003                    | .012                    | .006                    | .010                    | .005                 |
| TRE                            | .097                    | .051                    | .065                    | .038                    | .058                    | .048                 |
| G-D                            | .903<br>(.852,<br>.953) | .948<br>(.925,<br>.969) | .933<br>(.924,<br>.942) | .961<br>(.950,<br>.970) | .941<br>(.939,<br>.943) | .951 (.942,<br>.958) |
| $\hat{\sigma}_p^2$             | .132                    | 5.998                   | .812                    | 2.878                   | 3.647                   | 1.477                |
| $\hat{\sigma}_{general}^2$     | .118                    | 5.578                   | .748                    | 2.721                   | .017                    | 1.380                |
| $\hat{\sigma}_{group}^2$       | .013                    | .420                    | .064                    | .157                    | .181                    | .097                 |
| $\hat{\sigma}_{pi}^2$          | .001                    | .015                    | .004                    | .007                    | .153                    | .004                 |
| $\hat{\sigma}_{po}^2$          | .011                    | .300                    | .045                    | .095                    | .002                    | .066                 |
| $\hat{\sigma}_{pio,e}^2$       | .012                    | .086                    | .043                    | .072                    | .004                    | .033                 |
| $\hat{\sigma}_i^2$             | .000                    | .003                    | .000                    | .001                    | .000                    | .001                 |



|                             |               |              |              |              |              |                   |
|-----------------------------|---------------|--------------|--------------|--------------|--------------|-------------------|
| $\hat{\sigma}_o^2$          | .000          | .005         | .001         | .003         | 3.647        | .002              |
| $\hat{\sigma}_{io}^2$       | .000          | .000         | .000         | .000         | .017         | .000              |
| Instrument playing          |               |              |              |              |              |                   |
|                             | .879          | .959         | .917         | .953         | .926         |                   |
| G                           | (.747, 1.033) | (.943, .976) | (.892, .942) | (.943, .962) | (.920, .932) | .944 (.934, .952) |
|                             | .712          | .862         | .767         | .848         | .782         |                   |
| $\hat{\omega}_{HComposite}$ | (.574, .837)  | (.799, .915) | (.744, .790) | (.807, .884) | (.777, .787) | .832 (.797, .862) |
|                             | .167          | .097         | .150         | .104         | .144         |                   |
| $\hat{\omega}_{HSubscale}$  | (.041, .360)  | (.050, .156) | (.123, .179) | (.074, .141) | (.137, .150) | .112 (.087, .141) |
| SFE                         | .006          | .002         | .003         | .002         | .003         | .000              |
| TE                          | .063          | .030         | .046         | .029         | .045         | .039              |
| RRE                         | .051          | .009         | .033         | .017         | .027         | .017              |
| TRE                         | .121          | .041         | .083         | .047         | .074         | .056              |
|                             | .879          | .959         | .917         | .952         | .925         |                   |
| G-D                         | (.744, 1.028) | (.942, .975) | (.890, .941) | (.942, .962) | (.919, .931) | .943 (.933, .951) |
| $\hat{\sigma}_p^2$          | .186          | 4.799        | 1.128        | 4.898        | 1.636        | 1.636             |
| $\hat{\sigma}_{general}^2$  | .151          | 4.315        | .944         | 4.361        | 4.256        | 1.441             |
| $\hat{\sigma}_{group}^2$    | .035          | .484         | .184         | .537         | .783         | .194              |
| $\hat{\sigma}_{pi}^2$       | .005          | .043         | .016         | .031         | .064         | .003              |
| $\hat{\sigma}_{po}^2$       | .013          | .150         | .057         | .147         | .243         | .068              |
| $\hat{\sigma}_{pio,e}^2$    | .043          | .172         | .164         | .351         | .580         | .116              |
| $\hat{\sigma}_i^2$          | .000          | .004         | .001         | .003         | .004         | .001              |
| $\hat{\sigma}_o^2$          | .000          | .001         | .000         | .002         | .003         | .001              |
| $\hat{\sigma}_{io}^2$       | .000          | .001         | .000         | .000         | .000         | .000              |
| Reading music               |               |              |              |              |              |                   |
|                             | .902          | .972         | .926         | .960         | .934         |                   |
| G                           | (.766, 1.000) | (.959, .985) | (.901, .952) | (.951, .968) | (.928, .939) | .946 (.937, .954) |
|                             | .710          | .861         | .778         | .859         | .795         |                   |
| $\hat{\omega}_{HComposite}$ | (.573, .836)  | (.793, .917) | (.755, .802) | (.817, .895) | (.790, .800) | .829 (.794, .858) |
|                             | .192          | .111         | .148         | .101         | .139         |                   |
| $\hat{\omega}_{HSubscale}$  | (.061, .382)  | (.059, .176) | (.120, .177) | (.069, .139) | (.132, .145) | .117 (.092, .147) |

|                             |               |              |              |              |              |                   |
|-----------------------------|---------------|--------------|--------------|--------------|--------------|-------------------|
| SFE                         | .006          | .002         | .003         | .002         | .003         | .004              |
| TE                          | .042          | .018         | .041         | .024         | .039         | .041              |
| RRE                         | .049          | .008         | .029         | .014         | .024         | .009              |
| TRE                         | .098          | .028         | .074         | .040         | .066         | .054              |
|                             | .902          | .971         | .925         | .959         | .933         |                   |
| G-D                         | (.763, 1.000) | (.957, .984) | (.900, .950) | (.950, .967) | (.927, .939) | .945 (.936, .953) |
| $\hat{\sigma}_p^2$          | .192          | 7.984        | 1.145        | 4.202        | 5.077        | 1.647             |
| $\hat{\sigma}_{general}^2$  | .151          | 7.076        | .962         | 3.761        | 4.324        | 1.443             |
| $\hat{\sigma}_{group}^2$    | .041          | .908         | .183         | .441         | .753         | .204              |
| $\hat{\sigma}_{pi}^2$       | .005          | .074         | .015         | .031         | .059         | .029              |
| $\hat{\sigma}_{po}^2$       | .009          | .147         | .051         | .105         | .214         | .071              |
| $\hat{\sigma}_{pio,e}^2$    | .042          | .268         | .145         | .250         | .530         | .064              |
| $\hat{\sigma}_i^2$          | .000          | .008         | .002         | .007         | .011         | .004              |
| $\hat{\sigma}_o^2$          | .000          | .004         | .000         | .001         | .001         | .000              |
| $\hat{\sigma}_{io}^2$       | .000          | .002         | .000         | .001         | .002         | .001              |
| Listening                   |               |              |              |              |              |                   |
|                             | .800          | .909         | .869         | .912         | .888         |                   |
| G                           | (.667, .951)  | (.878, .939) | (.843, .895) | (.895, .928) | (.882, .894) | .893 (.876, .907) |
|                             | .499          | .660         | .579         | .654         | .606         |                   |
| $\hat{\omega}_{HComposite}$ | (.404, .600)  | (.578, .738) | (.560, .597) | (.595, .708) | (.602, .611) | .640 (.589, .685) |
|                             | .300          | .248         | .290         | .258         | .282         |                   |
| $\hat{\omega}_{HSubscale}$  | (.165, .462)  | (.181, .324) | (.264, .318) | (.210, .311) | (.275, .288) | .253 (.215, .295) |
| SFE                         | .000          | .000         | .002         | .001         | .001         | .001              |
| TE                          | .138          | .078         | .089         | .065         | .078         | .090              |
| RRE                         | .063          | .013         | .040         | .022         | .032         | .016              |
| TRE                         | .201          | .092         | .131         | .088         | .112         | .107              |
|                             | .799          | .907         | .868         | .911         | .886         |                   |
| G-D                         | (.663, .946)  | (.875, .938) | (.841, .894) | (.893, .927) | (.880, .892) | .891 (.874, .905) |
| $\hat{\sigma}_p^2$          | .168          | 6.950        | 1.019        | 2.338        | 4.407        | 2.119             |
| $\hat{\sigma}_{general}^2$  | .105          | 5.051        | .679         | 1.677        | 3.010        | 1.519             |
| $\hat{\sigma}_{group}^2$    | .063          | 1.899        | .340         | .661         | 1.397        | .600              |
| $\hat{\sigma}_{pi}^2$       | .000          | -.011        | .008         | .011         | .029         | .012              |
| $\hat{\sigma}_{po}^2$       | .029          | .598         | .105         | .165         | .389         | .214              |
| $\hat{\sigma}_{pio,e}^2$    | .053          | .413         | .189         | .227         | .638         | .149              |

|                             |              |              |              |              |              |                   |
|-----------------------------|--------------|--------------|--------------|--------------|--------------|-------------------|
| $\hat{\sigma}_i^2$          | .000         | .000         | .000         | .000         | .000         | .000              |
| $\hat{\sigma}_o^2$          | .000         | .012         | .002         | .004         | .010         | .005              |
| $\hat{\sigma}_{io}^2$       | .000         | .001         | .000         | .001         | .001         | .000              |
| Composing                   |              |              |              |              |              |                   |
|                             | .818         | .927         | .864         | .910         | .883         |                   |
| G                           | (.666, .995) | (.902, .951) | (.831, .899) | (.893, .924) | (.875, .891) | .889 (.873, .904) |
|                             | .419         | .592         | .520         | .592         | .564         |                   |
| $\hat{\omega}_{HComposite}$ | (.332, .515) | (.502, .677) | (.501, .539) | (.530, .650) | (.559, .568) | .600 (.548, .648) |
|                             | .399         | .335         | .344         | .317         | .319         |                   |
| $\hat{\omega}_{HSubscale}$  | (.246, .575) | (.260, .416) | (.311, .377) | (.266, .373) | (.312, .327) | .290 (.250, .333) |
| SFE                         | .007         | .003         | .007         | .003         | .008         | .003              |
| TE                          | .105         | .057         | .077         | .058         | .065         | .082              |
| RRE                         | .070         | .013         | .053         | .029         | .044         | .026              |
| TRE                         | .182         | .073         | .136         | .090         | .117         | .111              |
|                             | .817         | .925         | .862         | .908         | .881         |                   |
| G-D                         | (.661, .988) | (.899, .950) | (.828, .896) | (.890, .923) | (.873, .889) | .888 (.870, .902) |
| $\hat{\sigma}_p^2$          | .148         | 9.489        | .777         | 2.499        | 3.445        | 1.680             |
| $\hat{\sigma}_{general}^2$  | .076         | 6.063        | .468         | 1.627        | 2.200        | 1.133             |
| $\hat{\sigma}_{group}^2$    | .072         | 3.427        | .309         | 0.872        | 1.245        | .547              |
| $\hat{\sigma}_{pi}^2$       | .005         | .134         | .024         | .036         | .122         | .023              |
| $\hat{\sigma}_{po}^2$       | .019         | .588         | .069         | .160         | .253         | .154              |
| $\hat{\sigma}_{pio,e}^2$    | .051         | .519         | .189         | .319         | .693         | .197              |
| $\hat{\sigma}_i^2$          | .000         | .033         | .002         | .007         | .012         | .006              |
| $\hat{\sigma}_o^2$          | .000         | .007         | .001         | .004         | .005         | .002              |
| $\hat{\sigma}_{io}^2$       | .000         | .002         | .000         | .000         | .001         | .000              |

Note. MUSPI-S = shortened form of the Music Self-Perception Inventory, GT = generalizability theory, G = generalizability coefficient,  $\hat{\omega}_{HComposite}$  = Omega hierarchical composite coefficient,  $\hat{\omega}_{HSubscale}$  = Omega hierarchical subscale coefficient. SFE = specific-factor error, TE = transient error; RRE = random-response error, TRE = total relative error, G-D = global dependability coefficient, ULS = unweighted least squares estimation, WLSMV = diagonally weighted least squares estimation with robust standard errors and a mean and variance adjusted test statistic. Generalizability coefficients for bifactor model are equivalent to omega total coefficients. Values within parentheses represent 95% confidence interval limits.

Due to inclusion of additional sources of measurement error, G, global D coefficients, and most omega coefficients are uniformly lower and overall measurement error uniformly higher in the *pio* bifactor designs than in the *pi* designs. Patterns of general and group factor effects for the *pio* designs mirror those in the *pi* designs, with general factor effects being stronger and group factor effects being

weaker for the Overall Music Proficiency composite ( $\hat{\omega}_{HComposite}$ : 0.813-0.909;  $\hat{\omega}_{HSubscale}$ : 0.052-0.091), Instrument Playing subscale ( $\hat{\omega}_{HComposite}$ : 0.712-0.862;  $\hat{\omega}_{HSubscale}$ : 0.097-0.167), and Reading Music subscale ( $\hat{\omega}_{HComposite}$ : 0.710-0.861;  $\hat{\omega}_{HSubscale}$ : 0.101-0.192) than for the Composing ( $\hat{\omega}_{HComposite}$ : 0.419-0.600;  $\hat{\omega}_{HSubscale}$ : 0.290-0.399) and Listening ( $\hat{\omega}_{HComposite}$ : 0.499-0.660;  $\hat{\omega}_{HSubscale}$ : 0.248-0.300) subscales.

Confidence interval widths for estimating G, global D, and omega coefficients in Table 12 display patterns of precision in line with those described for the *pi* multivariate and bifactor designs and *pio* multivariate designs. Widths of the intervals are narrower for composite than for subscale scores, progressively narrower as numbers of item scale points increase with ULS estimation, and wider for two item scale points but similar for four and eight points with WLSMV estimation.

**Subscale VARs within GT multivariate and bifactor designs.** Consistent with many of the indices already reported, VARs for the *pio* multivariate and bifactor designs shown in Table 13 are very similar to each other and to those reported for the *pi* designs in Table 9. All MUSPI-S subscale scores exceed the threshold of 1.00 to support subscale viability, with the Composing subscale always yielding higher VARs than the other subscales, and the Listening subscale yielding higher VARs than Instrument Playing and Reading Music subscales in most instances.

**Table 13.** MUSPI-S subscale value-added ratios for GT *persons* × *items* × *occasions* multivariate and bifactor designs.

| Design/ subscale           | Metric/Estimator |       |         |       |         |       |
|----------------------------|------------------|-------|---------|-------|---------|-------|
|                            | 2-Point          |       | 4-Point |       | 8-Point |       |
|                            | ULS              | WLSMV | ULS     | WLSMV | ULS     | WLSMV |
| <b>Multivariate design</b> |                  |       |         |       |         |       |
| Instrument playing         | 1.184            | 1.162 | 1.151   | 1.117 | 1.140   | 1.137 |
| Reading music              | 1.232            | 1.178 | 1.164   | 1.123 | 1.149   | 1.148 |
| Listening                  | 1.187            | 1.157 | 1.191   | 1.191 | 1.191   | 1.157 |
| Composing                  | 1.369            | 1.303 | 1.291   | 1.303 | 1.262   | 1.219 |
| <b>Bifactor design</b>     |                  |       |         |       |         |       |
| Instrument playing         | 1.176            | 1.153 | 1.145   | 1.091 | 1.135   | 1.126 |
| Reading music              | 1.218            | 1.159 | 1.154   | 1.103 | 1.140   | 1.132 |
| Listening                  | 1.210            | 1.195 | 1.210   | 1.206 | 1.209   | 1.151 |
| Composing                  | 1.386            | 1.250 | 1.304   | 1.263 | 1.275   | 1.206 |

*Note.* MUSPI-S = shortened form of the Music Self-Perception Inventory, ULS = unweighted least squares estimation, WLSMV = diagonally weighted least squares estimation with robust standard errors and a mean and variance adjusted test statistic.

6. Discussion

6.1. Overview

Although introduced to the research community over 60 years ago [1], GT continues to be used widely in both research and practice to evaluate the psychometric properties for scores yielded by a

broad range of assessment procedures. Recent advances in computer technology and structural equation modeling techniques have further expanded accessibility to programs for doing GT analyses and increased the scope of such analyses. In the study reported here, we sought to integrate, synthesize, and apply newly developed GT-based SEM techniques for conducting complete GT analyses of scores from univariate, multivariate, and bifactor designs with varying numbers of item scale points. Illustrations were focused on objectively scored measures with items and occasions serving as universes of generalization, but the same techniques can be readily applied to subjectively scored assessments by substituting raters for either items or occasions.

6.2. Effectiveness of the indicator-mean method

**Univariate and multivariate designs.** A central part of the present analyses was to extend applications of the indicator-mean method for deriving absolute error variance components and related D coefficients to multivariate and bifactor designs and replicate analyses for univariate designs from Lee and Vispoel [107] using a reduced length form of the MUSPI (MUSPI-S). Across MUSPI-S scales and response metrics for observed scores, SEMs with ULS parameter estimates yielded composite and subscale score G coefficients, global D coefficients, and variance components essentially the same as those produced by the *mGENOVA* package, which still is considered the gold standard for performing GT multivariate analyses. However, *mGENOVA* is limited to analysis of one and two facet designs, and although not considered here, SEMs using the indicator-mean approach can be further extended to multivariate and bifactor designs with more than two measurement facets. When Lee and Vispoel [107] did so with three-facet univariate designs, they found that the indicator-mean approach yielded more accurate absolute error and D associated coefficients than did previous methods used in GT SEMs [98], and we would expect similar results to hold for multivariate and bifactor designs with three or more measurement facets.

**Bifactor designs.** Although Vispoel and colleagues [64, 102-104, 108-109, 111] recently extended GT techniques to bifactor designs using SEMs, there currently are no other formal GT-based computer packages for analyzing such designs for purposes of comparison. Nevertheless, the results found here are congruent with those from previous studies in showing that multivariate and bifactor analyses produce essentially the same G and D coefficients within *pi* and *pio* designs [64, 102, 109, 111]. We further demonstrated here that this congruence holds for multivariate and bifactor designs when integrating the indicator-mean method into appropriate GT-based SEMs. These overall results across studies are not surprising given that second-order hierarchical and bifactor models can be considered reparameterizations of each other [see, e.g., 126, 145-146].

6.3. Univariate GT analyses for shortened and full-length forms of the MUSPI-S

The full length MUSPI [142-144] and its shortened form (MUSPI-S; [116-119]), respectively, include twelve and four items per subscale. Consequently, analyzing scores from the MUSPI and MUSPI-S using the same dataset from Lee and Vispoel [107] facilitates comparisons of effects for numbers of subscale items on conventional and GT-based indices of score accuracy. Lee and Vispoel provided alpha and test-retest reliability estimates for the same four subscales analyzed here (Instrument Playing, Reading Music, Listening, and Composing) using response metrics with two, four, and eight points as well as G and global D coefficients for *pi* and *pio* designs for the Composing subscale. In Table 14, we provide those indices for MUSPI-S observed subscale scores here and for MUSPI observed subscale scores from Lee and Vispoel [107].

**Table 14.** Selected score accuracy coefficients for the MUSPI-S from this study and MUSPI from Lee and Vispoel (2024).

| Design/Scale | This Study MUSPI-S | Lee & Vispoel (2024) MUSPI |
|--------------|--------------------|----------------------------|
|--------------|--------------------|----------------------------|

|                                                                                     | $\alpha$<br>(occ<br>1) | $\alpha$<br>(occ<br>2) | Test-<br>Retest | G     | G-D   | $\alpha$<br>(occ<br>1) | $\alpha$<br>(occ<br>2) | Test-<br>Retest | G     | G-D   |
|-------------------------------------------------------------------------------------|------------------------|------------------------|-----------------|-------|-------|------------------------|------------------------|-----------------|-------|-------|
| Mean across subscales                                                               |                        |                        |                 |       |       |                        |                        |                 |       |       |
| 2-Point                                                                             | 0.933                  | 0.940                  | 0.858           |       |       | 0.957                  | 0.960                  | 0.912           |       |       |
| 4-Point                                                                             | 0.953                  | 0.963                  | 0.898           |       |       | 0.972                  | 0.976                  | 0.932           |       |       |
| 8-Point                                                                             | 0.963                  | 0.970                  | 0.913           |       |       | 0.976                  | 0.980                  | 0.936           |       |       |
| <i>persons</i> $\times$ <i>items</i> (Composing subscale)                           |                        |                        |                 |       |       |                        |                        |                 |       |       |
| 2-Point                                                                             | 0.911                  | 0.934                  | 0.826           | 0.911 | 0.911 | 0.942                  | 0.954                  | 0.894           | 0.943 | 0.940 |
| 4-Point                                                                             | 0.932                  | 0.950                  | 0.870           | 0.932 | 0.931 | 0.959                  | 0.971                  | 0.911           | 0.959 | 0.957 |
| 8-Point                                                                             | 0.937                  | 0.958                  | 0.891           | 0.937 | 0.936 | 0.965                  | 0.975                  | 0.919           | 0.965 | 0.962 |
| <i>persons</i> $\times$ <i>items</i> $\times$ <i>occasions</i> (Composing subscale) |                        |                        |                 |       |       |                        |                        |                 |       |       |
| 2-Point                                                                             |                        |                        |                 | 0.818 | 0.817 |                        |                        |                 | 0.884 | 0.882 |
| 4-Point                                                                             |                        |                        |                 | 0.864 | 0.862 |                        |                        |                 | 0.905 | 0.902 |
| 8-Point                                                                             |                        |                        |                 | 0.883 | 0.881 |                        |                        |                 | 0.913 | 0.911 |

*Note.* MUSPI-S = shortened form of the Music Self-Perception Inventory, MUSPI original full-length form of the Music Self-Perception Inventory, occ = occasion, G = generalizability coefficient, G-D = global dependability coefficient.

Mean alpha and test-retest coefficients shown in Table 14, respectively, range from 0.933-0.970 ( $M=0.954$ ) and 0.858-0.913 ( $M=0.890$ ) for MUSPI-S subscales compared to 0.957-0.980 ( $M=0.970$ ) and 0.912-0.936 ( $M=0.927$ ) for MUSPI subscales. Across the two inventories, alpha coefficients always exceed test-retest coefficients, and the magnitude of both coefficients increases with increases in numbers of item scale points. The greatest difference between the inventories is for test-retest coefficients with two item scale points (MUSPI: 0.912 vs MUSPI-S: 0.858) and the smallest is for alpha coefficients on occasion two with eight item scale points (MUSPI: 0.980 vs MUSPI-S: 0.970). In keeping with the conventional reliability estimates, G and global D coefficients for the Composing subscale across the two inventories in both the *pi* and *pjo* designs increase in magnitude with increases in numbers of item scale points. Differences in these indices range from 0.026 to 0.032 in the *pi* designs and from 0.030 to 0.066 in the *pjo* designs, always favoring the full-length MUSPI. The largest differences in the *pjo* designs are for G (0.066) and global D (0.065) coefficients with two item scale points and the smallest are for G (0.030) and global D (0.030) coefficients with eight item scale points. As a result, the best way to approximate the accuracy of original MUSPI scale scores when using its shortened form is to retain its 8-point response metric. When doing so, MUSPI-S scores come reasonably close to approximating the accuracy of MUSPI scores while using only one-third of its original items.

Another noteworthy finding within the present analyses was that G, global D, and cut-score specific D coefficients for the MUSPI-S Composing (e.g., making up your own music) and Listening (e.g., identifying characteristics of music by ear) subscales were lower than those for Instrument Playing and Reading Music subscales across designs and response metrics. This might have resulted from listening ability and composing skills being less concrete to conceptualize and/or less familiar

to the respondents. Nevertheless, overall patterns of results were highly consistent across subscales for observed scores in terms of accuracy improving with increases in item scale points, but with the greater improvements occurring when moving from two to four points than when moving from four to eight points. The clear message conveyed by these results is to avoid using dichotomous scales when measuring constructs like those considered here. Patterns of relative effects of different sources of measurement error within the *pio* designs also were highly consistent across subscales. In nearly all instances, transient error was highest, followed respectively by random response and specific-factor error. These results highlight the importance of retesting when measuring the current constructs, the likely overestimation of score accuracy when relying exclusively on single-occasion data, and the value of estimating effects for multiple sources of measurement error.

#### 6.4. Multivariate GT analyses of MUSPI-S scores

**Composite score results.** Important benefits of using multivariate GT designs include simultaneous univariate analyses for all embedded subscales as already discussed, derivation of more appropriate indices of score accuracy and measurement error for composite scores, and estimation of correlation coefficients between subscale scores corrected for all sources of measurement error estimated within a design. Over 70 years ago, Standard D 6.3 (*If a test is divided into sets of items of different content, internal consistency should be determined by procedures designed for such tests.*) from the Technical Recommendations for Psychological Tests and Diagnostic Techniques [147] underscored the importance of adjusting reliability estimates for composite scores to reflect subscale representation and interrelations. Accordingly, Cronbach et al. [148] subsequently developed alpha coefficients for composite scores stratified by content categories and later applied these same ideas to multi-facet multivariate GT designs to account for additional sources of measurement error [93, 95].

To illustrate the effects of content stratification on score accuracy indices using the present data, we can compare G coefficients for composites from the *pi* multivariate designs (0.977, 0.985, 0.988; see Table 7), which are equivalent to stratified alpha coefficients, to the non-stratified alpha coefficients representing composite scores on the first measurement occasion reported in Table 5 (0.954, 0.967, 0.971). Despite non-stratified alpha coefficients and observed subscale score intercorrelations being relatively high (see Tables 5 & 7), non-stratified alpha coefficients for composite scores are from 0.017 to 0.022 lower than corresponding stratified coefficients. As subscale score intercorrelations decrease, these differences would likely further increase [2, 148]. An important advantage of multivariate GT designs is that all derived G and D coefficients for composite scores are appropriately adjusted for subscale representation and their interrelationships. Unfortunately, in research studies and practical settings, stratified alpha coefficients and score accuracy indices from GT multivariate designs for composites are still rarely reported.

**Composite versus subscale scores.** Due to inclusion of more items, G and D coefficients for MUSPI-S observed composite scores were uniformly higher and corresponding proportions of measurement error uniformly lower than for observed subscale scores. However, the patterns of relative effects for overall designs (*pi* vs *pio*) and numbers of item scale items on observed composite score indices mirrored those for subscale scores, with higher G and D coefficients in the *pi* than in the *pio* designs and increasingly higher values for these coefficients as numbers of item scale points increased. As with subscales, greater increases in score accuracy and reductions in measurement error occurred for observed composite scores when moving from two to four item scale points than when moving from four to eight scale points, again illustrating the disadvantages of using dichotomously scored items within these self-report measures.

**Correlation coefficients.** The final unique feature of multivariate designs illustrated here, and one not shared with corresponding bifactor designs, is to produce correlation coefficients between pairs of subscale scores corrected for all sources of measurement error estimated within the analyzed design. Due to measurement error being present in all designs, corrected correlation coefficients always exceeded uncorrected ones, thereby implying that the underlying constructs measured by



each pair of subscale scores are more strongly related than the uncorrected coefficients would suggest. Because overall measurement error was always greater in the *pio* than in the *pi* designs, corrected correlation coefficients in the *pio* designs always exceeded those in the *pi* designs. These results illustrate the importance of including all relevant sources of measurement error affecting scores within a GT design, not only in assessing score accuracy, but also in gauging the concurrent validity of universe scores. Other consistent findings for corrected correlation coefficients across all designs were in showing that, among the constructs measured, self-perceptions of abilities to play a musical instrument and read music were most strongly related (*pi* design: corrected  $\bar{r} = 0.888$ ; *pio* design: corrected  $\bar{r} = 0.889$ ) and self-perceptions of abilities to compose music and read music were most weakly related (*pi* design: corrected  $\bar{r} = 0.646$ ; *pio* design: corrected  $\bar{r} = 0.665$ ).

#### 6.5. Bifactor GT analyses of MUSPI-S scores

Although first described by Holzinger and colleagues in the 1930s [149-150], only in recent years have uses of bifactor models truly proliferated [see, e.g., 101, 126-130]. A bifactor model is suitable for measures that represent hierarchically structured constructs in which a broad general domain factor affects responses to all items, and additional independent group factors affect responses only to those items intended to measure narrower subdomain constructs. In the present context, the broad factor represented self-perceptions of overall music proficiency and group factors represented self-perceptions of skill in playing musical instruments, reading music, listening, and composing. In contrast to GT univariate and multivariate analyses, universe scores in GT bifactor designs represent the additive sum of independent general and group factor effects. However, unlike the typical conventional single-occasion bifactor analyses that currently dominate the research literature (see, e.g., [128]), GT-based bifactor designs can produce global and cut-score-specific D coefficients at both composite and subscale levels and distinguish multiple sources of measurement error when items are administered over two or more occasions.

G coefficients, D coefficients, and proportions of measurement error for the present GT *pi* and *pio* bifactor designs for observed scores mirrored results from their corresponding multivariate designs, with score accuracy improving more between two and four item scale points than between four and eight scale points, and transient error being the predominant source of measurement error within the *pio* designs. The most important additional finding from the bifactor analyses, and one consistent with most bifactor analyses reported in the research literature (see, e.g., [101-103, 111, 128]), was that general factor exceeded group factor effects at both composite and subscale levels but to a greater extent at the composite level. Among subscales, Instrument Playing and Reading Music were more affected by general factor and less affected by group factor effects than were Composing and Listening. This result along with the correlations among subscale scores discussed in the previous section further verify that perceptions of overall music proficiency were more related to perceptions of performing and reading music than to listening to or composing music.

#### 6.6. Other noteworthy aspects of the GT SEM designs

**Assessing subscale added value.** We chose to report value-added indices as the primary basis for evaluating subscale viability here, because they are applicable to both GT multivariate and bifactor designs. In general, viability of scores from a given subscale is undermined by its overlap with scores from other subscales that comprise the composite. Yet despite the relatively high correlations between observed subscale scores in many instances, VARs for all subscales and designs exceeded 1.00, thereby supporting subscale viability within all contexts considered here. This is likely due, in part, to the typically high G coefficients for subscale scores across designs. Based on these results, reporting of both MUSPI-S subscale and composite scores would be justified for individuals like those sampled here.

**Evaluating effects of scale coarseness.** Applications of common statistical procedures (ANOVA, multiple linear regression, correlational indices, etc.) typically are governed by assumptions that scores are measured on equal interval scales. However, this is unlikely to be strictly true when using

Likert-style questionnaires. To evaluate effects of possible violations of this assumption more thoroughly, we analyzed MUSPI-S data using item scales with two, four, and eight points and compared the results to those obtained from SEMs using WLSMV parameter estimation in which observed item responses are transformed to continuous latent response variable metrics presumed to be equal interval in nature. Such transformations are based on estimation of item thresholds that typically alter distances between observed scale points to conform to those within continuous standard normal score distributions [62]. Results reported here revealed that G and D coefficients on observed score metrics deviated most from those on continuous latent response variable metrics when dichotomously scored items were used but increased in congruence as numbers of item scale points increased. These results again serve to discourage use of the dichotomous scales and illustrate a mechanism that can be used to evaluate effects of scale coarseness using any number of item scale points.

A finding observed here and elsewhere (see, e.g., [62, 107]) was that G and D coefficients on WLSMV metrics were somewhat higher for two than for four or eight observed scale points. These differences can occur for a variety of reasons including a positive bias sometimes observed when estimating score accuracy using dichotomously scored items (see, e.g., [141]), differences in the characteristics of the observed score distributions, and the after-the-fact conversion of the original eight-point item scale metric to two and four points. However, given the added information provided by individual descriptors of responses on the eight-point scale, results obtained on that metric using WLSMV estimation might be expected to best correspond to those that would be obtained from a scale that is truly continuous in nature [151].

**Confidence intervals.** The *semTools* package [140] can be linked to the *lavaan* package in R to derive Monte Carlo-based confidence intervals [139] for nearly any desired parameter. For sake of illustration, we derived 95% confidence intervals for G, global D, and omega hierarchical coefficients. In most instances, widths of the intervals for these indices narrowed with increases in numbers of observed item scale points but were generally much wider for two-point than for four- and eight-point item metrics. As with many other indices already discussed, these results again emphasize drawbacks in precision when using dichotomously scored items.

## 7. Final Conclusions

In writing this article, we sought to provide readers with a guide to analyzing a wide variety of GT-based designs by taking advantage of procedures made possible when using SEMs to conduct such analyses. The supplementary materials associated with this article include code in R and further detail in how to analyze all illustrated designs. We hope that these resources prove useful in evaluating and understanding the quality and nature of scores used for either norm- or criterion-referencing purposes.

**Supplementary Materials:** The following supporting information can be downloaded at the website of this paper posted on Preprints.org. Supplementary Materials File S1: Online Supplement to Structural Equation Modeling Techniques for Estimating Score Dependability within Generalizability Theory Based Univariate, Multivariate, and Bifactor Designs.

**Author Contributions:** Conceptualization, 1.1.1 and 2.2.; methodology, 1.1.1 and 2.2.; software, 2.2 and 3.3.; validation, 1.1.1., 2.2. and 3.3.; formal analysis, 2.2, 1.1.1., and 3.3.; investigation, 1.1.1., 2.2. and 3.3.; resources, 1.1.1.; data curation, 1.1.1. and 2.2.; writing—original draft preparation, 1.1.1., 2.2, and 3.3; writing—review and editing, 1.1.1., 2.2, and 3.3; visualization, 1.1.1., 2.2, and 3.3; supervision, 1.1.1.; project administration, 1.1.1; funding acquisition, 1.1.1. All authors have read and agreed to the published version of the manuscript.

**Funding:** This project received no external funding but did receive internal research assistant support from the Iowa Testing Programs.

**Data Availability Statement:** This study was not preregistered and inquiries about accessibility to the data should be forwarded to the lead author.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Cronbach, L.J.; Rajaratnam, N.; Gleser, G.C. Theory of generalizability: A liberalization of reliability theory. *Br. J. Stat. Psychol.* **1963**, *16*, 137–163.
2. Rajaratnam, N.; Cronbach, L.J.; Gleser, G.C. Generalizability of stratified-parallel tests. *Psychometrika* **1965**, *30*, 39–56.
3. Gleser, G.C.; Cronbach, L.J.; Rajaratnam, N. Generalizability of scores influenced by multiple sources of variance. *Psychometrika* **1965**, *30*, 395–418.
4. Cronbach, L.J.; Gleser, G.C.; Nanda, H.; Rajaratnam, N. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*; Wiley: New York, NY, USA, 1972.
5. Brennan, R.L. *Elements of Generalizability Theory (Revised Edition)*; American College Testing: Iowa City, IA, USA, 1992.
6. Fyans, L.J. *Generalizability Theory: Inferences and Practical Applications*; Jossey-Bass: San Francisco, 1983.
7. Shavelson, R.J.; Webb, N.M. *Generalizability Theory: A Primer*; Sage: Thousand Oaks, CA, USA, 1991.
8. Brennan, R.L. *Generalizability Theory*; Springer: New York, NY, USA, 2001.
9. Cardinet, J.; Johnson, S.; Pini, G. *Applying generalizability theory using EduG*. Routledge: New York, 2010.
10. Crocker, L.; Algina, J. *Introduction to Classical and Modern Test Theory*. Harcourt Brace: New York, 1986.
11. McDonald, R.P. *Test theory: A unified approach*. Erlbaum: Mahwah, NJ, 1999.
12. Raykov, T.; Marcoulides, G.A. *Introduction to psychometric theory*. Routledge: New York, NY, 2011.
13. Marcoulides, G.A. Generalizability theory. In *Handbook of Applied Multivariate Statistics and Mathematical Modeling*; Tinsley, H., Brown, S., Eds.; Academic Press: San Diego, CA, 2000; pp. 527–551.
14. Wiley, E.W.; Webb, N.M.; Shavelson, R.J. The generalizability of test scores. In *APA Handbook of Testing and Assessment in Psychology: Vol. 1. Test Theory and Testing and Assessment in Industrial and Organizational Psychology*; Geisinger, K.F., Bracken, B.A., Carlson, J.F., Hansen, J.C., Kuncel, N.R., Reise, S. P., Rodriguez, M.C., Eds.; American Psychological Association: Washington, DC, 2013; pp. 43–60.
15. Webb, N.M.; Shavelson, R.J.; Steedle, J.T. Generalizability theory in assessment contexts; In *Handbook on Measurement, Assessment, and Evaluation in Higher Education*; Secolsky, C., Denison, D.B., Eds.; Routledge: New York, NY, 2012; pp. 152–169.
16. Gao, X.; Harris, D.J. Generalizability theory. In *APA Handbook of Research Methods in Psychology, Vol. 1. Foundations, Planning, Measures, and Psychometrics*; Cooper, H., Camic, P.M., Long, D.L., Panter, A.T., Rindskopf, D., Sher, K.J., Eds.; American Psychological Association, 2012; pp. 661–681.
17. Allal, L. Generalizability theory. In *The International Encyclopedia of Educational Evaluation*; Walberg, H.J., Haertel, G.D., Eds.; Pergamon: Oxford, England, 1990; pp. 274–279.
18. Shavelson, R.J.; Webb, N.M. Generalizability theory. In *Encyclopedia of Educational Research*; Alkin, M.C., Ed.; Macmillan: New York, NY, 1992; Volume 2, pp. 538–543.
19. Brennan, R.L. Generalizability theory. In *The SAGE Encyclopedia of Social Science Research Methods*; Lewis-Beck, M.S., Bryman, A.E., Liao, T.F., Eds.; SAGE: Thousand Oaks, CA, 2004; Volume 2, pp. 418–420.
20. Shavelson, R.J.; Webb, N.M. Generalizability theory. In *Encyclopedia of Statistics in Behavioral Science*; Everitt, B.S., Howell, D.C., Eds.; Wiley, 2005; pp. 717–719.
21. Brennan, R.L. Generalizability theory. In *International Encyclopedia of Education (3rd ed.)*; Peterson, P., Baker, E.; McGaw, B., Eds.; Elsevier, 2010; Volume 4, pp. 61–68.
22. Matt, G.E.; Sklar, M. Generalizability theory. In *International Encyclopedia of the Social & Behavioral Sciences*; Wright, J.D., Ed.; Elsevier, 2015; Volume 9, pp. 834–838.
23. Franzen, M. Generalizability theory. In *Encyclopedia of Clinical Neuropsychology*; Kreutzer, J.S., DeLuca, J., Caplan, B., eds.; Springer International Publishing: Cham, 2018; pp.1554–1555.
24. Brennan, R.L.; Kane, M.T. (1979). Generalizability theory: A review. In *New Directions for Testing and Measurement: Methodological Developments (No.4)*; Traub, R.E., Ed.; Jossey-Bass: San Francisco, CA, 1979; pp. 33–51.
25. Brennan, R.L. Applications of generalizability theory. In *Criterion-Referenced Measurement: The State of the Art*; Berk, R.A., Ed.; The Johns Hopkins University Press: Baltimore, 1980; pp. 186–232.
26. Jarjoura, D.; Brennan, R.L. Multivariate generalizability models for tests developed according to a table of specifications. In *New Directions for Testing and Measurement: Generalizability Theory: Inferences and Practical Applications (No. 18)*; Fyans, L.J., Ed.; Jossey-Bass: San Francisco, CA, 1983; pp. 83–101.
27. Webb, N.M.; Shavelson, R.J.; Maddahian, E. Multivariate generalizability theory. In *New Directions in Testing and Measurement: Generalizability Theory (No. 18)*; Fyans, L.J., Ed.; Jossey-Bass: San Francisco, CA, 1983; pp. 67–82.
28. Brennan, R.L. Estimating the dependability of the scores. In *A Guide to Criterion-Referenced Test Construction*; Berk, R.A., Ed.; The Johns Hopkins University Press: Baltimore, 1984; pp. 292–334.
29. Allal, L. Generalizability theory. In *Educational Research, Methodology, and Measurement*; Keeves, J.P., Ed.; Pergamon: New York, 1988; pp. 272–277.
30. Feldt, L.S.; Brennan, R.L. Reliability. In *Educational Measurement (3rd ed.)*; Linn, R.L., Ed.; American Council on Education and Macmillan: New York, 1989; pp. 105–146.

31. Brennan, R.L. Generalizability of performance assessments. In *Technical Issues in Performance Assessments*; Phillips, G.W., Ed.; National Center for Education Statistics: Washington, DC, 1996; pp. 19-58.
32. Marcoulides, G.A. Applied generalizability theory models. In *Modern Methods for Business Research*; Marcoulides, G.A., Ed.; Erlbaum: Mahwah, NJ, 1998; pp. 1-21.
33. Strube, M.J. Reliability and generalizability theory. In *Reading and Understanding More Multivariate Statistics*; Grimm, L.G., Yarnold, P.R., Eds.; American Psychological Association: Washington, DC, 2000; pp. 23-66.
34. Haertel, E.H. Reliability. In *Educational Measurement (4th ed.)*; Brennan, R.L., Ed.; American Council on Education/Praeger: Westport, CT, 2006; pp. 65-110.
35. Kreiter, C.D. Generalizability theory. In *Assessment in Health Professions Education*; Downing, S.M., Yudkowsky, R., Eds.; Routledge: New York, 2009; pp. 75-92.
36. Streiner, D.L.; Norman, G.R.; Cairney, J. Generalizability theory. In *Health Measurement Scales: A Practical Guide to Their Development and Use*; Oxford University Press, 2014; pp. 200-226.
37. Shavelson, R.J.; Webb, N. Generalizability theory and its contribution to the discussion of the generalizability of research findings. In *Generalizing from Educational Research: Beyond Qualitative and Quantitative Polarization*; Ercikan, K., Roth, W., Eds.; Routledge: New York, 2019; pp. 13-32.
38. Kreiter, C.D.; Zaidi, N.L.; Park, Y. S. Generalizability theory. In *Assessment in Health Professions Education*; Yudkowsky, R., Park, Y. S., Downing, S.M., Eds.; Routledge: New York, NY, 2020; pp. 51-69.
39. Brennan, R.L. Generalizability theory. In *The History of Educational Measurement: Key Advancements in Theory, Policy, and Practice*; Clauser, B.E., Bunch, M.B., Eds.; Routledge: New York, NY, 2022; pp.206-231.
40. Cardinet, J.; Tourneur, Y.; Allal, L. The symmetry of generalizability theory: Applications to educational measurement. *J. Educ. Meas.* **1976**, *13*, 119-135.
41. Shavelson, R.J.; Dempsey Atwood, N. (1976). Generalizability of measures of teaching behavior. *Rev. Educ. Res.* **1976**, *46*, 553-611.
42. Cardinet, J.; Tourneur, Y.; Allal, L. Extension of generalizability theory and its applications in educational measurement. *J. Educ. Meas.* **1981**, *18*, 183-204.
43. Shavelson, R.J.; Webb, N.M. Generalizability theory: 1973-1980. *Brit. J. Math. Stat. Psy.* **1981**, *34*, 133-166.
44. Webb, N.M.; Shavelson, R.J. (1981). Multivariate generalizability of General Educational Development ratings. *J. Educ. Meas.* **1981**, *18*, 13-22.
45. Nußbaum, A. Multivariate generalizability theory in educational measurement: An empirical study. *Appl. Psych. Meas.* **1984**, *8*(2), 219-230.
46. Shavelson, R.J.; Webb, N.M.; Rowley, G.L. Generalizability theory. *Am. Psychol.* **1989**, *44*, 922-932.
47. Brennan, R.L. Generalizability theory. *Educ. Meas.-Issues Pra.* **1992**, *11*(4), 27-34.
48. Demorest, M.E.; Bernstein, L.E. Applications of generalizability theory to measurement of individual differences in speech perception. *J. Acad. Reh.* **1993**, *26*, 39-50.
49. Brennan, R.L.; Johnson, E.G. Generalizability of performance assessments. *Educ. Meas.-Issues Pra.* **1995**, *14* (4), 9-12.
50. Cronbach, L.J.; Linn, R.L.; Brennan, R.L.; Haertel, E. Generalizability analysis for performance assessments of student achievement for school effectiveness. *Educ. Psychol. Meas.* **1997**, *57*, 373-399.
51. Lynch, B.K.; McNamara, T.F. Using G-theory and many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Lang. Test.* **1998**, *15*(2),158-180.
52. Hoyt, W.T.; Melby, J.N. Dependability of measurement in counseling psychology: An introduction to generalizability theory. *The Counseling Psychologist* **1999**, *27*, 325-352.
53. Brennan, R.L. (2000). (Mis)conceptions about generalizability theory. *Educ. Meas.-Issues Pra.* **2000**, *19*(1), 5-10.
54. Brennan, R.L. (2000). Performance assessments from the perspective of generalizability theory. *Appl. Psych. Meas.* **2000**, *24*, 339-353.
55. Brennan, R.L. Generalizability theory and classical test theory. *Appl. Meas. Educ.* **2010**, *24*, 1-21.
56. Cronbach, L.J.; Shavelson, R.J. My current thoughts on coefficient alpha and successor procedures. *Educ. Psychol. Meas.* **2004**, *64*, 391-418.
57. Tavakol, M.; Brennan, R.L. Medical education assessment: A brief overview of concepts in generalizability theory. *Int. J. Med. Educ.* **2013**, *4*, 221-222.
58. Trejo-Meja, J.A.; Snchez-Mendiola, M.; Mndez-Ramrez, I.; Martnez-Gonzlez, A. Reliability analysis of the objective structured clinical examination using generalizability theory. *Med. Educ. Online* **2016**, *21*(1), 31650.
59. Vispoel, W.P.; Morris, C.A.; Kilinc, M. Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychol. Methods* **2018**, *23*(1), 1-26.
60. Vispoel, W.P.; Morris, C.A.; Kilinc, M. Practical applications of generalizability theory for designing, evaluating, and improving psychological assessments. *J. Pers. Assess.* **2018**, *100*(1), 53-67.
61. Vispoel, W.P.; Morris, C.A.; Kilinc, M. Using generalizability theory to disattenuate correlation coefficients for multiple sources of measurement error. *Multivar. Behav. Res.* **2018**, *53*, 481-501.
62. Vispoel, W.P.; Morris, C.A.; Kilinc, M. Using generalizability theory with continuous latent response variables. *Psychol. Methods* **2019**, *24*(2), 153-178.



63. Vispoel, W.P.; Xu, G.; Schneider, W.S. Using parallel splits with self-report and other measures to enhance precision in generalizability theory analyses. *J. Personal. Assess.* **2022**, *104*(3), 303–319.
64. Vispoel, W.P.; Lee, H.; Hong, H.; Chen, T. Applying multivariate generalizability theory to psychological assessments. *Psychol. Methods* **2023**, 1–23, *advance online publication*.
65. Andersen, S.A.W.; Nayahangan, L.J.; Park, Y.S.; Konge, L. Use of generalizability theory for exploring reliability of and sources of variance in assessment of technical skills: A systematic review and meta-analysis. *Acad. Med.* **2021**, *96*, 1609–1619.
66. Andersen, S.A.W.; Park, Y.S.; Sørensen, M.S.; Konge, L. Reliable assessment of surgical technical skills is dependent on context: An exploration of different variables using Generalizability Theory. *Acad. Med.* **2020**, *95*, 1929–1936.
67. Kreiter, C.; Zaidi, N.B.; Generalizability theory's role in validity research: Innovative applications in health science education. *Health Prof. Educ.* **2020**, *6*(2), 282–290.
68. Suneja, M.; Hanrahan, K.D.; Kreiter, C.; Rowat, J. Psychometric properties of entrustable professional activity-based objective structured clinical examinations during transition from undergraduate to graduate medical education: A generalizability study. *Acad. Med.* **2024**.
69. Anderson, T.N.; Lau, J.N.; Shi, R.; Sapp, R.W.; Aalami, L.R.; Lee, E.W.; Tekian, A.; Park, Y.S. The utility of peers and trained raters in technical skill-based assessments a generalizability theory study. *J. Surg. Educ.* **2022**, *79*, 206–215.
70. Jogerst, K.M.; Eurboonyanun, C.; Park, Y.S.; Cassidy, D.; McKinley, S.K.; Hamdi, I.; Phitayakorn, R.; Petrusa, E.; Gee, D.W. Implementation of the ACS/ APDS Resident Skills Curriculum reveals a need for rater training: An analysis using generalizability theory. *Am. J. Surg.* **2021**, *222*, 541–548.
71. Winkler-Schwartz, A.; Marwa, I.; Bajunaid, K.; Mullah, M.; Alotaibi, F.E.; Bugdadi, A.; Sawaya, R.; Sabbagh, A.J.; Del Maestro, R. A comparison of visual rating scales and simulated virtual reality metrics in neurosurgical training: A generalizability theory study. *World Neurosurg.* **2019**, *127*, e230–e235.
72. Kuru, C.A.; Sezer, R.; Çetin, C.; Haberal, B.; Yakut, Y.; Kuru, İ. *Acad. Radiol.* **2023**, *30*(10), 2290–2298.
73. Gatti, A.A.; Stratford, P.W.; Brisson, N.M.; Maly, M.R. How to optimize measurement protocols: An example of assessing measurement reliability using generalizability theory. *Physiother Can.* **2020**, *72*(2), 112–121.
74. O'Brien, J.; Thompson, M.S.; Hagler, D. Using generalizability theory to inform optimal design for a nursing performance assessment. *Eval. Health Prof.* **2019**, *42*, 297–327.
75. Peeters, M.J. Moving beyond Cronbach's alpha and inter-rater reliability: A primer on generalizability theory for pharmacy education. *Innov. Pharm.* **2021**, *12*, 14.
76. Atilgan, H. Reliability of essay ratings: A study on Generalizability Theory. *Eurasian J. Educ. Res.* **2019**, *19*, 1–18.
77. Chen, D.; Hebert, M.; Wison, J. Examining human and automated ratings of elementary students' writing quality: A multivariate generalizability theory application. *Am. Educ. Res. J.* **2022**, *59*(6), 1122–1156.
78. Deniz, K.Z.; Ilican, E. Comparison of G and Phi coefficients estimated in generalizability theory with real cases. *Int. J. Assess. Tools Educ.* **2021**, *8*(3), 583–595.
79. Wilson, J.; Chen, D.; Sandbank, M.P.; Hebert, M.; Graham, S. Generalizability of automated scores of writing quality in Grades 3–5. *J. Educ. Psychol.* **2019**, *111*(4), 619–640.
80. Eskin, D. Generalizability of Writing Scores and Language Program Placement Decisions: Score Dependability, Task Variability, and Score Profiles on an ESL Placement Test. *Studies in Applied Linguistics & TESOL* **2022**, *21*(2), 21–42.
81. Liao, Ray J. T. The use of generalizability theory in investigating the score dependability of classroom-based L2 reading assessment. *Lang. Test.* **2023**, *40*(1), 86–106.
82. Shin, J. Investigating and optimizing score dependability of a local ITA speaking test across language groups: A generalizability theory approach. *Lang. Test.* **2022**, *39*(2), 313–337.
83. Ford, A.L.B.; Johnson, L.D. The use of generalizability theory to inform sampling of educator language used with preschoolers with autism spectrum disorder. *J. Speech Lang. Hear. R.* **2021**, *64*, 1748–1757.
84. Hollo, A.; Staubitz, J.L.; Chow, J.C. Applying generalizability theory to optimize analysis of spontaneous teacher talk in elementary classrooms. *J. Speech Lang. Hear. R.* **2020**, *63*(6), 1947–1957.
85. Van Hooijdonk, M.; Mainhard, T.; Kroesbergen, E.H.; Van Tartwijk, J. Examining the assessment of creativity with generalizability theory: An analysis of creative problem solving assessment tasks. *Think. Ski. Creat.* **2022**, *43*, 100994.
86. Li, G.; Xie, J.; An, L.; Hou, G.; Jian, H.; Wang, W. A generalizability analysis of the mobile phone addiction tendency scale for Chinese college students. *Front. Psychiatry* **2019**, *10*, 241.
87. Kumar, S.S.; Merkin, A.G.; Numbers, K.; Sachdev, P.S.; Brodaty, H.; Kochan, N.A.; Trollor, J.N.; Mahon, S.; Medvedev, O. A novel approach to investigate depression symptoms in the aging population using generalizability theory. *Psychol. Assess* **2022**, *34*, 684–696. [[CrossRef](#)]
88. Truong, Q.C.; Krageloh, C.U.; Siegert, R.J.; Landon, J.; Medvedev, O.N. Applying generalizability theory to differentiate between trait and state in the Five Facet Mindfulness Questionnaire (FFMQ). *Mindfulness* **2020**, *11*, 953–963. [[CrossRef](#)]

89. Anthony, C.J.; Styck, K.M.; Volpe, R.J.; Robert, C.R.; Coddington, R.S. Using many-facet Rasch measurement and Generalizability Theory to explore rater effects for Direct Behavior Rating–Multi-Item Scales. *Sch. Psychol.* **2023**, *38*, 119–128.
90. Lyndon, M.P.; Medvedev, O.N.; Chen, Y.; Henning, M.A. Investigating stable and dynamic aspects of student motivation using generalizability theory. *Aust. J. Psychol.* **2020**, *72*(2), 199–210.
91. Sanz-Fernández, C.; Morales-Sánchez, V.; Castellano, J.; Mendo, A.H. Generalizability theory in the evaluation of psychological profile in track and field. *Sports (Basel)* **2024**, *12*(5), 127.
92. Mushquash, C.; O'Connor, B.P. SPSS and SAS programs for generalizability theory analyses. *Behav. Res. Methods* **2006**, *38*, 542–547.
93. Crick, J.E.; Brennan, R.L. *Manual for GENOVA: A Generalized Analysis of Variance System* (American College Testing Technical Bulletin No. 43). ACT, Inc.: Iowa City, IA, 1983.
94. Brennan, R.L. *Manual for urGENOVA* (Version 2.1). University of Iowa, Iowa Testing Programs: Iowa City, IA, 2001.
95. Brennan, R.L. *Manual for mGENOVA* (Version 2.1). University of Iowa, Iowa Testing Programs: Iowa City, IA, 2001.
96. Moore, C.T. *gtheory: Apply generalizability theory with R* (R package version 0.1.2). 2016.
97. Huebner, A.; Lucht, M. Generalizability theory in R. *Pract. Assess. Res. Eval.* **2019**, *24*(5).
98. Jorgensen, T.D. How to estimate absolute-error components in structural equation models of generalizability theory. *Psych* **2021**, *3*, 113–133.
99. Vispoel, W.P.; Xu, G.; Kilinc, M. Expanding G-theory models to incorporate congeneric relationships: Illustrations using the Big Five Inventory. *J. Pers. Assess.* **2021**, *104*, 429–442.
100. Vispoel, W.P.; Xu, G.; Schneider, W.S. Interrelationships between latent state-trait theory and generalizability theory in a structural equation modeling framework. *Psychol. Methods* **2022**, *27*, 773–803.
101. Vispoel, W.P.; Lee, H.; Xu, G.; Hong, H. Expanding bifactor models of psychological traits to account for multiple sources of measurement error. *Psychol. Assess.* **2022**, *32*, 1093–1111.
102. Vispoel, W.P.; Lee, H.; Xu, G.; Hong, H. Integrating bifactor models into a generalizability theory structural equation modeling framework. *J. Exp. Educ.* **2023**, *91*, 718–738.
103. Vispoel, W.P.; Lee, H.; Chen, T.; Hong, H. Extending applications of generalizability theory-based bifactor model designs. *Psych* **2023**, *5*, 545–575.
104. Vispoel, W.P.; Lee, H. **Merging generalizability theory and bifactor modeling to improve psychological assessments. Psychology and Psychotherapy: Review Study** **2023**, *7*(1), 1–4.
105. Vispoel, W.P.; Hong, H.; Lee, H.; Jorgensen, T.R. Analyzing complete generalizability theory designs using structural equation models. *Appl. Meas. Educ.* **2023**, *36*, 372–393.
106. Vispoel, W.P.; Lee, H.; Chen, T.; Hong, H. Using structural equation modeling to reproduce and extend ANOVA-based generalizability theory analyses for psychological assessments. *Psych* **2023**, *5*(2), 249–272.
107. Lee, H.; Vispoel, W.P. A robust indicator mean-based method for estimating generalizability theory absolute error and related dependability indices within structural equation modeling frameworks. *Psych* **2024**, *6*(1), 401–425.
108. Vispoel, W.P.; Hong, H.; Lee, H. Benefits of doing generalizability theory analyses within structural equation modeling frameworks: Illustrations using the Rosenberg Self-Esteem Scale [Teacher's corner]. *Struct. Equ. Model.* **2024**, *31*, 165–181.
109. Vispoel, W.P.; Lee, H.; Hong, H. Analyzing multivariate generalizability theory designs within structural equation modeling frameworks [Teacher's corner]. *Struct. Equ. Model.* **2024**, *31*(3), 552–570.
110. Vispoel, W.P.; Lee, H.; Chen, T. Multivariate structural equation modeling techniques for estimating reliability, measurement error, and subscale viability when using both composite and subscale scores in practice. *Mathematics* **2024**, *12*(8), 1–25.
111. Vispoel, W.P.; Lee, H.; Chen, T.; Hong, H. Analyzing and comparing univariate, multivariate, and bifactor generalizability theory designs for hierarchically structured personality traits. *J. Personal. Assess.* **2024**, *106*(3), 285–300.
112. Marcoulides, G.A. Estimating variance components in generalizability theory: The covariance structure analysis approach. *Struct. Equ. Model.* **1996**, *3*, 290–299.
113. Raykov, T.; Marcoulides, G.A. Estimation of generalizability coefficients via a structural equation modeling approach to scale reliability evaluation. *Int. J. Test.* **2006**, *6*, 81–95.
114. Brennan, R.L.; Kane, M.T. An index of dependability for mastery tests. *J. Educ. Meas.* **1977**, *14*, 277–289.
115. Kane, M.T.; Brennan, R.L. Agreement coefficients as indices of dependability for domain-referenced tests. *Appl. Psychol. Meas.* **1980**, *4*, 105–126.
116. Morin, A.J.S.; Scalas, L.F.; Vispoel, W.; Marsh, H.W.; Wen, Z. The Music Self-Perception Inventory: Development of a short form. *Psychol. Music* **2016**, *44*(5), 915–934.
117. Scalas, L.F.; Marsh, H.W.; Vispoel, W.; Morin, A.J.S.; Wen, Z. Music self-concept and self-esteem formation in adolescence: A comparison between individual and normative models of importance within a latent framework. *Psychol. Music* **2017**, *45*(6), 763–780.

118. Fiedler, D.; Hasselhorn, J.; Katrin Arens, A.; Frenzel, A.C.; Vispoel, W.P. Validating scores from the short form of the Music Self-Perception Inventory (MUSPI-S) with seventh- to ninth-grade school students in German. *Psychol. Music* **2024**, 1-22, *advance online publication*.
119. Vispoel, W.P.; Lee, H. Music self-concept: Structure, correlates, and differences across grade-level, gender, and musical activity groups. *Psychology and Psychological Research International Journal* **2024**, 9(1), 00413.
120. Schmidt, F.L.; Hunter, J.E. Measurement error in psychological research: Lessons from 26 research scenarios. *Psychol. Methods* **1996**, 1, 199–223.
121. Schmidt, F.L.; Le, H.; Ilies, R. Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychol. Methods* **2003**, 8, 206–224.
122. Le, H.; Schmidt, F.L.; Putka, D.J. The multifaceted nature of measurement artifacts and its implications for estimating construct-level relationships. *Organ. Res. Methods* **2009**, 12, 165–200.
123. Thorndike, R.L. Reliability. In *Educational Measurement*; Lindquist, E.F., Ed.; American Council on Education: Washington, DC, USA, 1951; pp. 560–620.
124. Spearman, C. The proof and measurement of association between two things. *Am. J. Psychol.* **1904**, 15, 72–101.
125. Spearman, C. Correlation calculated from faulty data. *Brit. J. Psychol.* **1910**, 3, 271–295.
126. Reise, S.P. The rediscovery of bifactor measurement models. *Multivar. Behav. Res.* **2012**, 47(5), 667–696.
127. Reise, S.P.; Bonifay, W.E.; Haviland, M.G. Scoring psychological measures in the presence of multidimensionality. *J. Personal. Assess.* **2013**, 95(2), 129–140.
128. Rodriguez, A.; Reise, S.P.; Haviland, M.G. Applying bifactor statistical indices in the evaluation of psychological measures. *J. Personal. Assess.* **2016**, 98(3), 223–237.
129. Rodriguez, A.; Reise, S.P.; Haviland, M.G. Evaluating bifactor models: Calculating and interpreting statistical indices. *Psychol. Methods* **2016**, 21(2), 137–150.
130. Zinbarg, R.E.; Revelle, W.; Yovel, I.; Li, W. Cronbach's  $\alpha$ , Revelle's  $\beta$ , and McDonald's  $\omega_H$ : Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika* **2005**, 70(1), 123–133.
131. Haberman, S.J. When can subscores have value? *J. Educ. Behav. Stat.* **2008**, 33, 204–229.
132. Haberman, S.J.; Sinharay, S. Reporting of subscores using multidimensional item response theory. *Psychometrika* **2010**, 75, 209–227.
133. Feinberg, R.A.; Jurich, D.P. Guidelines for interpreting and reporting subscores. *Educ. Meas.-Issues Pract.* **2017**, 36, 5–13. [[CrossRef](#)]
134. Sinharay, S. Added value of subscores and hypothesis testing. *J. Educ. Behav. Stat.* **2019**, 44, 25–44. [[CrossRef](#)]
135. Hjarne, M.S.; Lyrén, P.E. Group differences in the value of subscores: A fairness issue. *Front. Educ.* **2020**, 5, 55. [[CrossRef](#)]
136. Feinberg, R.A.; Wainer, H. A simple equation to predict a subscore's value. *Educ. Meas.-Issues Pra.* **2014**, 33, 55–56.
137. Rosseel, Y. lavaan: An R package for structural equation modeling. *J. Stat. Softw.* **2012**, 48, 1–36. [[CrossRef](#)]
138. Rosseel, Y.; Jorgensen, T.D.; De Wilde, L. Package 'lavaan'. R Package Version (0.6-17). 2023. Available online: <https://cran.r-project.org/web/packages/lavaan/lavaan.pdf> (accessed on 8 December 2024).
139. Preacher, K.J.; Selig, J.P. Advantages of Monte Carlo confidence intervals for indirect effects. *Commun. Methods Meas.* **2012**, 6, 77–98.
140. Jorgensen, T.D.; Pornprasertmanit, S.; Schoemann, A.M.; Rosseel, Y. semTools: Useful Tools for Structural Equation Modeling. R Package Version 0.5-6. 2022. Available online: <https://CRAN.R-project.org/package=semTools> (accessed on 8 December 2024).
141. Ark, T.K. Ordinal Generalizability Theory Using an Underlying Latent Variable Framework. Ph.D. Thesis, University of British Columbia, Vancouver, BC, Canada, 2015. Available online: <https://open.library.ubc.ca/soa/cIRcle/collections/ubctheses/24/items/1.0166304> (accessed on 8 December 2024).
142. Vispoel, W.P. Integrating self-perceptions of music skill into contemporary models of self-concept. *The Quarterly Journal of Music Teaching and Learning* **1994**, 5(4), 42–57.
143. Vispoel, W.P. Integrating self-perceptions of music skill into contemporary models of self-concept. *Visions of Research in Music Education* **2021**, 16, Article 33.
144. Vispoel, W.P. Measuring and understanding self-perceptions of musical ability. In *International Advances in Self Research*; Marsh, H.W., Craven, R.G., McInerney, D.M., Eds.; Information Age Publishing: Charlotte, NC, 2003; pp. 151–180.
145. Schmid, J.; Leiman, J.M. The development of hierarchical factor solutions. *Psychometrika* **1957**, 22, 53–61.
146. Schmid, J. The comparability of the bi-factor and second-order factor patterns. *J. Exp. Educ* **1957**, 25, 249–253.
147. American Psychological Association (APA) Technical recommendations for psychological tests and diagnostic techniques. *Psychol. Bull.* **1954**, 51(2, Pt.2), 1–38.



148. Cronbach, L.J.; Schönemann, P.; McKie, D. Alpha coefficients for stratified-parallel tests. *Educ. Psychol. Meas.* **1965**, *25*, 291–312.
149. Holzinger, K.J.; Harman, H.H. Comparison of two factorial analyses. *Psychometrika* **1938**, *3*(1), 45–60.
150. Holzinger, K.J.; Swineford, F. The bi-factor method. *Psychometrika* **1937**, *2*(1), 41–54.
151. Rhemtulla, M.; Brosseau-Liard, P.É.; Savalei, V. When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychol. Methods* **2012**, *17*, 354–373.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.