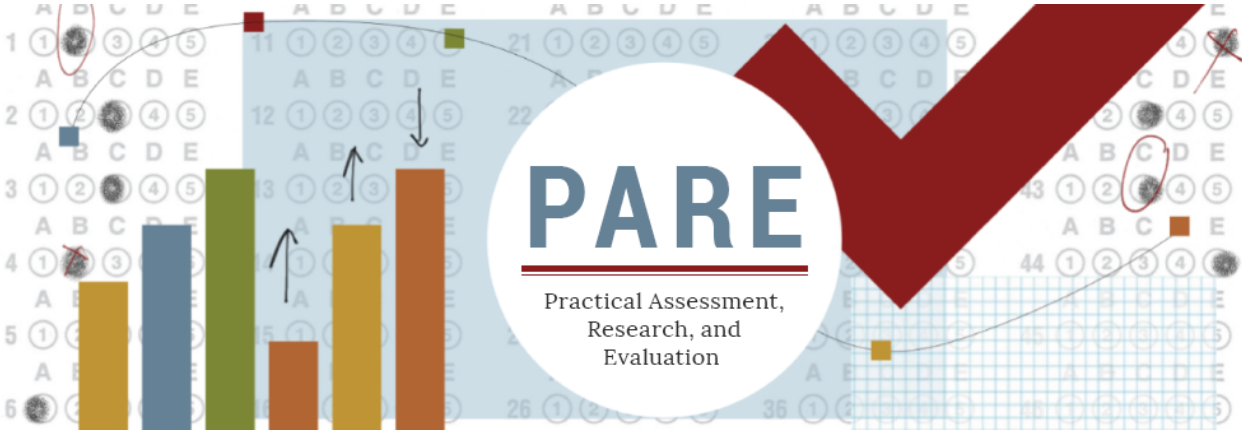


A peer reviewed, open-access electronic journal: ISSN 1531-7714

An R Package for Optimizing the Composite Reliability in Multivariate Nested Designs

Joyce M. W. Moonen - van Loon, *Maastricht University*
Jeroen Donkers, *Maastricht University*

Abstract: The reliability of assessment tools is critical for accurately monitoring student performance in various educational contexts. When multiple assessments are combined to form an overall evaluation, each assessment serves as a data point contributing to the student's performance within a broader educational framework. Determining composite reliability in such cases can be complex, particularly in naturalistic, unbalanced datasets in nested designs, which are common in programmatic and workplace-based assessments, where students are evaluated on unique, practical occasions. This paper introduces the `compositeReliabilityInNestedDesigns` package in R, designed to estimate composite reliability using multivariate generalizability theory and enhance the analysis of assessment data. The package produces extensive Generalizability and Decision study results with graphical interpretations. Composite reliability incorporates weights and covariance to integrate results across assessment tools. Weights can be optimized to minimize standard error of measurement or maximize reliability. Overall, the package's flexible use and optimization empowers assessment tailoring and robust insights into student performance. The approach is suitable for programmatic assessment. The package facilitates reliable, comprehensive evaluation across diverse assessments.

Keywords: Multivariate generalizability theory, Composite reliability, Nested design, Standard Error of Measurement (SEM), Programmatic assessment, R

Introduction

Generalizability theory, originally introduced by Cronbach, Rajaratnam, and Gleser (1963), is a statistical framework utilized to estimate the reliability of measurements by analyzing the sources of measurement errors caused by different conditions (Brennan, 2001; Cronbach et al., 1972; Rajaratnam, Cronbach, &

Gleser, 1965; Shavelson & Webb, 1991; Swanson, 1987). Often, the object of measurement is a person (e.g. a student). A set of similar conditions of measurement is called a facet. Examples of facets are items, assessors, test formats, and so on. By employing generalizability theory, we conduct Generalizability studies (G-studies) to gain a deeper understanding of the facets contributing to the measurements and the amount of error caused by each facet and their interactions. Additionally, Decision studies (D-studies) are designed to estimate the reliability of the same data collected under different conditions (Monteiro, Sullivan, & Chan, 2019). Generalizability theory has been applied to various fields, including educational assessment (Brennan, 2010; Crossley et al., 2002; Hays, Fabb, & van der Vleuten, 1995), sports (Heitman, Kovaleski, & Pugh, 2009; Sanz-Fernández et al., 2024), and social behavior (Vispoel, Morris, & Kilinc, 2018).

The designs of studies in Generalizability theory must specify whether facets are crossed or nested. 'Crossed' refers to the designs in which all the sampled levels of any facet are present for all levels of any other facet(s). In contrast, 'nested' refers to the fact that only some of the sampled levels of a given facet are present for each level of the other facet(s) (Li et al., 2015). In practice and in research literature, univariate designs are more commonly encountered.

Multivariate Generalizability Theory extends traditional Generalizability Theory and provides a systematic and rigorous method for evaluating the dependability and reliability of measurements involving multiple variables, e.g. multiple observations per person, to assess the consistency and stability of measurements across various conditions, considering the interrelationships among the measured variables. This nuanced approach is particularly pertinent in fields such as psychology, education, and other social sciences, where phenomena are often complex and multifaceted.

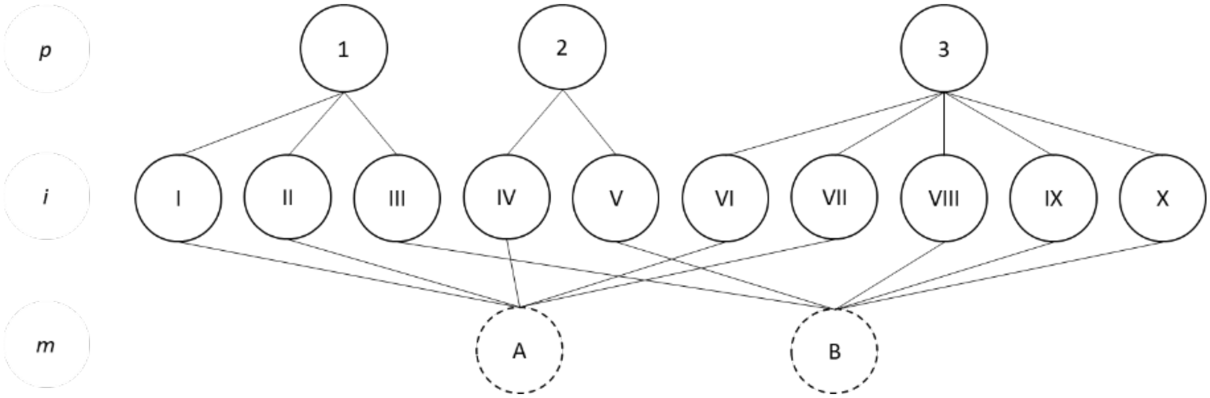
Generalizability Theory, both univariate and multivariate, is described and applied in many different settings. Huebner and Lucht (2019) describe the efficient application of generalizability theory in the statistical software environment R (R Core Team, 2019) using package `gtheory` (Moore, 2016), including data formatting, computing key quantities, and tabulating and visualizing results. This package can be applied to various univariate and multivariate models. Jiang et al. (2020) showed how selected multivariate Generalizability Theory designs can be analyzed using the `glmmTMB` package (Bolker, 2024), and discussed advantages and disadvantages of this approach compared to `mGENOVA` (Brennan, 2001). Vispoel, Lee, and Hong (2024) further demonstrated that multivariate GT designs also can be analyzed using structural equation modeling packages such as `lavaan` (Rosseel, 2012) in R. Brennan et al. (2022) extend multivariate generalizability theory (MGT) to tests with different random-effects designs for each level of a fixed facet, applied to mixed-format tests composed of multiple-choice and free-response items. Vispoel et al. (2023) describe Multivariate Generalizability Theory when scale scores are combined to form composites and compare MGT indices of score consistency and measurement error to those obtained using alternative GT-based procedures and across different software packages for analyzing multivariate GT designs. Vispoel, Lee and Chen (2024) extended multivariate G-theory to congeneric factor models with code in R provided for analyzing those and corresponding G-theory designs.

Most of these referenced studies assume that the objects of measurement are crossed with the other sources of measurement error. In this paper, we will focus on multivariate Generalizability Theory designs in which facet conditions are nested under the objects of measurement (Brennan, 2001).

Multivariate nested designs

We observe a multivariate model design in which each 'person' *p* is observed during a (varying) set of occasions receiving rating *i*, and each rating *i* is associated with only one observation type *m*. We regard person *p* as the object of measurement, rating *i* as random facet and observation type *m* as fixed facet. Figure 1 presents a graphical representation of this multivariate nested design, denoted by $i^{\circ}:p^{\bullet}$. Additionally, the relative error variance σ^2 (δ) captures the variability attributed to effects that include both the facet of interest and the object of measurement.

Figure 1. Graphical representation of the multivariate nested design.



This design is common in situations where learners gain expertise and competence during practice at the workplace. The concept of programmatic assessment (Heeneman et al., 2021), particularly exemplified by workplace-based assessment (WBA; Driessen et al., 2012), adopts a sophisticated methodology aimed at comprehensive learner evaluation. This approach seeks to glean insights into the capabilities of a learner *p* from diverse sources within unique, authentic professional settings. Commencing during the undergraduate phase, learners undergo frequent monitoring of their performance through a spectrum of assessments, each providing them with constructive and timely feedback within genuine contexts.

Feedback providers or assessors in this framework are teachers and professionals, each providing individualized feedback on a unique and specific authentic occasion of which the rating is denoted by *i*. Not only the number of occasions, and thus ratings, and assessments but also the number of assessors could vary between learners, with one learner for example receiving feedback or assessment from thirty and another one from only three different assessors. There are generally no unique learner-assessor combinations, as each assessor could assess one or multiple students and on several authentic occasions, or the assessor being unknown which is common in in Multisource Feedback (MSF; Lockyer & Sargeant, 2022). In addition, anonymous patients or clients can provide feedback on the performance of a learner during an interaction at the learner's workplace.

Various assessment tools, denoted by m, are deployed, each concentrating on distinct aspects and goals. For example, in medical education, performances are assessed using mini clinical evaluation exercises (MINICEXs), case-based discussions (CBDs), objective structured assessments of technical skill (OSATS), and Multisource Feedback (MSF), among others. These assessments provide valuable information on performance across diverse settings, contexts, and courses, with varying levels of stakes attributed to them. Importantly, while the assessments may differ in focus, they are harmonized within an overarching educational framework.

Crucially, learners' evaluation and guidance transcend the traditional singular examination approach. Instead, their progress is appraised based on the aggregated results derived from diverse assessments over longer periods of time in different learning contexts, contributing to a more nuanced and comprehensive understanding of their capabilities within the specified educational framework. Ensuring the quality of decision-making in the assessment process is paramount. Therefore, we aim to ensure that the workplace-based assessment outcomes accurately reflect the true level of learners' performance. To achieve this, we assess the reliability of the combined set of different assessment tools, as introduced by Moonen-van Loon et al. (2013).

While our primary focus is on workplace-based assessment, the $i^{\circ}{:}p^{\bullet}$ design and its variants have broad applications in multivariate nested designs. As outlined by Brennan (2001), this structure is particularly useful in contexts where individual measurements are grouped within higher-level units. For example, in educational assessment, students (*p*) may complete multiple essays graded by different raters resulting in grades (*i*) as part of a writing proficiency exam. In a balanced design, each student would have the same number of essays rated by an equal number of assessors, ensuring comparability in variance estimates. However, an unbalanced version may occur if students submit varying numbers of essays or if some raters grade more essays than others, introducing complexities in the reliability analysis. Similarly, in medical diagnostics, patients (*p*) undergoing diagnostic tests (*i*) may be assessed by multiple radiologists to ensure diagnostic consistency. A balanced design would involve each patient receiving evaluations from an equal number of radiologists, whereas an unbalanced design could arise if different radiologists read varying numbers of scans due to workload distribution. Another application arises in performance-based workplace evaluations, where employees (*p*) receive feedback, or ratings (*i*), from multiple supervisors or peer reviewers over time. Balanced designs might occur in structured review systems, such as standardized 360-degree feedback assessments, whereas unbalanced scenarios emerge when employees receive differing amounts of feedback due to role-specific variations or fluctuating workloads. These diverse applications illustrate the $i^{\circ}{:}p^{\bullet}$ design's flexibility in modeling complex, real-world measurement structures while maintaining robust reliability estimation.

Structure of the paper

In this paper, we introduce the `compositeReliabilityInNestedDesigns` package in R (R Core Team, 2019), a free software environment for statistical computing and graphics. The package allows researchers to conveniently input raw datasets that fit to the multivariate nested design $i^{\circ}{:}p^{\bullet}$ possibly naturalistic and unbalanced, and obtain comprehensive results from G- and D-studies in Generalizability theory, along with visually informative graphs. Notably, the easy-to-use code supports the inclusion of weights for different numbers of *m*, enabling the determination of overall composite reliability. Furthermore, the package contains optimization functionalities in which the optimal weighting is presented, both to maximize the composite reliability and to minimize the standard error of measurement (SEM). These features enhance the package's robustness, making it suitable for analyzing naturalistic and unbalanced datasets.

Section 2 formally describes the model, composite reliability, assumptions and requirements for applying the package to a dataset containing educational or professional development data in a programmatic, workplace-based assessment approach. In Section 3, we present the practical application of the `compositeReliabilityInNestedDesigns` package on a dataset containing educational data of International Medical Graduates; foreign doctors who follow a workplace-based assessment program to qualify for practicing medicine in Australia (Nair et al., 2017; Nair et al., 2021).

Implementation

Each person *p* may have one or more ratings *i* of each type *m*, where the number of ratings per person can vary, also among the types. Every rating *i* is 'scored' on the same scale. Consequently, the object of measurement *p* is crossed with the fixed multivariate variable *m*, and the facet *i* is nested within the fixed multivariate variables as each student is administered a unique sample of assessments, denoted as $i^{\circ}{:}p^{\bullet}$. Multiple universe score (co-)variances are estimated across types and error score variances. Assumptions of independent item sampling and uncorrelated residual effects lead to zero error score co-variances. Because of the varying number of raters, the absence of rater-person combinations and the anonymity in many cases, the raters are not represented by a facet in this model.

G-study

In generalizability theory, the initial step involves defining the universe of scores and facets that we aim to generalize across. For each person *p*, the universe score represents the expected value of the mean score. In a multivariate design, a universe score exists not only for each individual measure but also for the composite across all measures. The variance of universe scores across all persons *p* in the population is referred to as the universe score variance $\sigma^2(p)$, which conceptually aligns with the notion of true score variance in classical test theory. Additionally, the relative error variance $\sigma^2(\delta)$ captures the variability attributed to interactions between facets but excludes main effects.

To assess the reliability/generalizability of each individual observation type *m*, we calculate the reliability coefficient – a measure of the proportion of true variance present in a set of raw test scores (Aron & Aron, 2003). This coefficient is defined as the ratio of the universe score variance to the sum of the universe score variance and the error variance,

$$E\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p) + \sigma^2(\delta)}$$

where the relative and absolute error variance are equal in this design because the *i* and *i:p* terms are confounded.

The `compositeReliabilityInNestedDesigns` package determines the observed variance $\hat{\Sigma}_m^2(p)$ and covariance $\hat{\Sigma}_{mm\prime}(p)$ components of the persons *p* and ratings *i* nested in *p*, $\hat{\Sigma}_m^2(i\colon p)$, presented for all observation types *m*, *m'*. These components, calculated using package '`lme4`' (Bates et al., 2023), offer insights into the variability present within the data.

For each observation type m, the observed universe score variance corresponds to $\hat{\Sigma}_m^2(p)$ and the observed error score is calculated as $\hat{\Sigma}_m^2(\delta) = \frac{\hat{\Sigma}_m^2(i:p)}{n_m}$, where n_m denotes the number of ratings for observation type *m*. As each person *p* in the population may have a different number of ratings of type *m*, n_m (p), we employ the harmonic mean to determine the effective number of ratings n_m. Let $|P_m|$ be the number of persons with at least one rating of type *m*, then the harmonic mean for this type *m* is defined as $n_m = \frac{|P_m|}{\sum_{p=1}^{|P_m|} \frac{1}{n_m(p)}}$.

The estimated reliability coefficient is obtained by dividing the universe score variance by the sum of the universe score variance and the error variance. Additionally, the standard error of measurement (SEM) is calculated as the square root of the observed error score.

D-study

As previously discussed, the number of ratings *i* per observation type *m* plays a crucial role in determining reliability. In the D-study conducted by package `compositeReliabilityInNestedDesigns`, the reliability and the standard error of measurement (SEM) are graphically presented for various numbers of ratings per type.

Relating this to the practical setting, introduced in Section 1.1, these graphs provide valuable insights into how the reliability is affected when more or fewer assessments (*i*) per observation type (*m*) are required or recommended in an educational program. By visualizing the relationship between the number of assessments and reliability, educators and assessment designers can gain a deeper understanding of the impact of assessment quantity on the overall assessment quality for each assessment tool and make informed decisions regarding assessment program requirements and recommendations. For an example, see Figure 2 and Figure 3 in Section 3.2.

Composite reliability

To estimate the composite reliability of person *p* across ratings *i* with different observation types *m*, we employ multivariate generalizability theory. This approach enables the integration of ratings *i* from diverse observation types *m* to generate a comprehensive evaluation of person *p*'s observations. In addition to considering the observed variance $\hat{\Sigma}_m^2(p)$ for each observation type *m* and the observed error score $\hat{\Sigma}_m^2(\delta)$, we also need to include the covariance $\hat{\Sigma}_{mm'}(p)$ components representing the interrelationships between person *p* for different observation types *m* and *m'*.

When combining multiple observation types to calculate the composite reliability, each type m is assigned a positive weight *w_m*, where the sum of all the weights is equal to 1. If the weights are unknown, it is common practice to assign equal weights to each observation type. However, in many practical (educational) settings, there may already be predetermined weights or importance assigned to each observation type *m*.

For a specific number of ratings per observation type *m*, the composite universe score variance is the weighted sum of all the elements of the variance and covariance components. The estimated reliability coefficient $E\hat{\rho}^{\ 2}$ is defined as the ratio of the composite universe score variance to the sum of the composite universe score variance and the composite relative error variance:

$$E\hat{\rho}^{\ 2} = \frac{\sum_m w_m^2 \hat{\Sigma}_m^2(p) + \sum_m \sum_{m' \neq m} w_m w_{m'} \hat{\Sigma}_{mm'}(p)}{\sum_m w_m^2 \hat{\Sigma}_m^2(p) + \sum_m \sum_{m' \neq m} w_m w_{m'} \hat{\Sigma}_{mm'}(p) + \sum_m w_m^2 \hat{\Sigma}_m^2(\delta)}$$

The standard error of measurement (SEM) is equal to $\sqrt{\sum_m w_m^2 \hat{\Sigma}_m^2(\delta)}$.

By specifying different weight distributions, users can observe the resulting composite reliability coefficient and SEM for each possibility. This flexibility allows for the examination of how different weight assignments affect the overall outcomes and provides a means to tailor the setting (e.g., assessment program) to specific requirements or preferences.

By definition, when the standard error of measurement (SEM) is reduced, it leads to a narrower confidence interval. This reduction indicates a decrease in the expected distribution of error around the mean scores, which is an important objective in e.g. assessment design and evaluation. By minimizing the SEM, we aim to enhance the precision and accuracy of the results. A smaller SEM signifies less random error in the measurement process, resulting in more reliable and precise estimates of persons' true scores. This reduction in measurement error allows for a more confident interpretation of the outcomes and enhances the overall quality and validity of the setting at hand.

To minimize the SEM by adjusting the weights, subject to the constraint that the sum of (positive) weights equals 1, i.e. minimize $\sqrt{\sum_m w_m^2 \hat{\Sigma}_m^2(\delta)}$ subject to $\sum_m w_m = 1$, we use the Lagrange multiplier $\mathcal{L}(w_1, \dots, w_{|M|}, \mu) = \sum_m w_m^2 \hat{\Sigma}_m^2(\delta) - \mu(\sum_m w_m - 1)$, where $|M|$ is the number of different observation types. For each type *m*, we set the derivative equal to 0, thus

$$\frac{\partial \mathcal{L}}{\partial w_m} = 2w_m \hat{\Sigma}_m^2(\delta) - \mu = 0 \Leftrightarrow 2w_m \hat{\Sigma}_m^2(\delta) = \mu, \forall m$$

$$\frac{\partial \mathcal{L}}{\partial \mu} = -\left(\sum\nolimits_m w_m - 1\right) = 0 \Leftrightarrow \sum\nolimits_m w_m = 1$$

Using the first equality, rewriting gives

$$2w_m\hat{\Sigma}_m^2(\delta) = 2w_1\hat{\Sigma}_1^2(\delta) \Leftrightarrow w_m = \frac{w_1\hat{\Sigma}_1^2(\delta)}{\hat{\Sigma}_m^2(\delta)}, \forall m > 1$$

Rewriting the constraint of the minimization function $\sum_m w_m = 1$, using the equation above, yields

$$\sum\nolimits_m w_m = w_1 + \sum_{m>1} \frac{w_1\hat{\Sigma}_1^2(\delta)}{\hat{\Sigma}_m^2(\delta)} = \left(1 + \sum_{m>1} \frac{\hat{\Sigma}_1^2(\delta)}{\hat{\Sigma}_m^2(\delta)}\right) w_1 = 1$$

$$w_1 = \left(1 + \sum_{m>1} \frac{\hat{\Sigma}_1^2(\delta)}{\hat{\Sigma}_m^2(\delta)}\right)^{-1} = \left(\sum_m \frac{\hat{\Sigma}_1^2(\delta)}{\hat{\Sigma}_m^2(\delta)}\right)^{-1}$$

Concluding, to minimize the composite standard error or measurement, weights are set to $w_1 = \left(\sum_m \frac{\hat{\Sigma}_1^2(\delta)}{\hat{\Sigma}_m^2(\delta)}\right)^{-1}$ and $w_m = \frac{\hat{\Sigma}_1^2(\delta)}{\hat{\Sigma}_m^2(\delta)} w_1$ for all $m > 1$.

However, minimizing the SEM is not necessarily equal to maximizing composite reliability; a test can have high composite reliability but still have a high SEM if overall score variability is large. Likewise, reducing SEM does not always improve composite reliability, especially if the test has heterogeneous items. It depends on the exact setting which measure is preferred over the other. Minimizing SEM is preferred when absolute precision of individual scores is important. For example, in high-stakes testing or diagnostic assessments, reducing SEM ensures that observed scores closely approximate true ability levels, improving decision accuracy. Maximizing composite reliability is preferred when the goal is to assess internal consistency in multi-item measures. This is crucial in research settings where constructs are measured using multiple items (e.g., psychological scales, surveys). Higher composite reliability suggests that items are reliably capturing the intended construct. Because of this trade-off, the `compositeReliabilityInNestedDesigns` package provides a possibility to determine the weights that maximize the composite reliability coefficient, using function `solnp` for the nonlinear optimization using augmented Lagrange method of package `Rsolnp` (Ghalanos & Theussl, 2015).

Application and Discussion

In this section, we illustrate the functionality of the `compositeReliabilityInNestedDesigns` package using an educational dataset. The context for our study is the multisource feedback (MSF) part of the assessment program for the licensing requirements of International Medical Graduates (IMGs) in Australia. The IMGs originate from many different countries including Egypt, India, Sweden, South Africa, Brazil, Sri Lanka, Pakistan, Myanmar Burma, Iraq, among others. The dataset consists of all MSF assessments submitted between July 2019 and February 2024. Each form required assessors to evaluate trainee performance using a predefined set of criteria over the past 1 or 6 months, using a 5-point rating scale, and is filled out either by a Medical Colleague or another Co-worker in health care. The questionnaire is slightly different for the two types of assessors, but similar for the 1- and 6-month period. The assessors of an individual IMG are typically different in each period. The dataset comprises 1577 assessments of 130 IMGs.

In terms of Multivariate Generalizability Theory, each IMG *p* may have one or more graded assessments *i* of each assessment tool *m*, where the number of assessments per IMG can vary, also among the assessment tools. Every assessment *i* is scored on the same 1-5 Likert scale, of which the average was calculated.

G-study

The GStudy() function is designed to assess the reliability coefficient and the standard error of measurement (SEM) for each assessment tool. This function utilizes the harmonic mean of the number of assessments per tool as a measure of effective assessment quantity. By providing a dataset as input, users can obtain reliable estimates of the reliability coefficient and SEM for each assessment tool. The desired number of decimal places in the output (`nrDigitsOutput`) can be added as input.

```
R> library(compositeReliabilityInNestedDesigns)
> GStudy(mydata,nrDigitsOutput=4)
```

The second code line facilitates the execution of the G-study, which encompasses the calculation of composite reliability. It generates a descriptive statistics matrix. In our current example, the output is presented in Table 1, offering valuable information for each assessment type "Medical Colleague – Month 1" (MC1), "Co-Worker – Month 1" (CO1), "Medical Colleague – Month 6" (MC6) and "Co-Worker – Month 6" (CO6). The table presents the number of assessments (NrAssessments) and IMGs (NrStudents), the average score (MeanScore), and the standard deviation (StDev) of the assessment scores. Additionally, the table provides insights into the average number of assessments per student (AvgNrAssessments) and the harmonic mean (HarmonicMean) of the number of assessments per student.

Table 1. Describing statistics of the dataset for each of the four different assessment types.

	CO1	CO6	MC1	MC6
NrAssessments	395	395	394	393
NrAssessees	130	130	130	130
MeanScore	4.1849	4.31	3.8079	3.9047
StDev	0.6969	0.652	0.6809	0.6591
AvgNrAssessments	3.0385	3.0385	3.0308	3.0231
HarmonicMean	3.0116	2.9828	2.9771	2.9658
MaxNrAssessments	6	6	6	6
MinNrAssessments	2	1	1	1
MedianNrAssessments	3	3	3	3
ReliabilityCoeff_HM	0.2641	0.1242	0.4263	0.4254
SEM_HM	0.3798	0.3688	0.3534	0.3423

The harmonic mean for a given assessment tool in the synthetic dataset is determined by dividing the number of students who have at least one assessment of that tool by the sum of the reciprocals of the number of assessments per student.

Overall, these descriptive statistics and harmonic mean calculations provide valuable insights into the distribution of assessments and student performance within the dataset, contributing to a comprehensive understanding of the assessment program's characteristics and dynamics.

To view the observed variances, covariances, and error scores, the dataset, along with a vector specifying the number of assessments per assessment tool, can be added as input to the calculateVarCov() function.

This function plays a crucial role in the comprehensive analysis provided by the package, working in conjunction with other functions to offer a complete suite of statistical tools for researchers. While the function is an integral component of the package's functionality, it is presented here to address the needs of researchers seeking detailed insights into the measurement properties of the dataset.

```
R> varcov <- calculateVarCov(mydata, c("CO1"=3.0116, "CO6"=2.9828,
"MC1"=2.9771, "MC6"=2.9658))
> varcov$S_p
> varcov$S_iINp
> varcov$S_delta
```

Table 2 showcases the observed variance $\hat{\Sigma}_m^2(p)$ and covariance $\hat{\Sigma}_{mm'}(p)$ components of the students *p* and assessment scores *i* nested within students $\hat{\Sigma}_m^2(i:p)$ for each assessment tool *m* and *m'* within the dataset. The table provides a comprehensive overview of the variability and interrelationships among the assessment scores, enabling a detailed examination of the measurement properties associated with each assessment tool.

Specifically, for assessment tool MC6, the observed universe score ($\hat{\Sigma}_{MC6}^2(p)$) is equal to 0.0867, and the observed error score is equal to $\hat{\Sigma}_{MC6}^2(\delta) = \frac{\hat{\Sigma}_{MC6}^2(i:p)}{n_{MC6}} = \frac{0.3475}{2.9658} = 0.1172$, giving estimated reliability coefficient $E\hat{\rho}_{MC6}^2 = \frac{0.0867}{0.0867+0.1172} = 0.4253$, and SEM = $\sqrt{0.1172} = 0.3423$, as shown in Table 1.

Table 2. Observed variance and covariance components.

$\hat{\Sigma}_m^2(p)$	**CO1**	**CO6**	**MC1**	**MC6**
CO1	0.0518	0.0384	0.0607	0.0624
CO6	0.0384	0.0193	0.0328	0.0554
MC1	0.0607	0.0328	0.0928	0.0550
MC6	0.0624	0.0554	0.0550	0.0867

$\hat{\Sigma}_m^2(i:p)$	**CO1**	**CO6**	**MC1**	**MC6**
CO1	0.4345	0	0	0
CO6	0	0.4058	0	0
MC1	0	0	0.3719	0
MC6	0	0	0	0.3475

D-study

As previously discussed, the number of assessments per tool has a notable impact on the reliability estimation. To investigate the relationship between the number of assessments and reliability, the D-study functionality within the program provides an analysis of the reliability coefficient and standard error of measurement (SEM) for varying numbers of assessments per tool. To execute the D-study and obtain these insights, researchers can utilize

```
R> plots <- DStudy(mydata, maxNrAssessments = 60)
> plots$plotRel
> plots$plotSEM
```

The results for the synthetic dataset, spanning up to 60 assessments per assessment tool, are graphically presented in Figure 2 and Figure 3. Each graph depicts the number of assessments per tool on the x-axis. The first graph illustrates the reliability coefficient value on the y-axis, while the second graph represents the standard error of measurement (SEM).

These figures display the practical implications of varying assessment quantities on reliability and SEM, offering a visual representation of the relationship between these factors. This information can guide the

Figure 1. Reliability coefficient for each of the assessment tools for varying numbers of assessments (x-axis).

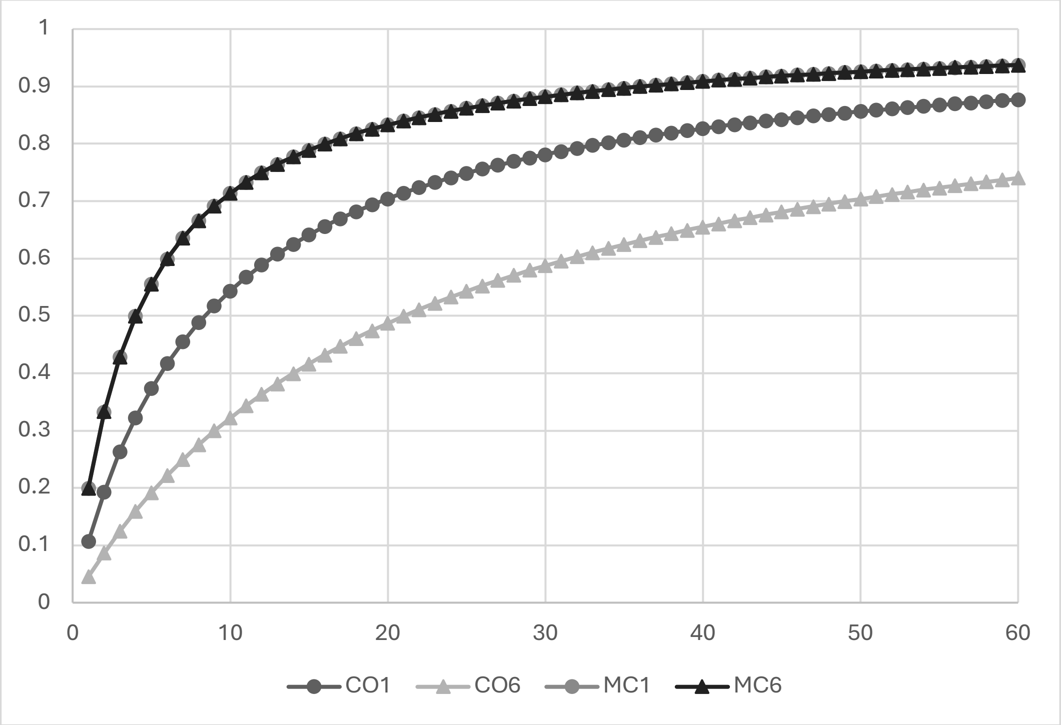
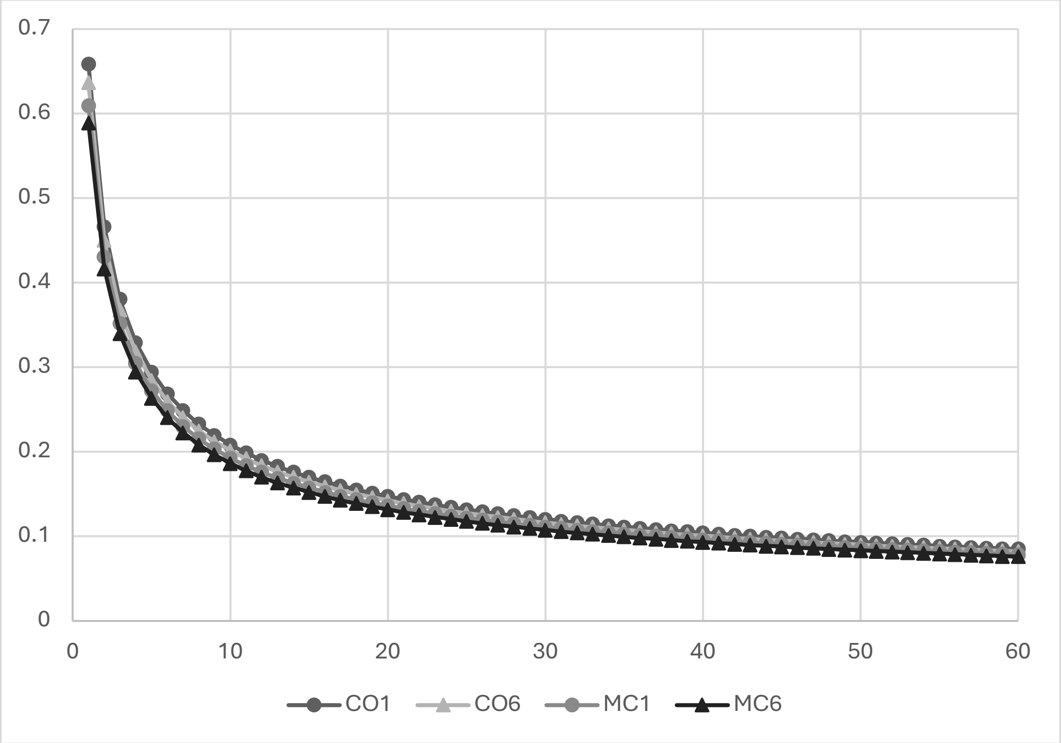


Figure 2. SEM for each of the assessment tools for varying numbers of assessments (x-axis).



development of assessment programs that strike a balance between reliability, precision, and efficiency. Solely in terms of reliability, it seems that the assessments of medical colleagues outperform the assessments of other co-workers. Although candidates are assessed on a similar scale, each group of assessors views a different aspect of the candidate's performance. Therefore, a combination of the assessments in the assessment program is highly beneficial to just leaning on either one of the tools.

Composite reliability

The estimation of composite reliability, encompassing the integration of student results from assessments of various tools, is performed utilizing multivariate generalizability theory. Leveraging the capabilities of `compositeReliabilityInNestedDesigns`, researchers can obtain a comprehensive analysis that includes the determination of weights optimizing the standard error of measurement (SEM) and the subsequent presentation of the composite reliability value.

The R package provides researchers with the flexibility to input different sets of numbers of assessments and weights as data, enabling the estimation of the composite reliability coefficient and the standard error of measurement (SEM). These metrics serve as crucial indicators of the extent to which the assessments accurately capture students' true performance across various assessment tools.

For the current educational program for IMGs, they are advised to collect 3 assessments per type. Currently, the weight attached to each assessment tool is equal.

```
R> compRel <- computeCompositeReliability(mydata,
+ n=c("CO1"=3,"CO6"=3,"MC1"=3,"MC6"=3),
+ weights = c("CO1"=0.25,"CO6"=0.25,"MC1"=0.25,"MC6"=0.25),
+ optimizeSEM=FALSE)
> compRel$reliability
> compRel$SEM
```

The composite reliability is 0.6232 and the SEM is 0.1803.

If the Boolean value optimizeSEM is set to TRUE, the package offers a corresponding set of weights that minimizes the SEM. These weights represent the relative importance assigned to each assessment tool when evaluating the overall performance of the student. By optimizing the weights, the precision and accuracy of the assessment outcomes are enhanced, effectively minimizing measurement error. The utilization of this weight optimization process provides researchers with a more refined and precise assessment of students' performance. By effectively minimizing measurement error, the package ensures that the assessment outcomes reflect the true abilities and achievements of the students better.

For the current example, running the same code with optimizeSEM=TRUE, improves the composite reliability to 0.6301 and SEM to 0.1796. The output `compRel$weights` gives 0.2227 for CO1, 0.2385 for CO6, 0.2603 for MC1, and 0.2785 for MC6.

By leveraging multivariate generalizability theory and the functionalities provided by the R package, researchers gain valuable insights into the composite reliability of student assessments. This allows for a comprehensive evaluation of student performance and supports informed decision-making in educational assessment contexts.

As an illustration, considering the number of assessments per tool depicted in the first two columns of Table 3, along with the corresponding weights in the subsequent two columns, the final two columns of the table display the composite reliability and composite standard error of measurement (SEM) resulting from this combination. This presentation allows researchers to examine the impact of different numbers of assessments and associated weights on the overall composite reliability and SEM. By analyzing the values presented in the last two columns, researchers can gain insights into the effectiveness and precision of the composite assessment outcomes. This information aids in understanding the interplay between the number of assessments, weight distribution, and the resulting composite reliability and SEM, thereby supporting informed decision-making in assessment design and evaluation.

Table 3. Composite reliability with various numbers of assessments and weights per tool.

CO1	**CO6**	**MC1**	**MC6**	**CO1**	**CO6**	**MC1**	**MC6**	**CompRel**	**CompSem**
N	**n**	**n**	**n**	**w**	**w**	**n**	**n**	**CompRel**	**CompSem**
3	3	3	3	0.25	0.25	0.25	0.25	0.6232	0.1803
3	4	3	4	0.25	0.25	0.25	0.25	0.6529	0.1690
3	3	4	4	0.25	0.25	0.25	0.25	0.6516	0.1695
4	4	4	4	0.20	0.20	0.30	0.30	0.6998	0.1568
6	6	6	6	0.20	0.20	0.30	0.30	0.7776	0.1280

For the same number of assessments, Table 4 presents the calculated weights per tool that effectively minimize the SEM, as well as the resulting composite reliability and SEM values associated with these optimized weights. When comparing the resulting reliability and SEM values in the last two rows of Table 3 and Table 4, it becomes apparent that minimizing the SEM does not necessarily maximize the composite reliability, as described in Section 2.3.

Table 4. Composite reliability with various numbers of assessments where the weights minimize the SEM.

CO1	**CO6**	**MC1**	**MC6**	**CO1**	**CO6**	**MC1**	**MC6**	**CompRel**	**CompSem**
N	**n**	**n**	**n**	**Opt w**	**Opt w**	**Opt w**	**Opt w**	**CompRel**	**optCompSEM**
3	3	3	3	0.2227	0.2385	0.2603	0.2785	0.6301	0.1796
3	4	3	4	0.19000	0.2712	0.2220	0.3168	0.6639	0.1659
3	3	4	4	0.1888	0.2022	0.2942	0.3148	0.6777	0.1654
4	4	4	4	0.2227	0.2385	0.2603	0.2785	0.6843	0.1555
6	6	6	6	0.2227	0.2385	0.2603	0.2785	0.7730	0.1270

For the any number of assessments, the package allows to determine the weights per assessment tool to maximize the composite reliability coefficient.

```
R> compMaxRel <- computeMaxcompositeReliability(mydata,
> + n=c("CO1"=3,"CO6"=3,"MC1"=3,"MC6"=3))
> compMaxRel$reliability
> compMaxRel$SEM
> compMaxRel$weights
```

Table 5 presents the calculated weights per tool that effectively maximize the composite reliability coefficient, as well as the resulting SEM values associated with these optimized weights.

Table 5. Composite reliability with various numbers of assessments where the weights maximize the composite reliability.

CO1	**CO6**	**MC1**	**MC6**	**CO1**	**CO6**	**MC1**	**MC6**	**CompRel**	**CompSem**
N	**n**	**n**	**n**	**Opt w**	**Opt w**	**Opt w**	**Opt w**	**CompRel**	**optCompSem**
3	3	3	3	0.2176	0.1641	0.2904	0.3279	0.6385	0.1828
3	4	3	4	0.1861	0.1913	0.2385	0.3841	0.6729	0.1691
3	3	4	4	0.1803	0.1366	0.3233	0.3598	0.6845	0.1679
4	4	4	4	0.2176	0.1641	0.2904	0.3279	0.7019	0.1583
6	6	6	6	0.2176	0.1641	0.2904	0.3279	0.7794	0.1292

The information conveyed in Table 4 and Table 5 facilitates the evaluation and comparison of different weight distributions, shedding light on the impact of weight optimization on the reliability and measurement error of the composite assessment. This empirical analysis supports researchers in making informed decisions regarding the selection and assignment of weights, leading to improved assessment practices and more robust interpretations of student performance.

As observed in the graphical output of the D-study (Figure 2 and Figure 3), to reach a reliability of 0.70, the individual assessment tools CO1, MC1, CO6 and MC6 require 20, 10, 50, and 10 assessments, respectively, totaling 90 assessments when considered separately. However, when these assessment tools are combined, the results displayed in Table 5 reveal that a mere four assessments per tool (16 in total) reach the same composite reliability.

This finding highlights the efficiency and effectiveness of combining multiple assessment tools to achieve a desired level of reliability. By leveraging the complementary strengths of different assessment tools, a significant reduction in the overall number of assessments can be achieved while maintaining the desired reliability threshold. This information is invaluable for assessment program planning and resource allocation, enabling educators and researchers to optimize the assessment process and minimize the burden on students and assessors without compromising the reliability of the outcomes.

Assumptions and Requirements

Generalizability theory, as a fundamental framework, typically assumes local independence of occasions or ratings, implying that each data point is independent of others. However, this assumption may not hold in authentic educational settings, where assessments can affect subsequent performances due to feedback provided to students. Despite this violation, such conditions are common in programmatic, workplace-based assessment settings. This underscores the need to interpret reliability indices within the dynamic nature of skill development. In this context, reliability estimates reflect the consistency of performance assessments within a structured learning trajectory rather than a static measurement framework. We accept this violation as the primary objective in such a setting is to differentiate between overall student performances over an extended period.

When combining different assessment tools, each focusing on specific aspects in education, to evaluate the performance of the student using multivariate generalizability theory, the students should be graded on the same rating scale using the same assessment standard throughout all assessments.

The data set utilized in the `compositeReliabilityInNestedDesigns` package must adhere to specific criteria. It should be an R dataframe with three columns labeled “ID”, “Type”, and “Score”. The

"Score" column must contain numeric data. To estimate the composite reliability of multiple assessment tools, each student must receive a grade for each assessment tool at least once. When including the desired number of assessments per tool, each tool must be included and receiving a numeric value. Furthermore, the sum of positive weights assigned to the assessment tools must be equal to one, ensuring a proper weighting scheme for the composite reliability calculation.

Conclusions

The `compositeReliabilityInNestedDesigns` R package provides a user-friendly and efficient solution for calculating composite reliability in settings with a multivariate *nested* design, particularly those characterized by naturalistic and unbalanced datasets. By extending the scope beyond the reliability of individual types, the package offers significant advancements in terms of feasibility to reach reliable results. It enables researchers to analyze the composite reliability of a comprehensive combination of types, employing multivariate generalizability theory and weighted optimization.

Relating this to the practical setting of programmatic workplace-based assessment programs, introduced in Section 1.1, utilizing the package, researchers gain flexibility in adjusting the WBA program according to specific requirements, while also accommodating the need to assign varying degrees of importance to different assessment tools. It captures the interplay between multiple assessment tools and their impact on the overall evaluation of student performance when (formative) assessment occurs in authentic workplace settings. This flexibility empowers users to tailor the assessment program to their specific needs, ensuring that certain tools of assessment receive appropriate emphasis and consideration within the overall educational framework.

For future improvements, we plan to incorporate raters as an additional facet, specifically for scenarios where raters are known and some provide multiple ratings, but without forming a fully crossed design. Another planned development is the creation of a Shiny app or web-based GUI to make the package more accessible to users unfamiliar with R, allowing for intuitive data input and clear visualization of composite reliability results.

In conclusion, package `compositeReliabilityInNestedDesigns` serves as a valuable tool for researchers and practitioners, e.g., those involved in educational assessment programs. Its ease of use, ability to handle naturalistic and unbalanced datasets, and incorporation of multivariate generalizability theory and weighted optimization offer a solution for analyzing the composite reliability of a diverse range of (programmatic assessment) programs. With its flexible features, the package empowers users to fine-tune their (assessment) programs and derive more robust and meaningful insights into persons' observations, e.g., students' performance.

Availability and requirements

Project name: `compositeReliabilityInNestedDesigns`

Project home page: https://github.com/jmoonen/compositeReliabilityInNestedDesigns

Operating system(s): Platform independent

Programming language: R

Other requirements: R 4.3.1 or higher

License: GNU General Public License version 3

Any restrictions to use by non-academics: no

List of abbreviations

G-study	Generalizability study
GT	Generalizability Theory
D-study	Decision study
MGT	Multivariate Generalizability Theory
SEM	Standard error of measurement
WBA	Workplace Based Assessment
MSF	Multi-Source Feedback

Declarations

Ethics approval and consent to participate

The use of the dataset was approved by the Health Services Research and Ethics committee of the Health Service (approval number A.U.- 201607-03).

Consent for publication

Not applicable.

Availability of data and materials

A small demo dataset is available in the GitHub repository, https://github.com/jmoonen/compositeReliabilityInNestedDesigns/tree/master/data. This demonstration dataset is also included in the compositeReliabilityInNestedDesigns package.

Results on another extensive dataset are presented by Nair et al. (2021).

Competing interests

The authors declare that they have no competing interests

Funding

Not applicable.

Authors' Contributions

The first author developed the software and wrote the manuscript. The second author provided input, contributed to the manuscript, and extensively tested the package.

Acknowledgements

The authors thank Prof. Dr. Cees van der Vleuten, Dr. Peter van Rosmalen, and Prof. Dr. Pim Teunissen for their valuable comments on the manuscript. The author thanks Prof. Dr. Balakrishnan Nair for sharing the dataset.

Received: 10/24/2023. **Accepted:** 4/15/2025. **Published:** 5/21/2025.

Citation: Moonen-van Loon, J. M. W., & Donkers, J. (2025). An R Package for Optimizing the Composite Reliability in Multivariate Nested Designs. *Practical Assessment, Research, & Evaluation*, 30(4). Available online: https://doi.org/10.7275/pare.2104

Corresponding Author: Joyce M. W. Moonen-van Loon, Maastricht University. Email: j.moonen@maastrichtuniversity.nl

References

Aron, A., & Aron, E. N. (2003). *Statistics for psychology* (3rd ed. ed.). Prentice Hall.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., Dai, B., Scheipl, F., Grothendieck, G., Green, P., Fox, J., Bauer, A., & Krivitsky, P. N. (2023). *lme4: Linear Mixed-Effects Models using 'Eigen' and S4*. In https://github.com/lme4/lme4/

Bolker, B. (2024). Getting started with the glmmTMB package. Retrieved July 19, 2024, from https://cran.r-project.org/web/packages/glmmTMB/vignettes/glmmTMB.pdf

Brennan, R. L. (2001). *Generalizability theory*. Springer.

Brennan, R. L. (2010). Generalizability Theory. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education*. Elsevier Ltd.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. Wiley.

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

Crossley, J., Davies, H. A., Humphris, G., & Jolly, B. (2002). Generalisability: a key to unlock professional assessment. *Medical Education*, *36*(10), 972-978.

Ghalanos, A., & Theussl, S. (2015). *Rsolnp: General Non-Linear Optimization*. In https://cran.r-project.org/web/packages/Rsolnp/

Hays, R. B., Fabb, W. E., & van der Vleuten, C. P. M. (1995). Reliability of the fellowship examination of the royal Australian college of general practitioners. *Teaching and Learning in Medicine*, *7*, 43-50.

Heitman, R. J., Kovaleski, J. E., & Pugh, S. F. (2009). Application of generalizability theory in estimating the reliability of ankle-complex laxity measurement. *Journal of Athletic Training*, *44*(1), 48-55. https://doi.org/10.4085/1062-6050-44.1.48

Huebner, A., & Lucht, M. (2019). Generalizability Theory in R. *Practical Assessment, Research, and Evaluation*, *24*(5).

Jiang, Z., Raymond, M., Shi, D., & DiStefano, C. (2020). Using a linear mixed-effect model framework to estimate multivariate generalizability theory parameters in R. *Behavior Research Methods*, *52*, 2383-2393. https://doi.org/10.3758/s13428-020-01399-z

Li, M., Shavelson, R. J., Yin, Y., & Wiley, E. (2015). Generalizability Theory. In. https://doi.org/10.1002/9781118625392.wbecp352

Lockyer, J., & Sargeant, J. (2022). Multisource feedback: an overview of its use and application as a formative assessment. *Canadian Medical Education Journal*, *13*(4), 30-35.

Monteiro, S., Sullivan, G. M., & Chan, T. M. (2019). Generalizability Theory Made Simple(r): An Introductory Primer to G-Studies. *Journal of Graduate Medical Education*, *11*(4), 365-370. https://doi.org/10.4300/JGME-D-19-00464.1

Moonen-van Loon, J. M. W., Overeem, K., Donkers, H. H. L., van der Vleuten, C. P. M., & Driessen, E. W. (2013). Composite reliability of a workplace-based assessment toolbox for postgraduate medical education. *Advances in Health Science Education*, *46*, 28-37.

Moore, C. T. (2016). *Apply Generalizability Theory with R: Package 'gtheory'*. In

Nair, B. R., Moonen-van Loon, J. M. W., Parvathy, M., Jolly, B. C., & van der Vleuten, C. P. M. (2017). Composite reliability of workplace-based assessment of international medical graduates. *Medical Journal of Australia*, *207*(10), 453-453.

Nair, B. R., Moonen-van Loon, J. M. W., Parvathy, M., & van der Vleuten, C. P. M. (2021). Composite Reliability of Workplace Based Assessment of International Medical Graduates. *MedEdPublish*, *10*(1), Article 104. https://doi.org/10.15694/mep.2021.000104.1

R Core Team. (2019). *R: A language and environment for statistical computing.* In R Foundation for Statistical Computing. https://www.R-project.org/

Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika*, *30*, 39-56.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2), 1-36. http://www.jstatsoft.org/v48/i02/

Sanz-Fernández, C., Morales-Sánchez, V., Castellano, J., & Hernández Mendo, A. (2024). Generalizability Theory in the Evaluation of Psychological Profile in Track and Field. *Sports*, *12*(5). https://doi.org/10.3390/sports12050127

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

Swanson, D. B. (1987). A measurement framework for performance-based tests. In I. Hart & R. Harden (Eds.), *Further developments in assessing clinical competence* (pp. 11-45). Can-Heal publication.

Vispoel, W. P., Lee, H., & Chen, T. (2024). Multivariate Structural Equation Modeling Techniques for Estimating Reliability, Measurement Error, and Subscale Viability When Using Both Composite and Subscale Scores in Practice. *Mathematics*, *12*(8), 1-25. https://doi.org/10.3390/math12081164

Vispoel, W. P., Lee, H., & Hong, H. (2024). Analyzing Multivariate Generalizability Theory Designs within Structural Equation Modeling Frameworks. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(3), 552-570. https://doi.org/10.1080/10705511.2023.2222913

Vispoel, W. P., Lee, H., Hong, H., & Chen, T. (2023). Applying Multivariate Generalizability Theory to Psychological Assessments. *Psychological Methods*, 1-23. Advance online publication. https://doi.org/10.1037/met0000606

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, *23*(1). https://doi.org/10.1037/met0000107

Appendix. Complete R Code

The following code presents the use of the package for the included mydata.rda dataset, which is a small dummy dataset. Please replace mydata your own dataset to apply the package.

```
library(compositeReliabilityInNestedDesigns)
GStudy(mydata,nrDigitsOutput=4)

varcov <- calculateVarCov(mydata, c("CO1"=3.0116, "CO6"=2.9828,
"MC1"=2.9771, "MC6"=2.9658))
varcov$S_p
varcov$S_iINp
varcov$S_delta

plots <- DStudy(mydata, maxNrAssessments = 60)
plots$plotRel
plots$plotSEM

compRel <- computeCompositeReliability(mydata,
+ n=c("CO1"=3,"CO6"=3,"MC1"=3,"MC6"=3),
+ weights = c("CO1"=0.25,"CO6"=0.25,"MC1"=0.25,"MC6"=0.25),
+ optimizeSEM=FALSE)
compRel$reliability
compRel$SEM

compMaxRel <- computeMaxcompositeReliability(mydata,
+ n=c("CO1"=3,"CO6"=3,"MC1"=3,"MC6"=3))
compMaxRel$reliability
compMaxRel$SEM
compMaxRel$weights
```