Received: 27 May 2025 | Accepted: 16 August 2026

DOI: 10.1111/bmsp.70074

the british psychological society
promoting excellence in psychology

ARTICLE

A taxonomy of reliability indices for cluster-level constructs measured by multiple individual-level indicators

Terrence D. Jorgensen | **Suzanne Jak**

University of Amsterdam, Amsterdam, The Netherlands

Correspondence
Terrence D. Jorgensen, Department of Child Development and Education, University of Amsterdam, Postbus 15776, 1001 NG Amsterdam, The Netherlands.
Email: T.D.Jorgensen@uva.nl

Funding information
Nederlandse Organisatie voor Wetenschappelijk Onderzoek, Grant/Award Number: VI.Vidi.201.009

Abstract

Generalizability theory (GT) provides a variety of intraclass correlation coefficients (ICCs) to quantify reliability across different conditions of measurement—for example, scale and interrater reliability quantify generalization across items and raters, respectively. In a multilevel modelling (MLM) context, an ICC is often used to quantify the proportion of a subject-level variable's variance that is attributable to cluster-level differences (e.g., between schools or neighbourhoods). When data from cluster members are their ratings of a group-level characteristic, the ICC is an inter-rater reliability coefficient, interpreted as the degree to which observations of a cluster generalize across members of that cluster. Structural equation modelling (SEM) is becoming a popular framework for modelling GT, and the 'omega' indices of composite reliability—which are functions of SEM parameters—are special cases of GT (i.e., how measurements of a construct generalize across indicators). The merger of MLM and SEM (i.e., ML-SEM) provides a unique opportunity to explore how well measurements of cluster-level constructs generalize across multiple measurement conditions (indicators, cluster members, or both). A taxonomy of ICCs is explicated to appropriately quantify reliability across aspects of a procedure for measuring cluster-level constructs, adding value to the limited ICCs currently considered in ML-SEM applications.

KEYWORDS

generalizability theory, inter-rater reliability, intraclass correlation coefficient, multilevel structural equation model, scale reliability

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2026 The Author(s). *British Journal of Mathematical and Statistical Psychology* published by John Wiley & Sons Ltd on behalf of British Psychological Society.

Theories of human behaviour typically involve psychological constructs that cannot be directly observed (i.e., latent constructs). Multiple-item scales or tests are typically used to capture individual differences due to a latent construct, where a scale composite (e.g., sum or average of scale items, number or proportion of correct test items) is used as an observed proxy for a latent variable. Classical test theory (CTT) has a long history of preoccupation with the reliability of such scale composites (Lord & Novick, 1968). Reliability represents the degree to which observed scores will be similar across repeated measurements (e.g., at different occasions, or using different tests) and is more formally defined in a later section. More modern test theories incorporate latent variables into statistical models—for example, person parameters in an item response theory (IRT) model, common factors in a confirmatory factor analysis (CFA) or structural equation model (SEM), or random effects in a linear mixed model (LMM). Modern measurement models for hypothesized latent constructs have enabled definitions of reliability to become more nuanced and for reliability coefficients to be estimated as functions of measurement-model parameters.

This article focusses on Generalizability Theory (GT), which is a measurement model that can be estimated using LMM (Jiang, 2007; Putka et al., 2008) or SEM (Jorgensen, 2021; Marcoulides, 1996; Raykov & Marcoulides, 2006; Vispoel et al., 2018). We apply the GT framework to a particular measurement scenario that has raised some controversy in the past few years: measuring a group-level construct using individual-level data. This measurement scenario is often encountered in scientific fields that study people who are 'clustered' into many natural groups—for example, employees nested within departments or companies; patients nested within doctors or clinics; and students nested within teachers, classes, or schools. When the construct of interest refers to a trait about the group (e.g., school climate, group norms), individual members constitute multiple indicators of the group, which compounds with multiple scale items to yield many latent variables in the same model. Controversy exists even in how constructs are conceptually defined or labelled in multilevel settings, but we refer to group-level constructs measured with error as 'shared' constructs (Jak et al., 2023; Stapleton et al., 2016; Stapleton & Johnson, 2019).

Many definitions of reliability have been proposed for multilevel data, some of which are directly relevant to the context discussed in this paper (Geldhof et al., 2014; Lai, 2021), which is when multiple cluster members serve as raters of their cluster's construct and provide measurements using multiple scale items. In GT notation, this is denoted as a $(r:c) \times i$ design, where $(r:c)$ represents the raters (r) nested in clusters (c), which are crossed with items (i). Other reliability coefficients have been defined for a highly analogous situation in which multiple item-responses are replaced by multiple raters of the subjects nested in clusters—a $(s:c) \times r$ design (as well as for more complex designs that fall beyond our scope; Lin et al., 2025; Ten Hove et al., 2022). All of the cited proposals defined level-specific scale-composite reliability *or* inter-rater reliability (IRR) coefficients for within-level constructs, where the corresponding between-level construct represents a cluster-level aggregation of the individual-level construct (i.e., a "configural" construct; Stapleton et al., 2016; Jak et al., 2023). There is generally consensus about how to define and interpret reliability coefficients for variables that represent Level-1 (or within-level) information but controversy about higher-level variables. Our paper focuses on measuring shared constructs, which are inherently group-level traits, so within-level variability represents a specific dimension of error rather than true scores.

When measuring group-level constructs with responses of group members to multiple-item scales, we argue that researchers should evaluate the degree to which both sources of error (raters and items)—both jointly and separately—can prevent accurately distinguishing between groups. Although Lai (2021) proposed reliability coefficients relevant for this situation that jointly consider both sources of error, he recommended it instead for configural constructs, proposing a very similar coefficient (but using a different model) for shared constructs. To our knowledge, the only source that brings together all relevant reliability indices for this scenario under a single umbrella is a somewhat obscure section in an authoritative GT textbook by Brennan (2001, section 4.4.4, equations 4.31–4.33). His motivating example indeed involved class-level constructs (teacher quality, provided by multiple students' responses to multiple questions), and he defined coefficients to quantify

the degree to which measurements generalize over students, items, and both. Apparently overlooked even among the recent articles proposing higher-level coefficients using GT (Lin et al., 2025; Ten Hove et al., 2022), Brennan's (2001, pp. 130–132) brief example provides insufficient guidance on its own for this relatively common context.

The primary goal of this article was to provide a more thorough introduction to useful reliability indices when measuring cluster-level constructs with multiple scale items, so readers can understand how to address important research goals—namely, to critically evaluate both dimensions of their measurement procedure in order to reveal the most efficient way(s) to minimize measurement error. To fulfil this goal, we will (a) show how the GT framework can elucidate relationships among existing reliability coefficients defined for this or similar contexts (though often given different labels), (b) develop a comprehensive notation to report such reliability coefficients, and (c) provide guidance about how to interpret and utilize such coefficients. The article begins with a motivating example[1] from education science. The principles of GT are introduced, focusing on decomposition of variance components that are then used to define various reliability coefficients. To facilitate interpretation of the coefficients defined using GT, we provide complementary definitions using the 'correlation approach' (Lin et al., 2025; Vangeneugden et al., 2005), which relies on covariance algebra to derive the same coefficients as (un)conditional correlations between repeated measurements of the same target. Using the motivating example data, we demonstrate how LMM and SEM can be used to estimate the variance components and reliability coefficients. We also illustrate how the coefficients can be used to inform how measurements can be made more reliable. We discuss how various previously proposed and commonly reported coefficients fit into the comprehensive GT framework, and we provide advice for researchers interested in validating scales used to measure cluster-level constructs.

1 | MOTIVATING EXAMPLE

Ye et al. (2024) evaluated the measurement of a dimension of school climate: school emphasis on academic success (SEAS). They analysed data available from the Trends in International Mathematics and Science Study (TIMSS; Mullis et al., 2020), provided with their supplementary materials on the associated Open Science Framework (OSF) project (Ye et al., 2022). Their goal was to compare measurements of the same construct from different sources (teachers or school principles) using 11 items that reference different members of the school (students, parents, or teachers). To avoid complications unrelated to this article's goals, we focus only on parental SEAS, measured by teacher responses to four of the 11 items on a 1 (*very high*) to 5 (*very low*) scale. An example item is 'How would you characterize parental expectations for student achievement within your school?' Complete descriptions of all items can be found in Ye et al. (2024, table 1), but our examples focus only on Items 5–8. Ye et al. (2024) evaluated the measurement instrument using multilevel CFA. Figure 1 shows a path diagram depicting such a CFA for the 4-indicator, 1-factor model relevant to the subset of parental SEAS items (as measured by teachers). The parameters depicted in Figure 1 will be introduced with the corresponding structural equations in a later section.

SEAS is a school-level construct, which is measured by perceptions of teachers within each school. Thus, SEAS is a shared construct (Jak et al., 2023), measured using both multiple items and multiple raters (teachers) per school. The construct of interest is therefore $\boldsymbol{\eta}^{(B)}$ in Figure 1, a proxy of which might be calculated to include in subsequent analyses or utilize for decision-making. For example, a school-level

[1]This example is used to facilitate communicating many nuanced concepts in a concrete way, but it is not the only such design to which the concepts are applicable. For example, items could be replaced by raters (Lin et al., 2025; Ten Hove et al., 2022) to measure group-level traits (e.g., independent raters observe students to measure classroom climate), or individual subjects could be the 'cluster' when each subject is repeatedly measured (or observed) using multiple items (or raters). For other such analogous designs, the reliability indices we define can still be applied, but different qualitative labels might be more suitable.

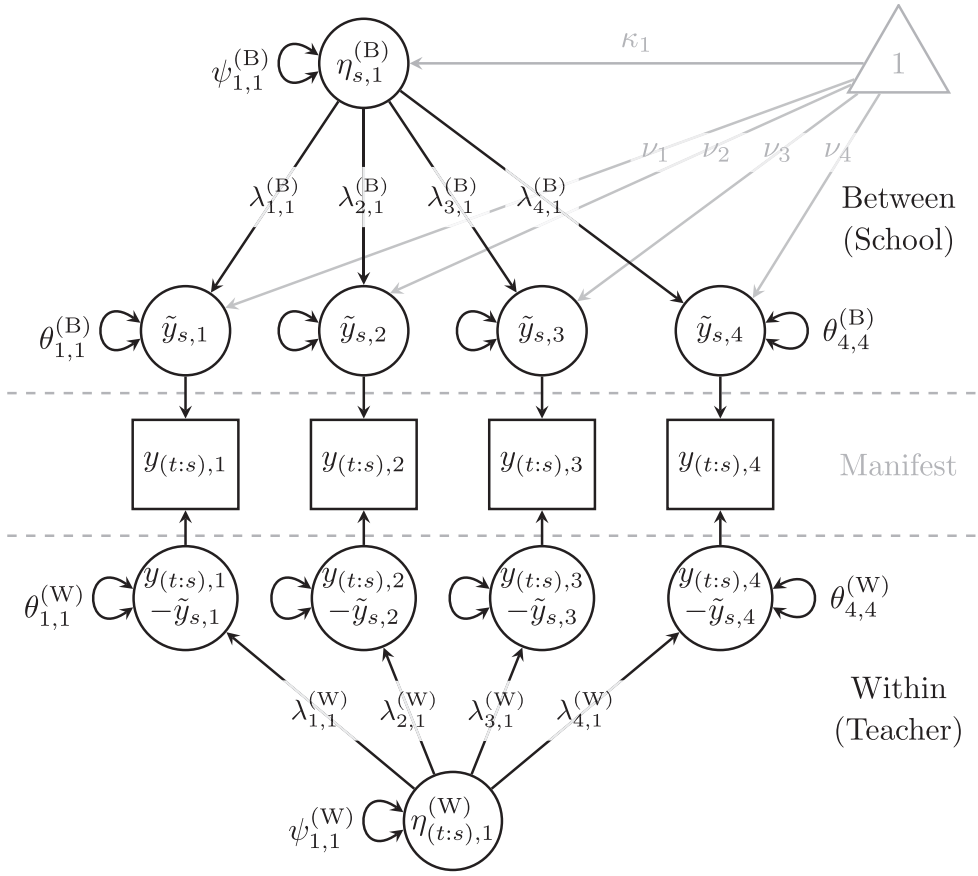


FIGURE 1 Path diagram depicting a multilevel CFA with 4 indicators (items) decomposed into latent school- ($\tilde{y}_s$) and teacher-level components ($y_{t:s} - \tilde{y}_s$), each measuring a single factor at that level.

composite ($\bar{y}_s$) could be calculated to rank-order schools in terms of their parental SEAS or to predict a school-level outcome variable in a regression model:

$$z_s = \beta_0 + \beta_{z|y}\bar{y}_s + \varepsilon_s \tag{1}$$

such as school-average student achievement, as analysed by Ye et al. 2024 using multilevel SEM (ML-SEM). The school-level composite for each School s could be calculated by taking the arithmetic mean $\bar{y}_s$ of responses $y_{(t:s)i}$ to Items $i = 1, \ldots, 4$ by Teachers $t = 1, \ldots, n_{t:s}$ within School s:

$$\bar{y}_s = \frac{1}{4 \times n_{t:s}} \sum_{i=1}^{4} \sum_{t=1}^{n_{t:s}} y_{(t:s)i}, \tag{2}$$

where $n_{t:s}$ is the number of teachers who rated parental SEAS in School s. The reliability ρ_y of this composite has practical importance due to downstream consequences of how it is used to make decisions (e.g., draw inferences). For example, it quantifies the degree to which a simple-regression coefficient (e.g., $\beta_{z|y}$ in Equation 1) would be attenuated due to measurement error, relative to the simple-regression coefficient β_T when regressing z_s on an error-free school-level predictor μ_s (McDonald, 1999, p. 135, equation 7.33):

$$\beta_{z|y} = \beta_T \sqrt{\rho_y} \;\Leftrightarrow\; \beta_T = \frac{\beta_{z|y}}{\sqrt{\rho_y}}. \tag{3}$$

The same form holds for correcting a correlation coefficient ρ_{yz}, which is equivalent to a standardized simple-regression coefficient (McDonald, 1999, pp. 133–134, equations 7.23 and 7.25). The formula can be extended to correct the correlation for attenuation by measurement error also present in z_s by dividing by (the square-root of the product of) both variables' reliabilities (McDonald, 1999, pp. 133–134, equations 7.24 and 7.26); however, correcting multiple-regression coefficients for attenuation becomes more complicated. Note that these principles are also applicable to between-level coefficients of multilevel models (e.g., cluster-randomized trials; S.-J. Cho & Preacher, 2016, pp. 775–778).

The school-level composite in Equation (2) contains multiple sources of measurement error, complicating how to conceive of its reliability from a CTT perspective. Geldhof et al. (2014) proposed resolving this issue by simply applying CTT reliability coefficients (α, ω) to the latent components at each level of analysis in Figure 1. However, as Lai (2021) aptly criticized, Geldhof et al.'s (2014) level-specific reliability coefficients do not reflect the reliability of the composite in Equation (2); instead, Geldhof et al.'s (2014) coefficients reflect hypothetical reliability of a composite that cannot be calculated from the unobserved school-level components:

$$\tilde{y}_s = \sum_{i=1}^{4} \tilde{y}_{s,i}, \tag{4}$$

where $\tilde{y}_{s,i} = \mathbb{E}_{\mathcal{T}}\left(y_{(t:s)i}\right)$ is the unobserved expected value (i.e., population mean) of teacher responses to Item i in School s, and the expectation $\mathbb{E}_{\mathcal{T}}(.)$ is taken over the population of teachers ($\mathcal{T}$, following notation specified in Lin et al., 2025, appendix A). Lai (2021) defined reliability coefficients that validly apply to observed-data composites. Although he derived coefficients from a CTT perspective, we find that many of Lai's (2021) coefficients (which we will discuss in this paper) can be understood quite informatively from a GT perspective, and are analogous to those proposed by Brennan (2001, equations 4.31–4.33) but defined using CFA parameters and estimated under less restrictive assumptions.

2 | GENERALIZABILITY THEORY

Whereas CTT only differentiates between true-score variance and measurement-error variance (Lord & Novick, 1968), GT extends CTT by allowing for variance to be decomposed between multiple sources of systematic and random error that arise from different design 'facets' (Brennan, 2001; Cronbach et al., 1963). The target of the measurement is the facet of *differentiation*, whose 'universe scores' are analogous to true scores in CTT. A given design for obtaining measurements of the target yields a universe of admissible observations that are contaminated with measurement error arising from the facets of *generalization* (e.g., multiple scale items, raters, or occasions). A generalizability coefficient quantifies the degree to which a single observed score—or a composite of n_{facet} measurements of the same target—can be generalized across measurement scenarios, analogous to a reliability coefficient in CTT. Here, we focus only on the design described in our motivating example. For more thorough introductions to GT, we refer to Brennan (2010) and Vispoel et al. (2018), and for more exhaustive coverage of GT we refer to Brennan (2001), Cronbach et al. (1972), and Shavelson and Webb (1991).

For the SEAS construct, the target of measurement is the school, whose SEAS measurements differ due to perspectives of different teachers and how the different scale items are interpreted. Because each teacher only provides measurements of one school, these facets are nested (represented symbolically as $t{:}\,s$). Because every teacher (within every school) responds to the same four items, the item facet is crossed with the other facets, yielding a $(t{:}\,s) \times i$ design. A linear GT model for this design decomposes observed scores into variations around a grand mean μ:

$$y_{(t:s)i} = \mu + \beta_s + \beta_i + \beta_{si} + \beta_{t:s} + \beta_{(t:s)i,\epsilon}, \tag{5}$$

including the main effects of school ($\beta_s = \mu_s - \mu$) and items ($\beta_i = \mu_i - \mu$), their interaction ($\beta_{si} = \mu_{si} - \mu_s - \mu_i + \mu$), and the main effect of teachers ($\beta_{t:s}$, which is confounded with any teacher × school interaction due to nesting). The teacher (within school) × item interaction ($\beta_{(t:s)i,\epsilon}$) is likewise confounded with the three-way interaction; furthermore, this is the highest-order term in the model, so it is also confounded with random measurement error (ϵ), as indicated by the subscript. Every facet's effect has an expected value of zero: $\mathbb{E}(\beta) = 0$.

2.1 | Variance decomposition

Following from the assumption that Equation (5) is a linear combination of independent components with finite variance, the variance of observed scores can be decomposed into independent variance components associated with each component in Equation (5):

$$\text{Var}(y_{(t:s)i}) = \text{Var}(\beta_s) + \text{Var}(\beta_i) + \text{Var}(\beta_{si}) + \text{Var}(\beta_{t:s}) + \text{Var}(\beta_{(t:s)i,\epsilon}), \tag{6}$$

$$\sigma^2_y = \sigma^2_s + \sigma^2_i + \sigma^2_{si} + \sigma^2_{t:s} + \sigma^2_{(t:s)i,\epsilon}. \tag{7}$$

Whereas the subscripts in Equations 5 and 6 indicate the dimension(s) across which each effect varies, the subscripts in Equation (7) merely identify a variance component, which is typically[2] assumed to be constant across units sampled from that dimension.

Although the linear model in Equation (5) does not require facet effects to follow a specific distribution, practical applications often involve assuming they are normally distributed. For ease of presentation, we therefore use the normal-distribution's variance parameter σ^2 (Equation 7) in place of the generic Var(*) operator (Equation 6) when defining coefficients in the following subsections.

2.2 | Relative and absolute error

Observations from a GT design can be used to compare scores between subjects/objects from the facet of differentiation, but such comparisons are complicated by measurement error. GT distinguishes between relative and absolute error. *Relative error* is relevant when making norm-based decisions (e.g., comparing 2 schools' relative standing, regardless of how high or low their actual scores are), whereas *absolute error* is relevant when making criterion-referenced decisions (e.g., comparing a school's score to a specified threshold; Vispoel et al., 2018). McGraw and Wong (1996) referred to indices of *consistency* and *agreement* as being defined using relative and absolute error, respectively, which are also called generalizability and dependability coefficients, respectively (Vispoel et al., 2018).

From the motivating example's $(t:s) \times i$ design, observed scores $y_{(t:s)i}$ and $y_{(t':s')i'}$ can differ due to any effects in Equation (5) that include the s subscript, not only to 'true' differences between schools. The shared SEAS construct's *universe scores* (the GT analogue of true scores in CTT) are represented in Equation (5) as the sum of the grand mean μ with the school effect β_s, which captures school-specific differences that cannot be attributed to differences between teachers' (i.e., raters') perceptions of the school ($\beta_{t:s}$), nor to how items are interpreted differently between individual teachers ($\beta_{(t:s)i,\epsilon}$) or on average between schools (β_{si}). All variance components that include the target's (i.e., school) subscript contribute to the relative error term: $\sigma^2_{si} + \sigma^2_{t:s} + \sigma^2_{(t:s)i,\epsilon}$.

[2]Our sections illustrating LMM and SEM provide more detail about the homoskedasticity assumption, and how SEM allows for heteroskedasticity.

Item-specific factors (β_i) can make School's scores higher or lower, depending on the item(s) used, but they do not affect the rank-ordering of schools. Thus, its variance component (σ_i^2) is omitted from indices of consistency but included in the absolute error term to define indices of agreement (McGraw & Wong, 1996; Vispoel et al., 2018). If teachers rated all schools rather than only their own (i.e., a crossed rather than nested facet), then teacher-specific factors would not affect rank ordering of schools. However, due to the nesting of teachers within schools, the confounded variance component $\sigma_{t:s}^2 = \sigma_t^2 + \sigma_{ts}^2$ means that the absolute-error component σ_t^2 cannot be disentangled from the relative-error component σ_{ts}^2 (Brennan, 2001). This led McGraw and Wong (1996) to drop the consistency–agreement distinction from their notation for indices of nested (one-way model) designs.

2.3 | Random versus fixed facets

Although the number of teachers in a school is finite in practice, the teachers who are selected to respond to the SEAS scale constitute a random sample of teachers from that school, just as the schools constitute a random sample of schools sampled for the TIMSS project (Mullis et al., 2020). One can also argue that these four items are a random sample from a population of possible items to include in the measurement tool, in which case the measurement properties only differ due to sampling error, so reliability coefficients proposed in the next section reflect the consistency (or agreement) expected of scores measured using other teachers and items.

However, the TIMSS project (Mullis et al., 2020) does not randomly sample new items to measure SEAS when it samples new schools and teachers every 4 years. Thus, these $n_i = 4$ items could instead be considered a *fixed* facet, so generalizations are only expected to apply to new samples that use the same items. We adopt the convention of Brennan (2001, p. 59) by using a capital $N_i \ll \infty$ to indicate when we consider our items to be the entire 'population' of items used for measurement. Treating these $N_i = 4$ items as a fixed facet has consequences for how the reliability coefficients defined in the next section for random effects must be adapted (compare Brennan, 2001, figure 4.4 for a fixed facet on p. 123 to figure 4.1 on p. 106 when the same facet is random).

Although Shrout and Fleiss (1979) and McGraw and Wong (1996) discussed ICCs for fixed raters, Ten Hove et al. (2024, p. 973) argued it is difficult to defend in research settings where raters are unlikely to remain constant across studies. Likewise, we do not think it is reasonable to consider lower-level units (here, teachers) as fixed when measuring cluster-level constructs because generalizability is meant to apply to any new (random) sample of the target facet (here, schools). It follows that in such a multistage sampling design, sampling new clusters would imply that in practice, new cluster members must also be sampled within those clusters. Nonetheless, an index of generalizability *only* over scale items implicitly considers the nested facet fixed (Brennan, 2001, section 4.4.4), but we discuss its interpretation after defining that index. We do, however, explicitly consider the case of fixed items in the following section, which is more realistic in the design of our motivating example.

3 | INTRACLASS CORRELATION COEFFICIENTS FOR A SHARED CONSTRUCT

In CTT, the classic formula for reliability expresses the ratio of true-score variance (σ_T^2) to total variance (σ_y^2) of an observable variable y:

$$\rho_{y,y'} = \rho_{y,T}^2 = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} = \frac{\sigma_T^2}{\sigma_y^2}, \tag{8}$$

where σ_E^2 is composed of all measurement-error variance. This can be described as an intraclass correlation coefficient (ICC), which is equivalently interpreted as the squared correlation between true

and observed scores. The commonly used symbol $\rho_{y,y'}$ follows from its interpretation as the correlation between parallel observed scores y and y'. The ICC has long been used as a reliability coefficient in CTT (Bartko, 1966), particularly to quantify IRR (Gwet, 2021; McGraw & Wong, 1996; Shrout & Fleiss, 1979; Ten Hove et al., 2022).

Reliability coefficients in GT are also ICCs, but a greater variety of definitions follow from the distinction between relative and absolute error, decomposition of error into multiple components, and whether to quantify reliability of individual or composite (e.g., averaged) measurements. A nomenclature introduced by Shrout and Fleiss (1979) was improved by McGraw and Wong (1996) to report ICCs in a format that indicates these varieties. However, the conventions were only developed for single-facet designs that capture a single source of measurement error. For example, ICC(1) reflects the reliability of single scores in a nested design, whereas ICC(k) reflects the reliability of composite scores calculated by averaging over k items, raters, etc.—for example, ICC(4) would indicate a composite that averaged over $k = 4$ measurements. For crossed-facet designs, ICC(C,k) reflects consistency of composite scores (or single scores when $k = 1$), whereas ICC(A,k) reflects absolute agreement among composite scores.

The more complex $(t\colon s) \times i$ design of our motivating example necessitates extending the current convention to accommodate more facets. When defining reliability coefficients in the following sections, we will replace k with more immediately identifiable symbols corresponding to the number of measurements *per facet of generalization*. For example, $\text{ICC}_s(\text{C}, n_{t:s} = 20, n_i = 1)$ would represent the consistency (indicated by 'C') of school scores (facet of differentiation indicated by subscript s) calculated by averaging across $n_{t:s} = 20$ teachers per school, each of whom would respond to a single item ($n_i = 1$). Some of the ICCs we define in upcoming sections have more complex subscripts because they apply to joint facets of differentiation, such as teachers within schools $(t\colon s)$ or item responses about schools (si).

3.1 | Generalizability and dependability with random facets

We first define ICCs representing generalizability and dependability, considering both facets as random. We then discuss implications of fixed facets to define a coefficient of IRR[3] conditional on fixed items, followed by an inter-item[4] reliability coefficient conditional on fixed raters, as originally proposed by Brennan (2001, section 4.4.4). To illustrate the connections and distinctions among these coefficients (and others discussed in a later section on Connections with Past Research), Table 1 lists each index (per row) and indicates which variance components from Equation (7) appear in its denominator.

3.1.1 | Generalizability

A generalizability coefficient (G-coef) expresses the target variance proportional to relative error (the parenthetical term in the denominator):

[3]As mentioned in Footnote 1, each of these indices is applicable in other analogous designs where the target is at Level 2 and fully crossed with a facet of generalization, but other qualitative labels might be more suitable. For example, if items are replaced by raters (a $(s\colon c) \times r$ design; Ten Hove et al., 2022; Lin et al., 2025) to measure cluster-level traits, our inter-item reliability coefficients should instead be labelled IRR. In the case of repeatedly measured subjects (who are then the 'clusters'), our IRR should instead be labelled as an index of stability or test–retest reliability (Brennan, 2001, p. 112).

[4]This is more commonly referred to as 'scale reliability', or specifically as 'composite reliability' (coefficient ω), limiting the appropriateness of those terms when quantifying reliability of a single item, as we do here using GT. The term 'internal consistency reliability' is also frequently used, particularly to describe what is quantified by coefficient α, which is also more limited than what GT allows us to quantify (i.e., not only consistency but absolute agreement). Thus, for the remainder this paper, we follow the recommendation of an anonymous reviewer to refer to this as 'inter-item' reliability, analogous to IRR.

TABLE 1 Role of each variance component in ICC definitions for shared constructs.

		Generalizes over …		Denominator includes				
ICC	Equation(s) in paper	Items	Raters	σ^2_i	σ^2_{si}	$\sigma^2_{t:s}$	$\sigma^2_{(t:s)i,\epsilon}$	Other notations
$ICC_s(C, n_{t:s}, n_i)$	9	✓	✓		✓	✓	✓	G-coef, $\mathbb{E}(\rho^2)$, α^b, ω^b
$ICC_s(A, n_{t:s}, n_i)$	11	✓	✓	✓	✓	✓	✓	D-coef, Φ
$ICC_{si}(A, n_{t:s}, N_i)$	14		✓		✓[a]	✓	✓	
$ICC_{t:s}(C, n_{t:s}, n_i)$	16	✓			✓	✓[a]	✓	α^{2l}, ω^{2l} (when $n_{t:s} = 1$)
$ICC_{t:s}(A, n_{t:s}, n_i)$	18	✓		✓	✓	✓[a]	✓	
ICC_O	40, 41, 42		✓			[b]		ICC(1, 1)[c], ICC(1)[d]
$ICC(n_{t:s})$	46		✓			[b]		ICC(1, k)[c], ICC(k)[d]
ICC_L	44, 41		✓			✓		
[e]	45	✓			✓			$\tilde{\alpha}^b$, $\tilde{\omega}^b$

Note: ICC = intraclass correlation coefficient, α implies the fitted model assumes parallel (or at least tau-equivalent) items, ω implies the model allows items to be congeneric. The top section (2 rows) contains coefficients that generalize over both dimensions of error; the middle section (5 rows) contains reliability of an observed (composite of) variable(s) over only one dimension of error; the bottom section (2 rows) contains hypothetical reliability of a latent variable or unobservable composite.

[a]Component also included in numerator.

[b]These are single-item IRR coefficients, so an analogous variance component would be included from a univariate model.

[c]Shrout and Fleiss (1979).

[d] McGraw and Wong (1996), who both used k instead of $n_{t:s}$.

[e]Our paper does not explicitly define indices proposed by Geldhof et al. (2014), which Lai (2021) denoted with $\tilde{\alpha}^b$ and $\tilde{\omega}^b$ to contrast with his α^b and ω^b.

$$ICC_s(C, n_{t:s}, n_i) = \frac{\sigma^2_s}{\sigma^2_s + \left(\frac{\sigma^2_{t:s}}{n_{t:s}} + \frac{\sigma^2_{si}}{n_i} + \frac{\sigma^2_{(t:s)i,\epsilon}}{n_{t:s} \times n_i}\right)}, \tag{9}$$

where the highest-order variance term ($\sigma^2_{(t:s)i,\epsilon}$) is reduced by averaging over $n_{t:s} \times n_i$ measurements. This is equivalent to Brennan (2001, equation 4.31), and its interpretation as a 'generalizability' coefficient follows from it quantifying the degree to which a composite of $n_{t:s}$ teachers' responses to n_i items can be generalized to a composite of $n_{t:s}$ different teachers' responses to n_i different items, each sampled from the same populations of teachers (from the same schools) and of items, respectively. In other words, how correlated could one expect the rank-ordering of schools to be with a composite using $n_{t:s}$ different teachers and n_i different items?

Consistency across averaged scores corresponds to the CTT internal-consistency reliability coefficient α (Cronbach, 1951; Sijtsma, 2009), indicating the expected squared correlation $\mathbb{E}(\rho^2)$ between a school's composite and its true (universe) score. The G-coef is a conditional correlation because average differences between items are partialed out by omitting σ^2_i from Equation (9). By setting $n_{t:s} = n_i = 1$ in Equation (9), we obtain the degree to which a single Teacher t's response to Item i about School s can be generalized to the response of a different teacher (from the same school) to a different item. Thus, $ICC_s(C, n_{t:s} = 1, n_i = 1)$ is the squared (conditional) correlation between a school's single score ($y_{(t:s)i}$) and its universe score, or the correlation between parallel single measurements ($y_{(t:s)i}$ and $y_{(t':s)i'}$). In the notation used by Vangeneugden et al. (2005) and Lin et al. (2025) for the correlation approach:

$$\begin{aligned} ICC_s(C, n_{t:s} = 1, n_i = 1) &= \text{Cor}(y_{(t:s)i}, y_{(t':s)i'} \mid i, i') \\ \text{(only when } \sigma^2_i = 0) &= \text{Cor}(y_{(t:s)i}, y_{(t':s)i'} \mid i, i', t, t'). \end{aligned} \tag{10}$$

It bears repeating that this design does not permit separating relative from absolute teacher error because σ^2_t and σ^2_{ts} are confounded in the nested component's $\sigma^2_{t:s}$. Thus, except when $\sigma^2_t = 0$ (which cannot be verified when this facet is nested), Equation (9) simultaneously quantifies consistency across items and agreement across teachers. Hence, the correlation in Equation (10) can only condition on items in practice, not on teachers (see Lin et al., 2025, equation 19).

3.1.2 | Dependability

Consistency indices are appropriate to quantify the degree to which a correlation or unreliable predictor's simple-regression coefficient is attenuated due to measurement error (Cho & Preacher, 2016). Such coefficients only concern how individuals are rank-ordered and are not affected by centring the variables higher or lower on the number line (as might occur when using more 'difficult' items or more 'strict' raters). When scores are instead compared to an absolute numeric criterion (e.g., when used to classify the targets of measurement), then absolute error enters the ICC definition (compare the first 2 rows of Table 1). A dependability coefficient (D-coef) includes the variability of items on average (σ^2_i), as well as the degree to which a numeric classification criterion[5] ξ differs from the mean μ:

$$\mathrm{ICC}_s(\mathrm{A}, n_{t:s}, n_i) = \frac{\sigma^2_s + (\xi - \mu)^2}{\sigma^2_s + (\xi - \mu)^2 + \left(\frac{\sigma^2_{t:s}}{n_{t:s}} + \frac{\sigma^2_i + \sigma^2_{si}}{n_i} + \frac{\sigma^2_{(t:s)i,\epsilon}}{n_{t:s} \times n_i} \right)}. \tag{11}$$

A global[6] D-coef (often denoted $\mathbf{\Phi}$ in GT literature, e.g., Brennan, 2010, equation 13) is obtained by setting $\xi = \mu$ in Equation (11), which corresponds to a (no longer conditional) correlation between parallel composites, or between parallel measurements when setting $n_{t:s} = n_i = 1$:

$$\mathrm{ICC}_s(\mathrm{A}, n_{t:s} = 1, n_i = 1) = \mathrm{Cor}(y_{(t:s)i}, y_{(t':s)i'}), \tag{12}$$

using notation from the correlation approach (Lin et al., 2025; Vangeneugden et al., 2005).

3.1.3 | Estimating ICCs for future data

Point estimates of Equations (9) and (11) can be obtained by plugging in estimates of the relevant variance components. The sample sizes of each design facet need not correspond to the same design employed to estimate the variance components. Given unbiased estimates of the variance components in Equation (7), the reliability of hypothetical composites can be calculated for other designs[7], to inform an optimal design for obtaining future measurements. In fact, a single sample size might not even apply to the observed data when unbalanced designs are used. In our SEAS example data, the number of teachers ($N_t = 418$) per school ($N_s = 156$) ranged from $n_{t:s} = 1$ (44 schools) to 6

[5]In our SEAS example, a numeric criterion might be used if a school must meet a threshold to qualify for funding—for example, for organising workshops to improve parental involvement in students' academic success. A cut-score-specific D-coef can be reported by indicating the cut-score about which agreement is evaluated—for example, a cut score of $\xi = 4$ could be reported as $\mathrm{ICC}_s(\mathrm{A} = 4, n_{t:s}, n_i)$.

[6]Dependability increases as more extreme cut scores are used for classification. We refrain from further discussion of cut-score D-coefs because manipulating the cut score does not help illustrate how the measurement protocol might be improved. Furthermore, the IRT literature has shown that conditional reliability (Nicewander, 2018) is typically lower for more extreme scores, casting doubt on the gain in dependability by classifying based on more extreme scores.

[7]G-Theory literature (Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991) refers to this as a *decision study* (D-study), often using n' to indicate the hypothetical sample size of a facet, in place of the observed sample size from the generalizability study (G-study) used to estimate the variance components.

teachers (6 schools). In this case, a composite of the observed data involves averaging over a different number of teachers per school, so the reliability of that composite is calculated using the harmonic-mean number of teachers per school $n_{t:s}$ (see also Putka et al., 2008, about ICCs for unbalanced crossed designs):

$$\tilde{n}_{t:s} = \left(\frac{1}{N_s} \sum_{s=1}^{N_s} n_{t:s}^{-1} \right)^{-1}, \tag{13}$$

which is $\tilde{n}_{t:s} = 1.908$ in our sample. Substituting $n = 1$ for either sample size in Equation (9) (or 11) quantifies how generalizable (or dependable) scores are expected to be across individual teachers or in response to a single item.

The advantage of the G-coef is that it quantifies reliability—and its implications for attenuation—simultaneously across all sources of measurement error in the GT design. This paper focusses on a specific scenario in which a cluster-level shared construct is measured using multiple-item scales responded to by members of each cluster. It can also be informative to quantify both inter-item reliability and IRR indices, each of which indicate the degree to which a measurement tool or protocol could be improved to yield better measurements. If improving the reliability of individual measurements is infeasible, reliability of a composite could also be improved by designing a future study that uses more raters or items. Given the relative costs of each possibility, an optimal design might be sought to yield a minimum acceptable reliability at the lowest possible cost.

3.2 | Generalizability and dependability with fixed facets

When only a subset of multiple dimensions of error (i.e., facets of generalization) are to be generalized over, this can be seen as synonymous with considering the complementary subset of facets of generalization to be fixed (Brennan, 2001, section 4.4.4). We begin with IRR given fixed items, which allows us to discuss the implications of treating a facet as fixed. We then discuss the difficulties of considering nested raters as fixed, but nonetheless useful to define an index of inter-item reliability to evaluate that dimension of the measurement design.

3.2.1 | Interrater reliability

The utility of quantifying IRR even in the presence of item-specific error is that one can diagnose the degree to which unique rater perspectives contaminate item or composite scores. This allows researchers to consider the potential impact of reducing heterogeneity in rater perceptions, independent of considerations about inter-item reliability. In our SEAS example, an IRR coefficient would be intended to quantify how measurements generalize across different teachers who respond to the scale, but not how scale responses would generalize to different items. Thus, item-related error variance must be included in the numerator of the ICC, reflecting it as part of the expected universe-score variance. Our ICC's subscript is therefore *si* to denote it as generalizability of school scores using these particular items. Given that teacher effects cannot be disentangled from the teacher × school interaction to distinguish relative from absolute error, we define this IRR index as an agreement index, since it certainly always reflects absolute agreement among raters (Brennan, 2001, section 4.4.4 also noted the equivalence of G- and D-coefs for this index). Only under the restrictive assumption that $\sigma_t^2 = 0$ (which is empirically unverifiable and theoretically different to defend) would this IRR coefficient reflect consistency, but in that case, it could not be distinguished from absolute agreement anyway.

According to Brennan (2001, compare figure 4.4 for a fixed facet on p. 123 to figure 4.1 on p. 106 when the same facet is random), a fixed-facet's main effect is removed from the total variance, whereas its interaction with the target effect becomes part of the universe-score variance used in the numerator. Thus, treating our $N_i = 4$ items as fixed has consequences for how the D-coef in Equation (11) must be adapted. First, the denominator must *exclude* the variance component for fixed items (σ^2_i) because the same item effects will be present when different schools (and teachers within them) are sampled. Second, the numerator must now *include* the school × item interaction component (σ^2_{si}) as part of true-score variance:

$$\mathrm{ICC}_{si}(\mathrm{A}, n_{t:s}, N_i) = \frac{\sigma^2_s + \frac{\sigma^2_{si}}{N_i}}{\sigma^2_s + \frac{\sigma^2_{si}}{N_i} + \left(\frac{\sigma^2_{t:s}}{n_{t:s}} + \frac{\sigma^2_{(t:s)i,e}}{n_{t:s} \times N_i}\right)}. \tag{14}$$

This is a G-coef that is equivalent to a D-coef (Brennan, 2001, p. 130, equation 4.33) due to the nesting of raters (teachers) under the facet of differentiation (schools).

One may quantify the degree to which a single teacher's composite of item responses generalizes to another teacher's by setting $n_{t:s} = 1$. Following the notation of Lin et al. (2025, appendix E), the correlation among composites of all fixed items i can be defined as:

$$\mathrm{ICC}_{si}(\mathrm{A}, n_{t:s} = 1, N_i) = \mathrm{Cor}(y_{(t:s)+}, y_{(t':s)+} \mid \forall_i), \tag{15}$$

where the subscript '+' indicates taking a sum or average across all items, and conditioning on all of them (represented by the 'for all' symbol: ∀). The interpretations differ between the interrater agreement coefficient in Equation 14 and the D-coef in Equation (11), as reflected by comparing the rows of Table 1 corresponding to those equations. Equation (11) does not condition on items—thus treating items as random—so they reflect the degree to which a composite of $n_{t:s}$ teachers' (sampled from the same school s) responses to *any randomly sampled set* of n_i items can be generalized to a composite of $n_{t:s}$ *different* teachers' responses to *another set* of n_i randomly sampled items. In other words, how correlated could one expect schools' scores to be, when scores are a composite of $n_{t:s}$ different teacher responses to different items?

In contrast, the IRR coefficient in Equation (14) conditions on items, thus treating them as fixed. So it reflects the degree to which a composite of $n_{t:s}$ teachers' responses to *these* $N_i = 4$ items can be generalized to a composite of $n_{t:s}$ different teachers' responses to *the same* $N_i = 4$ items, when sampled from the same school s.

If the teacher main effect were not confounded due to nesting, then its interpretation could be stated as how correlated one could expect the rank-ordering of schools to be (i.e., consistency) with a composite using $n_{t:s}$ different teachers but the same $N_i = 4$ items. However, although fixed items do not contribute to absolute error, average differences between teachers would affect rank-ordering of schools, due to teachers being nested within schools.

3.2.2 | Inter-item reliability

An index of inter-item reliability can be defined similar to the G-coef in Equation (9). However, a inter-item reliability coefficient would not be intended to quantify how measurements generalize across different teachers who respond to the scale. Thus, teacher-related relative[8] error variance must be included also in the numerator of the ICC:

[8]Only variance that reflects between-school differences (i.e., σ^2_{ts}) should be included in the numerator and denominator of a consistency index. Unfortunately, the nesting of teachers within schools does not allow disentangling pure teacher effects (σ^2_t) from the confounded variance component $\sigma^2_{t:s} = \sigma^2_t + \sigma^2_{ts}$, so any $\sigma^2_t > 0$ cannot be excluded from the numerator or denominator.

$$\text{ICC}_{t:s}(\text{C}, n_{t:s}, n_i) = \frac{\sigma_s^2 + \frac{\sigma_{t:s}^2}{n_{t:s}}}{\sigma_s^2 + \frac{\sigma_{t:s}^2}{n_{t:s}} + \left(\frac{\sigma_{si}^2}{n_i} + \frac{\sigma_{(t:s)i,\epsilon}^2}{n_{t:s} \times n_i}\right)}, \tag{16}$$

where the ICC subscript $t{:}s$ indicates this coefficient's target of measurement is teachers within schools. Despite the target of measurement actually being clusters for a shared construct, the index in Equation (16) is only meant to quantify generalizability over items (i.e., inter-item reliability, or internal consistency), not cluster members, as reflected in those rows of Table 1 (see also an equivalent definition in Brennan, 2001, equation 4.32). One can interpret this as the expected correlation between two measurements of the same randomly sampled school given by the same randomly sampled teacher within that school, but using two different items (Lin et al., 2025; Vangeneugden et al., 2005):

$$\begin{aligned} \text{ICC}_{t:s}(\text{C}, n_{t:s} = 1, n_i = 1) &= \text{Cor}(y_{(t:s)i}, y_{(t:s)i'} \mid t, i, i') \\ \text{(due to nesting)} &= \text{Cor}(y_{(t:s)i}, y_{(t:s)i'} \mid i, i'). \end{aligned} \tag{17}$$

Note that despite this coefficient implicitly treating teachers as fixed, we refrain from using a capital $N_{t:s}$ because we do not take seriously the assumption that teachers are actually fixed when gathering future samples. The practical purpose of this coefficient is merely to evaluate the degree to which measurements are contaminated specifically by item-specific error.

Whereas the G-coef reflects how reliably measurements can distinguish between clusters (schools)—jointly across cluster members (teachers) and items—the ICC in Equation (16) narrows this evaluation of reliability on a particular dimension of measurement error (items) so that researchers can consider how to adjust their measurement procedure or instrument in order to minimize that specific source of error.

One may quantify the degree to which individual responses (or a composite of $n_{t:s} \geq 2$ teachers' responses) to a single scale item generalize to another item by setting $n_i = 1$ in Equation (16). Likewise, one may quantify the degree to which a single teacher's item response (or scale composite) generalizes to another (composite of) item(s) by setting $n_{t:s} = 1$. One may also quantify absolute agreement between (composites of) items (i.e., not conditioning the correlation in Equation (17) on particular items i and i') by including the item-effects variance component (σ_i^2) in the denominator, as it appears in Equation (11):

$$\begin{aligned} \text{ICC}_{t:s}(\text{A}, n_{t:s}, n_i) &= \frac{\sigma_s^2 + \frac{\sigma_{t:s}^2}{n_{t:s}}}{\sigma_s^2 + \frac{\sigma_{t:s}^2}{n_{t:s}} + \left(\frac{\sigma_i^2 + \sigma_{si}^2}{n_i} + \frac{\sigma_{(t:s)i,\epsilon}^2}{n_{t:s} \times n_i}\right)} \\ &= \text{Cor}(y_{(t:s)i}, y_{(t:s)i'}). \end{aligned} \tag{18}$$

The utility of quantifying inter-item reliability even in the presence of rater error is that one can diagnose the degree to which item-specific properties contaminate individual or composite scores. This allows researchers to consider the potential impact of reducing measurement error in their scales, independent of considerations about rater error.

4 | ILLUSTRATION USING A LMM

Having defined reliability indices in the previous section, we use this section to report the results of an illustrative analysis in R (version 4.6.1; R Core Team, 2026), which readers can reproduce using

the data and R-syntax files available[9] from our OSF project. We analysed the parental SEAS data (responses to $N_i = 4$ items by $N_t = 418$ teachers nested in $N_s = 156$ schools) using the mixed-modelling package `lme4` (version 1.1-38; Bates et al., 2015), although the same methods can be applied using other LMM software or to any shared construct(s) measured by multiple-item scales, which are completed by individual members responding about the group to which they belong. More extensive treatments of LMM are available in various textbooks (e.g., Molenberghs & Verbeke, 1997; Snijders & Bosker, 2012; Hox et al., 2018; Jiang, 2007), and briefer papers demonstrate estimating variance components for G-coefs with `lme4` (Jiang, 2018; Putka et al., 2008; Jorgensen, 2021; Ten Hove et al., 2025).

We interpret point (and interval) estimates because that is the focus of this article, without testing hypotheses about meeting minimal criteria (e.g., Bonett, 2002; Bonett & Wright, 2015; McGraw & Wong, 1996). We used restricted maximum likelihood (REML) estimation because the LMM contains only a fixed intercept and random intercepts associated with components in Equation (5), and variance components of the latter (Equation 7) are our priority for unbiased estimation. Whereas full ML estimation can produce (often downward) biased estimates of variance components by not accounting for the *df* lost when estimating fixed effects, REML provides more accurate, unbiased variance-component estimates by 'partialing out' the fixed intercept in Equation (5) (Snijders & Bosker, 2012).

We report 90% CIs for all ICCs, obtained using a resampling method often referred to as a Monte Carlo CI (Preacher & Selig, 2012), implemented in the `semTools` package (version 0.5-9; Jorgensen et al., 2026). Ten Hove et al. (2025) demonstrated via simulations that Monte Carlo CIs often yield nominal coverage rates when applied to LMMs (see also Jiang et al., 2022, who investigated other resampling methods). A 90% confidence level was chosen for two-sided CIs, which corresponds to 95% confidence using only the lower bound (e.g., for interested readers to test a H_0 that reliability exceeds a minimal criterion), but we report both lower and upper limits for completeness.

4.1 | Practical application

The LMM of interest is represented in Equation (5), which contains a single fixed parameter (the intercept μ), four random intercepts (β_s, β_i, β_{si}, and $\beta_{t:s}$), and the residual term ($\beta_{(t:s)i,\epsilon}$). The distributional assumptions are that the random intercepts and residuals are independent and normally distributed, with the variance components from Equation (7):

$$\beta_s \sim \mathcal{N}(0, \sigma_s^2), \tag{19}$$

$$\beta_i \sim \mathcal{N}(0, \sigma_i^2), \tag{20}$$

$$\beta_{si} \sim \mathcal{N}(0, \sigma_{si}^2), \tag{21}$$

$$\beta_{t:s} \sim \mathcal{N}(0, \sigma_{t:s}^2), \tag{22}$$

$$\beta_{(t:s)i,\epsilon} \sim \mathcal{N}(0, \sigma_{(t:s)i,\epsilon}^2). \tag{23}$$

[9] https://osf.io/vg5uq/.

TABLE 2 Maximum likelihood point estimates of variance components for SEAS data using LMM and SEM.

	LMM	SEM_{Par}	SEM_{Con}
$\hat{\sigma}^2_s$	.151	.149	.127
$\hat{\sigma}^2_i$	.107	.108	.108
$\hat{\sigma}^2_{si}$	.045	.044	.046
$\hat{\sigma}^2_{t:s}$	.285	.285	.302
$\hat{\sigma}^2_{(t:s)i,\epsilon}$	.295	.295	.282

Note: LMM = linear mixed model, using `lme4::lmer()` to obtain restricted ML estimates. SEM_{Par} = structural equation model with essentially parallel constraints, using `lavaan::cfa()` to obtain ML estimates. SEM_{Con} = structural equation model for congeneric items, calculated from ML estimates using weights $\mathbf{w} = \frac{1}{4}\mathbf{1}$ to average over $n_i = 4$ items.

We used the R package `lme4` (version 1.1.38; Bates et al., 2015) to fit this LMM to the SEAS data provided by Ye et al. (2022). Along with SEM results presented in the next section, the resulting variance-component estimates are reported in Table 2. Point estimates were plugged into Equations (9, 11, 14), and (16) to obtain reliability coefficients reported in Table 3.

Comparing variance components in Table 2, one can see that there is more school-level variance ($\sigma^2_s = 0.151$) than item variance ($\sigma^2_i = 0.107$) or their interaction ($\sigma_{si} = 0.045$). This is encouraging because the shared SEAS construct's variance should primarily reflect differences between schools, not differences in how schools differentially respond to items. However, teacher-level variances are overwhelmingly more influential than school-level variance, indicating that composites would be quite saturated with measurement error, which seriously undermines comparing schools based on composites of teacher ratings. Indeed, the reliability coefficients in Table 3 can illuminate what conditions would be necessary for such a composite to yield reliable (i.e., generalizable or dependable) data.

Each of the first four rows of Table 3 is a G-coef (Equation 9) indicating the expected consistency of single or composite scores. The first row of Table 3 indicates that a single teacher's response to a single item of the SEAS parental subscale would not be very generalizable, $\text{ICC}_s(\text{C}, n_{t:s} = 1, n_i = 1) = .194$, 90% CI [.122, .260]—that is, we could not expect the rank-ordering of schools to be consistent using single-teacher, single-item responses. Composite scores are more consistent, by averaging over $n_i = 4$ items, $\text{ICC}_s(\text{C}, n_{t:s} = 1, n_i = 4) = .289$, 90% CI [.186, .380], or over $\tilde{n}_{t:s} = 1.908$ teachers, $\text{ICC}_s(\text{C}, n_{t:s} = 1.908, n_i = 1) = .301$, 90% CI [.199, .387], but even averaging over both items and teachers, $\text{ICC}_s(\text{C}, n_{t:s} = 1.908, n_i = 4) = .430$, 90% CI [.298, .532], leaves over half of the composite's variance contaminated by measurement error. Because item-specific error ($\sigma^2_i = 0.107$ in Table 2) is not negligibly smaller than other variance components, absolute agreement (Equation 11) is noticeably lower than consistency, $\text{ICC}_s(\text{A}, n_{t:s} = 1.908, n_i = 4) = .400$, 90% CI [.271, .508], indicating that composite scores calculated from this number of items and teachers per school (and using the mean as a cut score) would not be a reliable (or 'dependable') way to classify schools.

One can see in Table 2 that teacher-related error is the overwhelming majority of variance. Equations 14 and 16 can be used to quantify the (un)reliability of scores due to specific error components. Inter-item reliability, $\text{ICC}_{t:s}(\text{C}, n_{t:s} = \tilde{n}_{t:s}, n_i = 4) = .857$, 90% CI [.826, .881], is quite high—even the lower 90% confidence limit exceeds .80—whereas IRR is quite poor, $\text{ICC}_{si}(\text{A}, n_{t:s} = \tilde{n}_{t:s}, N_i = 4) = .462$, 90% CI [.336, .560], indicating that teacher error is the greater threat to validity of any school-level inferences drawn from an SEAS composite.

It is well known from CTT that reliability can be improved by adding more scale items, but GT can give researchers a good intuition about why reliability can be improved with larger samples in any facet of generalization. Increasing the number of items (n_i) or teachers per school ($n_{t:s}$) in Equation (9) reduces the size of those variance components contributing to a composite. The same variance-component

TABLE 3 Point and interval estimates of reliability coefficients for SEAS data using LMM and SEM.

	LMM[a]		SEM_{Par}[b]		SEM_{Con}[b,c]	
Coefficient	$\widehat{\rho}^2$	**90% CI**[d]	$\widehat{\rho}^2$	**90% CI**	$\widehat{\rho}^2$	**90% CI**
$ICC_s(C, n_{t:s} = 1, n_i = 1)$	.194	[.122, .260]	.193	[.122, .258]	—	—
$ICC_s(C, n_{t:s} = 1, n_i = 4)$	.289	[.186, .380]	.287	[.187, .376]	.249	[.158, .340]
$ICC_s(C, n_{t:s} = \tilde{n}_{t:s}, n_i = 1)$	.301	[.199, .387]	.300	[.200, .383]	—	—
$ICC_s(C, n_{t:s} = \tilde{n}_{t:s}, n_i = 4)$	.430	[.298, .532]	.428	[.299, .528]	.381	[.259, .489]
$ICC_s(A, n_{t:s} = \tilde{n}_{t:s}, n_i = 4)$	.400	[.271, .508]	.397	[.274, .494]	.352	[.237, .455]
$ICC_{si}(A, n_{t:s} = \tilde{n}_{t:s}, N_i = 4)$[e]	.462	[.336, .560]	.459	[.336, .556]	.415	[.298, .519]
$ICC_{t:s}(C, n_{t:s} = \tilde{n}_{t:s}, n_i = 4)$[f]	.857	[.826, .881]	.857	[.827, .880]	.855	[.825, .880]
$ICC_s(C, n_{t:s} = 7, n_i = 4)$	.707	[.576, .784]	.705	[.578, .782]	.663	[.528, .753]
$ICC_s(C, n_{t:s} = 14, n_i = 4)$	.803	[.698, .861]	.803	[.699, .860]	.770	[.654, .840]

Note: LMM = linear mixed model, SEM_{Par} = structural equation model with essentially parallel constraints, SEM_{Con} = structural equation model for congeneric items, CI = Monte Carlo confidence interval estimated using `semTools::monteCarloCI()`, ICC = intraclass correlation coefficient (generically symbolized as $E(\widehat{\rho}^2)$ for consistency or $\boldsymbol{\Phi}$ for agreement in GT literature), $\tilde{n}_{t:s} = 1.908$ is the harmonic-mean number of teachers per class in the SEAS data.

[a]Restricted MLE using `lme4::lmer()`.

[b] MLE using `lavaan::cfa()`.

[c]Congeneric items are assumed fixed in new samples; thus, ICCs should be reported using a capital $N_i = 4$ (or $N_i^C = 4$ to additionally indicate items are congeneric).

[d] Robust asymptotic covariance matrix estimated using `merDeriv::vcov.lmerMod()` (Wang & Merkle, 2018).

[e] Interrater agreement (G- or D-coef given fixed items).

[f] Inter-item reliability (G-coef given fixed raters).

estimates in Table 2 can be plugged into Equation (9) (or 11) with hypothetical sample sizes, to estimate the expected reliability of a composite calculated from more teachers per school. The bottom two rows of Table 3 hold the number of items constant, but increasing $n_{t:s}$ shows that a minimally acceptable reliability of .70 could be achieved with $n_{t:s} = 7$ teachers per school. However, using a stricter criterion for minimally acceptable reliability, $n_{t:s} = 14$ yields $ICC_s(C, n_{t:s} = 14, n_i = 4) = .803$, 90% CI [.698, .861]. Even larger samples would be required for a more conservative approach of using (lower) confidence limits to test whether the population reliability meets a minimum standard (e.g., Bonett, 2002; Bonett & Wright, 2015; McGraw & Wong, 1996), as shown in our OSF project's R syntax.

5 | ILLUSTRATION USING SEM

We now present a multilevel CFA, which can be specified analogous[10] to the LMM in Equation (5). We briefly introduce the multilevel factor model below, using the traditional LISREL notation (Bollen, 1989). We continue using the subscripts associated with our motivating example, but the model description below generalizes to any sort of observations nested within clusters. After briefly discussing similarity of CFA and LMM results, we distinguish conceptual issues about treating items as a fixed or random facet, with consequences for homogeneity assumptions implicit to LMM. We present analogous formulas for the variance components in Equation (7) as functions of (potentially less constrained) CFA parameters, yielding more general definitions of factor-model-based reliability coefficients analogous to

[10]Although the parameters have equivalent interpretations, estimates can differ negligibly between LMM and CFA due to maximizing different likelihoods (Jorgensen, 2021; Vispoel et al., 2023): the former univariate, the latter multivariate (items stored in wide format).

those in Equations 9, 11, 14, and 16, which allows reliability to be estimated under less-restrictive measurement conditions than LMM (Vispoel et al., 2021).

5.1 | Multilevel common factor model

The responses of each teacher in each school to the n_i scale items is stored in a vector $\mathbf{y}_{t:s}$, which is decomposed into latent components at the school level (i.e., school means; Equation 4) and teacher level (i.e., teacher deviations from their school's mean):

$$\mathbf{y}_{t:s} = \tilde{\mathbf{y}}_s + (\mathbf{y}_{t:s} - \tilde{\mathbf{y}}_s), \tag{24}$$

where the parenthetical term is the within-school (teacher) component. Each component vector is modelled as a function of $f \geq 1$ common factors $\boldsymbol{\eta}$ at that level of analysis:

$$\mathbf{y}_{t:s} - \tilde{\mathbf{y}}_s = \boldsymbol{\Lambda}^{(\mathrm{W})}\boldsymbol{\eta}_{t:s}^{(\mathrm{W})} + \boldsymbol{\varepsilon}_{t:s}^{(\mathrm{W})}, \tag{25}$$

$$\tilde{\mathbf{y}}_s = \boldsymbol{\nu} + \boldsymbol{\Lambda}^{(\mathrm{B})}\boldsymbol{\eta}_{s}^{(\mathrm{B})} + \boldsymbol{\varepsilon}_{s}^{(\mathrm{B})}, \tag{26}$$

where $\boldsymbol{\Lambda}$ is a $n_i \times f$ matrix of factor loadings and $\boldsymbol{\varepsilon}$ is a vector of residuals (a 'unique factor' per item). Because lower-level components are mean centered, intercepts $\boldsymbol{\nu}$ are only modelled at the highest level of analysis. The common factors are assumed normally distributed at each level of analysis:

$$\boldsymbol{\eta}_{t:s}^{(\mathrm{W})} \sim \mathcal{MVN}(0, \boldsymbol{\Psi}^{(\mathrm{W})}), \tag{27}$$

$$\boldsymbol{\eta}_{s}^{(\mathrm{B})} \sim \mathcal{MVN}(\boldsymbol{\kappa}, \boldsymbol{\Psi}^{(\mathrm{B})}), \tag{28}$$

where $\boldsymbol{\Psi}$ is a $f \times f$ covariance matrix and $\boldsymbol{\kappa}$ is a vector of factor means. Likewise, unique factors are assumed normally distributed at each level of analysis:

$$\boldsymbol{\varepsilon}_{t:s}^{(\mathrm{W})} \sim \mathcal{MVN}(0, \boldsymbol{\Theta}^{(\mathrm{W})}), \tag{29}$$

$$\boldsymbol{\varepsilon}_{s}^{(\mathrm{B})} \sim \mathcal{MVN}(0, \boldsymbol{\Theta}^{(\mathrm{B})}), \tag{30}$$

where $\boldsymbol{\Theta}$ is a (typically diagonal) $n_i \times n_i$ covariance matrix. The model-implied mean structure of observed variables is:

$$\mathbb{E}(\mathbf{y}_{t:s}) = \boldsymbol{\nu} + \boldsymbol{\Lambda}^{(\mathrm{B})}\boldsymbol{\kappa}, \tag{31}$$

while the model-implied covariance structure at each level is:

$$\mathrm{Var}(\mathbf{y}_{t:s} - \tilde{\mathbf{y}}_s) = \boldsymbol{\Lambda}^{(\mathrm{W})}\boldsymbol{\Psi}^{(\mathrm{W})}\boldsymbol{\Lambda}^{(\mathrm{W})} + \boldsymbol{\Theta}^{(\mathrm{W})}, \tag{32}$$

$$\mathrm{Var}(\tilde{\mathbf{y}}_s) = \boldsymbol{\Lambda}^{(\mathrm{B})}\boldsymbol{\Psi}^{(\mathrm{B})}\boldsymbol{\Lambda}^{(\mathrm{B})} + \boldsymbol{\Theta}^{(\mathrm{B})}. \tag{33}$$

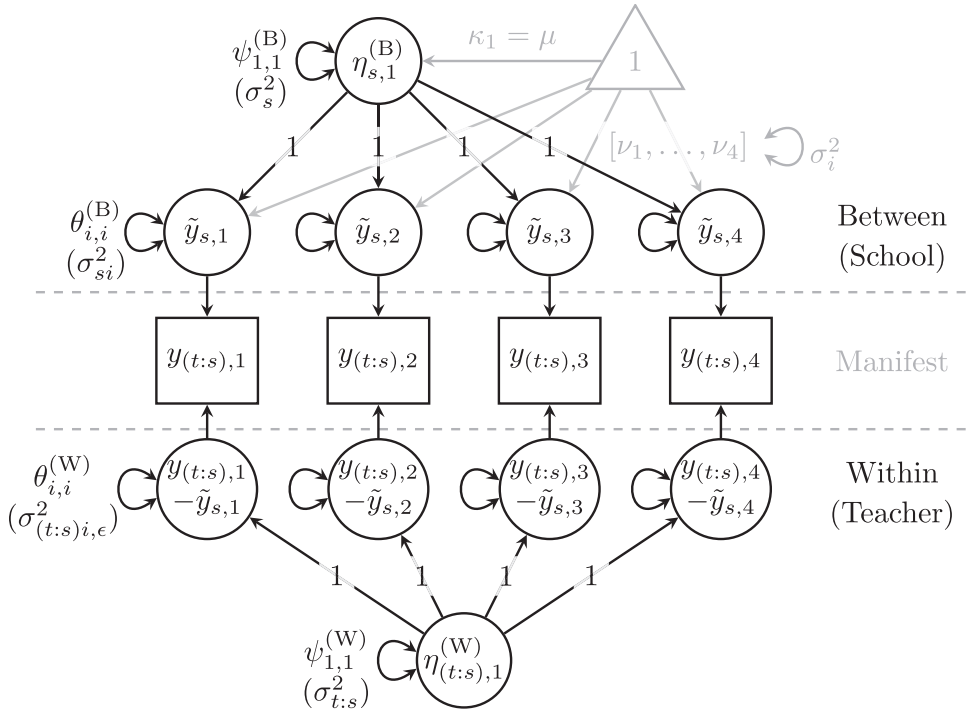


FIGURE 2 Path diagram depicting a multilevel CFA with the assumption of 4 indicators being essentially parallel items (i.e., equal factor loadings and residual variances across item components) at each level of analysis. Each CFA parameter is labelled using traditional LISREL notation (Bollen, 1989), with analogous GT parameters depicted in parentheses. The unique-factor variances are constrained to equality, the item intercepts are constrained to average zero to identify the factor mean (grand mean $\boldsymbol{\mu}$ from GT), and the variance of intercepts represents the variance component $\boldsymbol{\sigma}^2_i$ (main effect of items).

When fitting the congeneric CFA in Figure 1 to the SEAS data, we constrain factor loadings to equality across levels ($\boldsymbol{\Lambda}^{(W)} = \boldsymbol{\Lambda}^{(B)} = \boldsymbol{\Lambda}$), so the school- and teacher-level factors remain interpretable as the GT components indicated in Figure 2 (Jak et al., 2013, 2023; Jak & Jorgensen, 2017). But the formulas we present in the following sections do not strictly require cross-level[11] metric invariance.

The same variance components from Equation (7) can be estimated as variances of common and unique factors in Figure 1, although the absolute-error component $\boldsymbol{\sigma}^2_i$ would not be directly estimable. When all factor loadings = 1 (see Figure 2), $\boldsymbol{\sigma}^2_i$ can be represented as a function of estimated item intercepts:

$$\hat{\sigma}^2_i = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} \hat{\nu}^2_i, \tag{34}$$

as depicted by the mean structure in Figure 2. The factor mean is estimable by constraining the sum of intercepts to be zero (Jorgensen, 2021), yielding intercepts that are interpretable as item main effects ($\boldsymbol{\beta}_j$, i.e.,

[11]Upon rejecting a H_0 of full metric invariance, the interpretability of factors could be preserved by establishing partial invariance—for example, by testing individual indicators for differential item functioning (DIF). However, interpreting the meaning of different loadings across levels remains difficult. Violations of metric invariance across clusters actually implies that Level-1 loadings should be specified as random parameters that vary across Level-2 units, which would imply cluster-specific reliabilities. This methodology needs further exploration and lies beyond the scope of this paper.

item-mean deviations around the grand mean $\kappa = \mu$). Figures 1 and 2 are equivalent when all factor loadings are fixed to 1 and residual variances (at each level of analysis) are constrained to equality across items. The constraints in Figure 2 represent *essentially*[12] *parallel* indicators, reflecting an assumption that items are randomly sampled from a theoretically infinite population of items measuring the same construct (Vispoel et al., 2021).

We used ML estimation in the `lavaan` package (version 0.7-2; Rosseel, 2012; Rosseel et al., 2026) to fit the CFA in Figure 2 to the SEAS data provided by Ye et al. (2022). Table 2 shows how similar the variance-component estimates are between the two modelling frameworks (Jorgensen, 2021; Vispoel et al., 2023), with differences $\leq .002$ in absolute value across components. Likewise, the reliability coefficients in Table 3 differ $\leq .003$ in absolute value between those obtained from LMM and SEM estimates. All interpretations and applications are consistent with the LMM reported in the previous section, given the fitted SEM reflects the same essentially parallel assumptions made by LMM. The rest of this section discusses how SEM can be advantageous by relaxing the essentially parallel assumption.

5.2 | Congeneric items

The multilevel CFA in Figure 2 incorporates essentially parallel assumptions: equality of loadings and residual variances across items. Relaxing the equality constraints on residual variances yields the less strict assumption of essential[13] *tau-equivalence*, which only involves equality of factor loadings. The most general CFA in Figure 1 does not impose any equality constraints, conforming to the least restrictive measurement model for *congeneric* items.

The essentially parallel assumption of the previous LMM and CFA applications is easily defensible (and arguably necessary) for the teacher facet, given that teachers constitute a random sample of lower-level units within the randomly sampled target of measurement (schools). However, items could be considered fixed when the same items are expected to be used for measurements provided in new samples, as we discussed when adapting the G-coef (Equation 9) to assume fixed items (Equation 14). We view a congeneric model to be conceptually consistent with estimating variance components of fixed facets because fixed-item parameters are estimated separately rather than assuming item-related effects are randomly sampled from a common distribution.

Similar to past work (e.g., Brennan, 2001, tables 4.1 & 4.2; McGraw & Wong, 1996, tables 4 & 5), our ICC formulas reflect the same essentially parallel assumption associated with randomly sampled facet levels. Whereas σ_i^2 is obtained indirectly as a function of CFA parameters (item means; see Equation 34) regardless of whether items are treated as random or fixed, the remaining components that capture item-related variance (see distributional assumptions in Equations 21 & 23) would still be estimated by LMM (as well as the CFA in Figure 2) assuming normality and homoskedasticity across levels of the fixed facet. The congeneric CFA in Figure 1 instead relaxes equality assumptions for components β_{si} and $\beta_{(t:s)i,\epsilon}$ across items. Because a CFA with essentially parallel constraints is nested under a congeneric CFA, the essentially parallel assumption can be tested[14] as a H_0 using a likelihood ratio test.

Calculating reliability coefficients using congeneric CFA parameters requires defining the variance components in Equation (6) in terms of CFA parameters. We provide these definitions in the next

[12]CFA models for continuous indicators include three types of measurement parameters: intercepts, factor loadings, and residual variances. Parallel indicators are assumed to be homogeneous with respect to all three types of parameter, whereas essentially parallel indicators only assume equal factor loadings and residual variances. The intercepts in Figure 2 can vary across items, reflecting the essentially parallel assumption.

[13]Similar to the parallel assumption described in Footnote 12, full versus essential tau-equivalence can be distinguished by whether intercepts are also assumed equal across items.

[14]Our supplementary R syntax on OSF shows that our data provide ample evidence to reject this H_0, $\chi^2(9) = 99.071, p < .0001$.

section. Furthermore, using a fixed set of items implies only generalizing to situations in which a composite is calculated from those items, in which case the IRR with $N_i = 4$ in Table 3 should be interpreted. Although we view a congeneric measurement model to be more consistent with treating items as fixed (in which case, no coefficients should include σ^2_i), we nevertheless include the G- and D-coefs from Equations 9 and 11 in Table 3. Similar to estimating inter-item reliability as though nested raters could be fixed, a researcher may also want to consider items as random even when relaxing parallel or tau-equivalence assumptions.

5.3 | Variance components implied by CFA parameters

The variance component for schools in Figure 2 is the school-level factor variance ($\sigma^2_s = \psi^{(B)}_{1,1}$). This is actually a simplification of the more general formula for congeneric items, in which the school-level variance contribution to a composite is a function of both the factor covariance matrix ($\boldsymbol{\Psi}^{(B)}$) and factor loadings ($\boldsymbol{\Lambda}^{(B)}$) depicted in Figure 1:

$$\sigma^2_s = \mathbf{w}'\boldsymbol{\Lambda}^{(B)}\boldsymbol{\Psi}^{(B)}\boldsymbol{\Lambda}^{(B)}\mathbf{w}, \tag{35}$$

where $\mathbf{w}$ is a vector of n_i weights applied to the items before summing. Optimal weights can be chosen to maximize the composite's reliability (Li, 1997), but most frequently the (implied) selected weights are a vector of ones ($\mathbf{1}$, yielding a scale sum) or the inverse of the number of summed items ($n_i^{-1}\mathbf{1}$, yielding a scale mean).

Other variance components constitute measurement error. The variance component for teachers within schools in Figure 2 is the teacher-level factor variance ($\sigma^2_{t:s} = \psi^{(W)}_{1,1}$). Again, this is a reduction in the more general formula for congeneric items, in which the teacher-level variance contribution to a composite is a function of both the factor covariance matrix ($\boldsymbol{\Psi}^{(W)}$) and factor loadings ($\boldsymbol{\Lambda}^{(W)}$) depicted in Figure 1:

$$\sigma^2_{t:s} = \mathbf{w}'\boldsymbol{\Lambda}^{(W)}\boldsymbol{\Psi}^{(W)}\boldsymbol{\Lambda}^{(W)}\mathbf{w}. \tag{36}$$

The variance component for the school × item interaction in Figure 2 is the residual variance of school-level item components ($\sigma^2_{si} = \theta^{(B)}_{i,i}$), constrained to equality across items. Without requiring equality constraints, the more general formulation takes the weighted sum of the residual covariance matrix:

$$\sigma^2_{si} = \mathbf{w}'\boldsymbol{\Theta}^{(B)}\mathbf{w}, \tag{37}$$

where $\boldsymbol{\Theta}^{(B)}$ can include residual covariances. Likewise, the variance component for the highest-order term in Equation (5) is the teacher-level residual variance ($\sigma^2_{(t:s)i,\epsilon} = \theta^{(W)}_{i,i}$), constrained to equality across items. But the general formulation is less constrained:

$$\sigma^2_{(t:s)i,\epsilon} = \mathbf{w}'\boldsymbol{\Theta}^{(W)}\mathbf{w}, \tag{38}$$

which accommodates tau-equivalent or congeneric items.

The final variance component is σ^2_i, which is only relevant for estimating absolute error in an agreement/dependability coefficient (Equations 11 or 18). When the factor mean is identified under a sum-to-zero constraint on intercepts, their estimates from the parallel model are interpretable as item main effects $\boldsymbol{\beta}_i$ in Equation (5) (Jorgensen, 2021); however, model-implied means

$\widehat{\mu}$ from the congeneric model are functions not only of the intercepts and factor mean, but also the factor loading (i.e., Equation 31). Instead, we can estimate σ_i^2 as a sum of squared (potentially weighted) deviations of model-implied item means from the model-implied grand mean (i.e., the mean of means: $\overline{\mu} = \frac{1}{n_i} \sum_{i=1}^{n_i} \widehat{\mu}_i$):

$$\sigma_i^2 = \frac{1}{n_i - 1} \sum_{i=1}^{n_i} [\, w_i \, (\widehat{\mu}_i - \overline{\mu})]^2 . \tag{39}$$

Note that Equation (39) is equivalent to Equation (34) when using unit weights and the intercepts are already deviations from the grand mean.

5.4 | Reliability for fixed congeneric items

Calculating reliability coefficients from estimated factor-model parameters has a long history (Bentler, 1968; Heise & Bohrnstedt, 1970; Jöreskog, 1971; McDonald, 1970). Although these coefficients have often been denoted ω (Cho, 2021) to contrast with coefficient α (Cronbach, 1951), Cho (2016) argued for a more systematic naming convention that denotes reliability estimates ($\widehat{\rho}$) with subscripts that indicate whether estimation assumes parallel ($\widehat{\rho}_{\mathrm{P}}$), tau-equivalent[15] ($\widehat{\rho}_{\mathrm{T}} = \alpha$), or congeneric items ($\widehat{\rho}_{\mathrm{C}} = \omega$). Our suggested ICC notation builds on conventions provided by McGraw and Wong (1996), whose notation did not distinguish between (a) fixed and random facets, nor between (b) parallel, tau-equivalent, or congeneric assumptions. Doing so can quickly become complicated in designs with more than one facet of generalization, as these distinctions can apply independently to different facets—for example, our continuing example below will treat items as fixed and congeneric, but teachers as random and parallel.

In this paper, we follow the convention of using a capital N to represent a fixed-facet's sample size, and we advocate treating fixed facets as congeneric in order to relax distributional assumptions. In such a case, the capital N could simultaneously imply the facet is fixed and congeneric. However, we recognize that researchers may want to treat items as fixed while fitting a model that assumes they are essentially parallel, especially if they employ software that does not permit relaxing the equality constraints necessary to reflect congeneric measurement. We do not attempt to resolve this notational distinction here, but the subscripts proposed by Cho (2016) could be applied as superscripts to each facet's sample size (e.g., $N_i^{\mathrm{C}} = 4$).

5.5 | Practical application

We used `lavaan::cfa()` to fit the CFA in Figure 1 (with metric-invariance constraints; Jak & Jorgensen, 2017; Jak et al., 2023) to the SEAS data, reporting results in the rightmost columns of Tables 2 and 3. The congeneric estimates in Table 2 are calculated using weights $\mathbf{w} = \frac{1}{N_i}\mathbf{1}$, so the composite is a mean across items, thus on the same scale as the items. Variance components in our R syntax on OSF are scaled to make their estimates from the congeneric model comparable to those from the essentially parallel model, but this scaling is not generally necessary to calculate ICCs because the scaling weights in the numerator and denominator cancel out.

The G-coefs in the first four rows of Table 3 continue to indicate unacceptably low reliability, although congeneric reliability estimates (e.g., .38 when averaging across teachers and items) are notably

[15]It is well-documented that the famous CTT coefficient α (Cronbach, 1951) is equivalent to model-based reliability when a one-factor model imposes tau-equivalence constraints, although α was derived without assuming an underlying factor model (Cronbach, 1951; Sijtsma, 2009; Sijtsma & Pfadt, 2021).

lower than corresponding parallel reliability estimates (i.e., .43). This seems to be driven largely by less school variance yet more teacher-level error, as inter-item reliability only differs in the third decimal (.855 vs. .857), whereas IRR (which is also a G-coef assuming fixed items) is lower for congeneric (.415) than parallel (.46) ICCs. The bottom two rows of Table 3 indicate the number of teachers per school necessary for parallel-model estimates to meet common reliability benchmarks ($\rho > .70$ or .80). The congeneric coefficients are lower, failing to meet these benchmarks.

6 | CONNECTIONS WITH PAST RESEARCH

For the context under consideration in our paper (shared constructs measured by cluster members' item responses), Brennan (2001, equations 4.31–4.33) originally proposed ICCs equivalent to our Equations 9, 16, and 14 to reflect generalizability of scores across cluster members, items, or both. He interpreted the coefficients that generalize over only one of the two facets of generalization (which we label as inter-item reliability and IRR) as treating the other facet as fixed, which coincides with interpretations of our coefficients as conditional correlations (Lin et al., 2025; Vangeneugden et al., 2005). In our view, the inter-item reliability and IRR coefficients have practical utility even if both facets of generalization are considered random, insofar as they can provide more information (i.e., beyond merely considering different sample sizes in a decision study) about which sources of error pose greater risks to valid comparisons between schools, or to comparing a single school to a given criterion.

Several related reliability coefficients and ICCs have been proposed for analogous multilevel designs. Ten Hove et al. (2022, Design 1) proposed IRR coefficients for designs in which raters observe targets who are nested in clusters, and measurements are made using a single item (see also Lin et al., 2025, who proposed different coefficients for Design 1). Our paper focusses on designs in which the raters are the Level-1 units, the group they are nested in is the target, and measurements are made using multiple items. In this section, we review existing multilevel reliability coefficients that can fall within that scope. Table 1 shows how and under what conditions they are related or equivalent to coefficients defined in Equations 9, 11, 14, 16, and 18 for given choices of n_i and $n_{t:s}$, and we provide additional formulas and point/interval estimates where necessary (e.g., when n_i or $n_{t:s}$ differ from Table 3). ICCs estimates in the following section that are excluded from Table 3 are available from our supplementary materials on OSF.

6.1 | Coefficients for a single multilevel variable

Textbooks on LMM (or multilevel modelling) often begin with the so-called empty or intercept-only model, in order to estimate an outcome variable's ICC (Hox et al., 2018, equation 2.9; Singer & Willett, 2003, equation 4.7; Snijders & Bosker, 2012, equations 3.2, 3.19, and 4.7). In univariate 2-level models, this ICC represents the proportion of variance in the outcome that is attributable to differences between clusters, and so can be explained by Level-2 covariates. Notation varies across textbooks, but the format always matches Equation (8) with σ^2_T replaced by the variance of Level-2 means (i.e., between-level variance: σ^2_B) and σ^2_E replaced by the remaining variance within Level-2 units (σ^2_W):

$$\mathrm{ICC_O} = \frac{\sigma^2_\mathrm{B}}{\sigma^2_\mathrm{B} + \sigma^2_\mathrm{W}}. \tag{40}$$

Bliese (2000) referred to this as ICC(1), which should not be confused with the reliability coefficient that McGraw and Wong (1996) also coined ICC(1), which we discuss below. We use the 'O' subscript for 'observed', following Kush et al. (2022, equation 11), who distinguished ICCs for observed and latent variables in multilevel SEM. We discuss the latter in an upcoming section.

When the outcome variable is meant to measure a cluster-level construct (as in our motivating example), the cluster members can be seen as 'raters' of their cluster. In such a case, Equation (40) can

be interpreted as an IRR coefficient, which Shrout & Fleiss (1979, p. 423) referred to as ICC(1,1) and McGraw & Wong (1996, table 4) referred to as ICC(1). Both of these express the reliability of a single rater's scores on a single indicator of the construct—in our motivating example, the reliability of scores from one teacher per school, using only a single item to measure parental SEAS. Table 1 lists this coefficient in the same section as the earlier IRR and inter-item coefficients because they reflect reliability of observed (or observable future) variables, although in this case it is only a single variable.

The univariate IRR coefficient in Equation (40) is related to the IRR coefficient for a $(t\colon s) \times i$ design in Equation (14). When only one scale item is used to measure the parental SEAS construct, between-item differences cannot be modelled in Equation (5), so the variance component σ^2_{si} becomes confounded with σ^2_s, and $\sigma^2_{(t:s)i,\epsilon}$ becomes confounded with $\sigma^2_{t:s}$. Therefore, those variance components disappear from Equation (14) when using only $n_i = 1$ item, reducing $\text{ICC}_{si}(\text{C}, n_{t:s} = 1, n_i = 1)$ to:

$$\text{ICC(1)} = \frac{\sigma^2_s}{\sigma^2_s + \sigma^2_{t:s}}, \tag{41}$$

which superficially appears equivalent to ICC_O in Equation (40), although the confounding makes its value equivalent to ICC_{si}. The absence of item-specific variance components from this calculation does not indicate that the single-indicator measurements lack item-specific error; rather, a single-indicator design does not allow item-specific error to be disentangled from other sources of variance.

Across the four items in our example, the range of ICC_O estimates was .198–.295, the mean of which ($M = .249$) is quite close to the expected value implied by our estimated LMM parameters—$\text{ICC}_{si}(\text{C}, n_{t:s} = 1, n_i = 1) = .252$, 90% CI [.181, .317]—and by the SEM with parallel constraints: $\text{ICC}_{si}(\text{C}, n_{t:s} = 1, n_i = 1) = .249$, 90% CI [.181, .313]. Model-implied ICC(1) per item can also be calculated from the congeneric SEM estimates, as functions of item-specific CFA parameters:

$$\text{ICC(1) for Item } i = \frac{\lambda^{(\text{B})}_{i,1}\psi^{(\text{B})}_{1,1}\lambda^{(\text{B})}_{i,1} + \theta^{(\text{B})}_{i,i}}{\lambda^{(\text{B})}_{i,1}(\psi^{(\text{B})}_{1,1} + \psi^{(\text{W})}_{1,1})\lambda^{(\text{B})}_{i,1} + \theta^{(\text{B})}_{i,i} + \theta^{(\text{W})}_{i,i}}. \tag{42}$$

We imposed metric-invariance constraints in our congeneric SEM, so $\lambda^{(\text{B})}_{i,1} = \lambda^{(\text{W})}_{i,1}$. These item-specific IRR estimates had a smaller range (.223–.238, $M = .230$) than the models that assumed parallel indicators.

6.2 | Coefficients for latent components

Hypothetically, one could imagine an indicator that does not contain any item-specific measurement error, or imagine being able to directly observe the latent multilevel variable η in Figures 1 and 2 and Equations 27 and 28:

$$\eta_{t:s} = \eta^{(\text{W})}_{t:s} + \eta^{(\text{B})}_s. \tag{43}$$

The same coefficient in Equations 40 and 41 could then hypothetically be calculated for $\eta_{t:s}$, indicating the (hypothetical) IRR for measurements free of item-specific error. Despite the latent variable $\eta_{t:s}$ not being observable (hence shown in the lowest section of Table 1), one can calculate this 'hypothetical reliability' as a function of the variance components for $\eta^{(\text{W})}_{t:s}$ and $\eta^{(\text{B})}_s$:

$$\text{ICC}_\text{L} = \frac{\psi^{(\text{B})}_{1,1}}{\psi^{(\text{B})}_{1,1} + \psi^{(\text{W})}_{1,1}}. \tag{44}$$

Analogous to ICC_O, Kush et al. (2022, equation 14) labelled this ICC_L to indicate it would be the latent variable's ICC. For configural constructs (i.e., representing characteristics of Level-1 units; Stapleton et al., 2016; Jak et al., 2023), ICC_L is useful to interpret as the proportion of individual differences in the construct that can be attributed to differences between clusters. For a shared construct (e.g., our parental SEAS example), it can be interpreted analogous to the IRR coefficient ICC(1), if raters were able to directly observe the latent group-level construct.

Note that interpreting $\eta_s^{(B)}$ as the cluster-level means of $\eta_{i:s}$ implies metric invariance (equality of factor loadings) across clusters, which further implies equivalent factor loadings across levels (Jak et al., 2013; Jak & Jorgensen, 2017). In our supplemental materials on OSF, the congeneric SEM we fitted to obtain estimates in Table 2 reflected this metric-invariance assumption. Furthermore, the latent scale can be identified by estimating both the within- and between-level factor variance under the constraint that their sum is 1, in which case the estimated $\widehat{\psi}_{1,1}^{(B)} = .296$, 95% CI [.171, .422], is already a direct estimate of ICC_L. By comparison, the SEM with parallel constraints yielded a higher estimated $ICC_L = .343$, 90% CI [.234, .452], consistent with parallel vs. congeneric coefficients in Table 3.

Geldhof et al. (2014) also proposed reliability coefficients for latent components of multilevel variables, but rather than IRR coefficients[16], they proposed (hypothetical) inter-item reliability coefficients for configural constructs. They proposed applying single-level composite-reliability formulas (α or model-based ω) to the latent components at each level of analysis (right-hand side of Equation 24). Level-2 model-based reliability would thus be calculated only from measurement parameters in Equations 26, 28, and 30, omitting all within-level variance that reflect differences between observations within a given cluster. In terms of the inter-item reliability coefficient we proposed in Equation (16) for our motivating example, the between-level coefficient proposed by Geldhof et al. (2014, see also Lai, 2021, equation 7) omits variance among teachers within the same school:

$$\frac{\sigma_s^2}{\sigma_s^2 + \frac{\sigma_{si}^2}{n_i}}. \tag{45}$$

For our example data, Geldhof et al.'s Level-2 $\widehat{\alpha} = .928$ and $\widehat{\omega} = .917$ both indicate very little item-specific error among school averages. This is a common finding, consistent with only minor violations of scalar invariance across clusters (i.e., differences in intercepts across clusters are captured by nonzero residual variances of indicators' Level-2 components; Jak & Jorgensen, 2017).

Compared to the ICC_s(C, $n_{i:s} = \tilde{n}_{i:s}, n_i = 4$) estimates in Table 2, Geldhof et al.'s (2014) high Level-2 'reliability' coefficients have a dubious interpretation. Instead of measurements that are free of item-specific error, imagine measurements that are free of rater error, such that any teacher within the same school would provide the same score (using a particular item) as any other teacher from the same school. One might then interpret the cluster-level construct as 'global' rather than "shared" (Jak et al., 2023). Granted, the multilevel CFA does not assume this because within-level variance is modelled. However, the coefficient in Equation (45) reflects the hypothetical reliability of a composite calculated from each item's latent cluster mean ($\tilde{J}$), which is not a composite one could calculate in practice, and greatly exaggerates how reliable an actual Level-2 composite would be (Lai, 2021). Because it is not a true reliability coefficient (i.e., it does not reflect reliability of an observed or at least observable variable), it is also included in the bottom section of Table 1.

[16]Ten Hove et al. (2022, Design 1, pp. 653–654) proposed IRR coefficients for a multilevel design comparable to the one under discussion here—$(s{:}\,c) \times r$—one where the raters r are independent observers of subjects s nested within clusters c (e.g., evaluators of student–teacher relationships who observe multiple students per teacher). Whereas we propose ICCs for shared constructs, Ten Hove et al. (2022, table 3) proposed cluster-level ICCs for configural constructs that are analogous to those of Geldhof et al. (2014), but with raters (fully crossed with the nested facets) interchanged with items. Thus, all discussion of Geldhof et al.'s (2014) coefficients also applies to those of Ten Hove et al. (2022).

6.3 | Coefficients for a cluster composite

Relative to sampling a single teacher's rating per school, cluster means are more reliable estimates of a cluster-level construct, as the error variance per school is proportionally minimized relative to the target variance present in each teacher's rating. The reliability of cluster means is expressed by dividing σ^2_{W} in Equation (40) (or $\sigma^2_{t:s}$ in Equation 41) by the number of Level-1 units per cluster (e.g., $n_{t:s}$ in our motivating example):

$$\mathrm{ICC}(n_{t:s}) = \frac{\sigma^2_{B}}{\sigma^2_{B} + \frac{\sigma^2_{W}}{n_{t:s}}} = \frac{\sigma^2_{s}}{\sigma^2_{s} + \frac{\sigma^2_{t:s}}{n_{t:s}}} = \frac{n_{t:s} \times \mathrm{ICC}_{O}}{1 + (n_{t:s} - 1)\mathrm{ICC}_{O}}. \tag{46}$$

Bliese (2000) referred to this as ICC(2), regardless of the cluster size, which can equivalently be expressed as a function of ICC_O and the (average) cluster size, as in the far-right term of Equation (46) (see also Snijders & Bosker, 2012, equation 3.21). Shrout and Fleiss (1979) referred to Equation (46) as ICC(1,k) and McGraw & Wong (1996, table 5) referred to it as ICC(k), where k is the number of raters. For unbalanced cluster sizes (as in our motivating example), the harmonic-mean cluster size is used ($k = \tilde{n}_{t:s} = 1.908$; Putka et al., 2008).

Across the four items in our example, the range of ICC(1.908) estimates was .320–.445, the mean of which ($M = .387$) is quite close to our estimated $\mathrm{ICC}_{si}(\mathrm{C}, n_{t:s} = 1.908, n_i = 1) = .391$, 90% CI [.300, .468], obtained from LMM parameters, as well as from the SEM with parallel constraints: $\mathrm{ICC}_{si}(\mathrm{C}, n_{t:s} = 1, n_i = 1) = .388$, 90% CI [.296, .465]. Model-implied ICC(1.908) per item was also calculated from the congeneric SEM estimates, dividing the within-level variances ($\psi^{(W)}_{1,1}$ and $\theta^{(W)}_{i,i}$) by 1.908 in Equation (42). As with ICC_O = ICC(1), these item-specific ICC(1.908) estimates had a smaller range (.354–.373, $M = .363$) than the models that assumed parallel indicators.

6.4 | Coefficients for a composite of items

Lai (2021) proposed some valid reliability formulas for observable composites calculated from multilevel, multiple-item scales, some of which are equivalent to ones we proposed here specifically for cluster-level constructs. Lai (2021) criticized[17] Geldhof et al. (2014) for applying composite-reliability formulas to a situation where a composite cannot actually be calculated. Although Geldhof et al. (2014, p. 75) referred to GT as a guideline for variance decomposition, they did not propose a G-coef for a composite at either level of analysis. Lai (2021) did not refer to GT, making reference only to CTT; however, he defined some reliability coefficients that are consistent with G-coefs, despite the CTT framework not allowing for multiple sources of error. The lack of connection to GT in these two papers might follow from mistakenly assuming that GT requires the parallel assumption, given Vispoel et al. (2021) had not yet published their paper showing how GT can incorporate congeneric measurement models via SEM (see also Vispoel et al., 2022).

When calculating a composite to represent a shared construct, by averaging over both raters and items, we propose the G-coef in Equation 9 (see also Brennan, 2001, equation 4.31), reported as $\mathrm{ICC}_s(\mathrm{C}, n_{t:s} = 1.908, n_i = 4)$ in Table 3. Our G-coef is equivalent[18] to what Lai (2021, equations 24 & 17

[17] Lin et al. (2025) raised an analogous critique of cluster-level ICCs proposed by Ten Hove et al. (2022). Just as Ten Hove et al. (2022, for raters) and Geldhof et al. (2014, for items) proposed analogous hypothetical cluster-level reliability coefficients (for latent components; see Footnote 16), Lin et al. (2025, equations 19 and 20) proposed valid reliability formulas for observable composites of (crossed) rater responses about subjects within clusters, and their index of consistency is analogous to the reliability formulas proposed by Lai (2021) for composites of item responses about clusters by their members (nested raters).

[18] Our G-coef and D-coef are analogously equivalent to the manifest cluster-level reliability coefficients proposed by Lin et al. (2025, equations 19 and 20), when replacing items by independent raters of the subjects within clusters (where raters are crossed with clusters, rather than nested within clusters as in our motivating example).

for α^b & ω^b, respectively) proposed for a Level-2 composite of a configural construct (see Table 1). For a shared construct, Lai (2021, equation 20) proposed a similar formula that replaces within-level CFA parameters with a saturated within-level covariance matrix. This follows from Stapleton et al. (2016) proposing that measurement of shared constructs should be modelled without imposing constraints, which is a recommendation we criticize in an upcoming section.

Regarding the same composite (i.e., averaging over both raters and items to represent a shared construct), we proposed a inter-item reliability coefficient in Equation (16) to represent the proportion of composite-score variance not associated with item-specific measurement error, which is reported in Table 3 as $\text{ICC}_{t:s}(\text{C}, n_{t:s} = \tilde{n}_{t:s} = 1.908, n_i = 4)$. One could also be interested in an inter-item reliability coefficient if a single rater per cluster were sampled, so the composite only averages over items. We would label this $\text{ICC}_{t:s}(\text{C}, n_{t:s} = 1, n_i = 4)$, which is equivalent to what Lai (2021, equations 23 and 13 for α^{2l} and ω^{2l}, respectively) (see Table 1) proposed for an 'overall' composite of a configural construct (i.e., to represent Level-1 scores without cluster-mean centering to decompose the Level-1 and -2 components). From the SEM with parallel constraints, the estimated $\text{ICC}_{t:s}(\text{C}, n_{t:s} = 1, n_i = 4) = .837$, 90% CI [.811, .858], is quite similar to the estimate from the congeneric SEM, $\text{ICC}_{t:s}(\text{C}, n_{t:s} = 1, n_i = 4) = .839$, 90% CI [.814, .860], which is consistent with inter-item reliabilities being quite similar in Table 3.

Note that the $\text{ICC}_{t:s}(\text{C}, \tilde{n}_{t:s} = 1, n_i = 4)$ estimates above are asymptotically equivalent to simply applying a inter-item reliability formula for single-level data to the sample of N_t teachers, ignoring that they are nested in schools. Geldhof et al. (2014) advised against doing so, expressing concern that this estimate of reliability would be biased if reliability differed across levels—a hypothesis they confirmed when simulating data from a model in which the metric-invariance assumption did not hold. However, their conception of level-specific reliability presupposes that the construct exists independently at each level of analysis, whereas we view the measurement of a shared construct as necessitating the metric-invariance assumption[19] by definition (Jak et al., 2023). From the GT perspective, ignoring the nesting of teachers in schools reduces the design to $t \times i$, leaving variance components confounded (see our discussion of IRR in Equation 41). This reduces $\text{ICC}_{t:s}(\text{C}, n_{t:s}, n_i)$ in Equation (16) to

$$\text{ICC}_t(\text{C}, n_i) = \frac{\sigma^2_t}{\sigma^2_t + \frac{\sigma^2_{ti}}{n_i}}, \tag{47}$$

where σ^2_s and $\sigma^2_{t:s}$ are confounded in σ^2_t, and σ^2_{si} and $\sigma^2_{(t:s)i,\epsilon}$ are confounded in σ^2_{ti}. In the population, this leads to the same inter-item reliability coefficient one would obtain by modelling the multilevel structure and including those (unconfounded) variance components in Equation (16). In finite samples, however, the unconfounded variance components in ML-SEM can vary independently (leading to differences in sample estimates) and will generally be estimated with less precision because a ML-SEM is more highly parameterized than a single-level SEM, which estimates fewer (but confounded) variance components. In Table 3, the $\text{ICC}_{t:s}(\text{C}, \tilde{n}_{t:s} = 1, n_i = 4) = .855$, 90% CI [.825, .879], CI width = .054, estimated from ML-SEM is slightly higher than a inter-item reliability coefficient estimated from a single-level SEM, $\text{ICC}_t(\text{C}, n_i = 4) = .842$, which has a narrower 90% CI [.820, .861], CI width = .041.

6.5 | Coefficients from a shared-and-configural model

In his appendix, Lai (2021, equation A1) also proposed how to calculate a reliability coefficient for a shared-construct composite ($\omega^{b(s)}$), using estimates from the 'shared-and-configural' model proposed

[19]It is worth noting that our online supplementary R syntax provides some evidence against this assumption, in that a likelihood ratio test showed significantly better fit without equating loadings across levels, $\chi^2(3) = 10.76, p = .013$, indicating it might be informative to probe which item's loadings vary across schools.

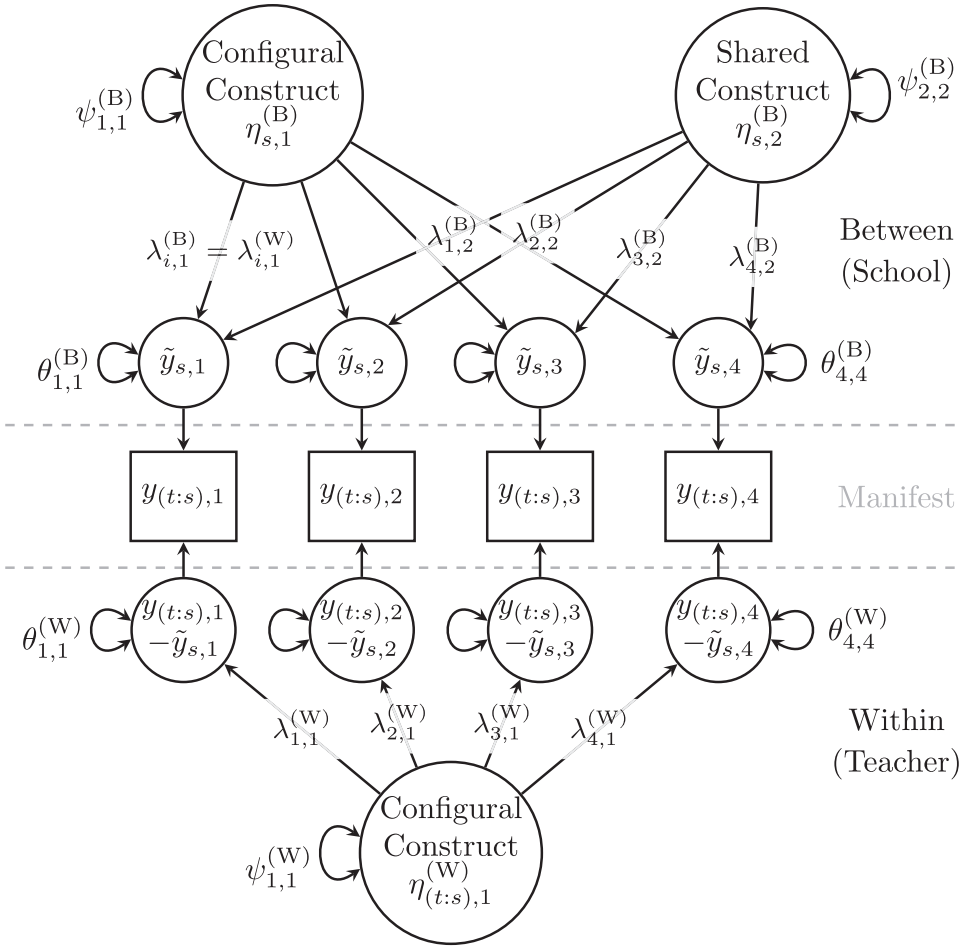


FIGURE 3 Path diagram depicting a 4-indicator multilevel CFA with both shared and configural factors, as proposed by Stapleton et al. (2016) and advocated by Stapleton and Johnson (2019).

by Stapleton et al. (2016, see also Stapleton & Johnson, 2019). Some conceptual and methodological problems of this model were discussed by Jak et al. (2023), most of which are only indirectly relevant to defining reliability coefficients for shared constructs (e.g., insofar as they reveal threats to the model's validity). Here, we discuss limitations of the shared-and-configural model (depicted in Figure 3) that become apparent from the GT perspective, casting doubt on the usefulness of a reliability coefficient calculated from that model.

6.5.1 | Substantive interpretation

The GT model in Equation (5) includes a random effect of schools ($\boldsymbol{\beta}_s$), which is analogous to CTT's true scores for a shared construct measured with this $(t{:}\,s) \times i$ design. This effect is represented by the between-level common factor ($\boldsymbol{\eta}^{(B)}_{s,1}$) in Figures 1 and 2. The shared-and-configural model includes two such random effects, represented as two orthogonal common factors in the Level-2 model of Figure 3. This Level-2 factor structure resembles an exploratory factor analysis (EFA) model, with the restriction that the loadings of one factor (i.e., labelled the 'configural' factor) are equated with the within-level

factor loadings. Stapleton and Johnson (2019) argued that the configural factor reflects individual perceptions of the shared construct ($\eta^{(W)}_{s,1}$), rather than the actual shared construct itself, yet Stapleton and Johnson (2019) argue it still has a cluster-level component ($\eta^{(B)}_{s,1}$) that is independent from the shared construct (now represented by $\eta^{(B)}_{s,2}$).

When a shared construct is measured by group members, we agree that measurements reflect the raters' individual perceptions—otherwise, it would be a global rather than shared construct (Jak et al., 2023). Controlling for item-specific differences, the decomposition in Equation (5) has terms for what is shared among individual perceptions (β_s) as well as how those perceptions differ among individuals rating the same cluster ($\beta_{t:s}$). Neither CTT nor GT require interpreting $\beta_s = \eta^{(B)}_s$ as strictly reflecting the target construct; rather, a true (or universe) score is the average across hypothetically or actually repeated measurements (e.g., Raykov & Marcoulides, 2011, p. 118). Accordingly, reliability (or generalizability) coefficients reflect the squared correlation of individual scores with true (or universe) scores, not necessarily with scores on the actual construct of interest.

Naturally, it is often of theoretical interest to determine the proportion of variance that is not merely reliable, but also valid (i.e., attributable to the target construct). Bifactor modelling, for example, allows for computing not only a reliability coefficient (so-called 'total' ω_T) but also so-called "hierarchical" ω_H (McDonald, 1999; Zinbarg et al., 2006), which is the proportion of variance due to the target construct, omitting any reliable but nuisance variance of composite scores attributable to 'method' factors. Stapleton and Johnson (2019) suggested that the shared-and-configural model could disentangle 'actual, objective safety of the neighbourhood (a shared cluster-level characteristic) and the individual's tolerance for unsafe conditions' (p. 314), but doing so would require a measurement design that allows multiple sources of (co)variance to be disentangled. Their illustrative example also analysed TIMSS data: a teacher-level construct *engaging teaching*, measured by multiple students per teacher with $n_i = 6$ items. Stapleton and Johnson (2019) suggested the within-level factor could represent raters' *acquiescence*, though that is just one among 'a range of possible nuisance characteristics or conditions' (p. 320). A long methodological history of modelling acquiescence bias would suggest that an acquiescence factor should load on items across more than one construct, with loadings constrained to equality (similar to parallel constraints in Figure 2; Billiet & McClendon, 2000; Maydeu-Olivares & Coffman, 2006; Savalei & Falk, 2014). Although we acknowledge that acquiescence is a theoretically justifiable source of variance to expect in students' responses to the TIMSS indicators of engaging teaching, we would caution against concluding that a configural factor $\eta^{(B)}_{s,1}$ (distinct from the shared construct $\eta^{(B)}_{s,2}$ in Figure 3) has captured that (or any specific source of) variance. In fact, given the propensity of a secondary factor (as in EFA) to capture any source of multidimensionality, nonlinearity, etc., we are more inclined to interpret Stapleton and Johnson's (2019) 'configural' factor as representing the target construct and the factor they labelled 'shared' as representing nuisance variance.

6.5.2 | Identifiability

Taking the GT model in Equation (5) as a starting point, the two between-level factors of the shared-and-configural model would be analogous to having two random intercepts for schools:

$$y_{(t:s)i} = \mu + (\gamma_s + \delta_s) + \beta_i + \beta_{si} + \beta_{t:s} + \beta_{(t:s)i,\epsilon}, \tag{48}$$

where γ_s and δ_s represent the shared and configural factors, respectively. Equation (48) reveals a practical limitation that the model would be underidentified, as the random effects γ_s and δ_s would

be perfectly confounded. The congeneric model allows the shared and configural factors to be distinguished only in the case that one factor's pattern of loadings is not proportionally equivalent to the other factor's pattern. It may be unlikely in practice that two distinct factors have exactly proportional patterns of factor loadings[20], but even approximately proportional patterns could yield predictable convergence limitations observed by Stapleton and Johnson (2019). For example, if the configural construct in Figure 3 had loadings $[.8, .8, .6, .4]'$ and the shared construct had loadings $[.5, .5, .4, .3]'$, the two factors might not be empirically distinguishable.

For Stapleton and Johnson (2019) to be able to estimate the shared-and-configural model parameters, they imposed certain identifiability conditions that we think render their reported solution(s) suspect. They fixed the first indicator's factor loading of each factor to 1, estimating all 3 factor variances,[21] but under the constraint that the configural factor's ICC_L had a fixed value. In fact, their model would not converge unless they fixed $ICC_L < .15$ (Stapleton et al., 2016, p. 323), which we suspect is a sign that $\eta^{(B)}_{s,1}$ already captured the shared-construct variance, rendering the secondary $\eta^{(B)}_{s,2}$ factor redundant. This would be hidden by artificially limiting the degree to which $\eta^{(B)}_{s,1}$ could contribute shared variance to the indicators, thus allowing for $\eta^{(B)}_{s,2}$ to contribute the 'missing' between-level (co)variance.

Stapleton and Johnson (2019) did acknowledge the arbitrariness of fixing ICC_L to unknown values, suggesting that 'additional items, measured at the cluster level' (p. 325) could be used to identify a model, but the existence of such an item (i.e., not determined by individual perceptions) might render individual-level responses unnecessary. Perhaps if directly obtaining such cluster-level measurements were 'expensive', then collecting it for a subset of clusters could allow valid construct variance to be distinguished from aggregated individual perceptions, analogous to a two-method planned missing-data design (Graham et al., 2006).

7 | DISCUSSION

This paper used GT to propose a taxonomy of ICCs (summarized by Table 1) that quantify composite reliability of cluster-level constructs measured by cluster members' responses to multiple items, the latter of which can be treated as random parallel measurements or as fixed, congeneric items with uniquely estimable parameters. We demonstrated how to estimate variance components and ICCs using open-source R packages for LMM and SEM, with data and R scripts available on the OSF. We showed that many previously proposed reliability coefficients for this context fit within our taxonomy, criticising the validity of some (e.g., Geldhof et al., 2014; Ten Hove et al., 2022) yet defending the validity of others (e.g., inter-item reliability of a cluster composite, ignoring the clustering of individual responses) on the condition they are interpreted correctly.

7.1 | Advice

Researchers should clearly communicate their measurement design, which informs which ICC nomenclature to use when reporting reliability estimates. For example, residents nested in neighbourhoods

[20] As a reviewer pointed out, even if two such factors were able to be extracted in an EFA, one would struggle to interpret them as being different factors at all. This is the point we made in the previous section on Substantive Interpretation.

[21] Other identification methods would be equivalent, such as our congeneric example: by estimating both factor variances under the constraint that they sum to 1 (so the between-level factor variance is ICC_L). Alternatively, the within-level factor's variance can be fixed to $\psi^{(W)}_{1,1} = 1$, allowing the between-level factor variance $\psi^{(B)}_{1,1}$ to be freely estimated because its loadings are equivalent to within-level loadings (i.e., Stapleton & Johnson, 2019, also imposed metric-invariance constraints in their illustrative example).

(r: n) might rate their neighbourhood safety (a group-level target construct) with multiple items (i)—this could be described as a (r: n) × i design, for which the G-coef in Equation (9) could be labelled $ICC_n(C, n_{r:n}, n_i)$.

Various ICCs can be reported, depending on the context to which it is meant to apply. For example, if a researcher is using the current sample to estimate reliability of data they plan to obtain in the future, then varying n_i and $n_{t:s}$ in Equations 9, 11, 14, 16, and 18 can inform how the design protocol could balance costs associated with more items or individuals in order to obtain more reliable composites. However, researchers often use the same data to estimate reliability and to test substantive hypotheses about how the measured construct relates to other variables. In this case, ICCs should be calculated using n_i and $n_{t:s}$ that reflect the actual design.

Researchers should bear in mind that only the G-coef in Equation (9) reflects the degree to which estimated correlation/regression coefficients would be attenuated by all sources of measurement error accounted for by the measurement design. The IRR and inter-item reliability coefficients can inform which design features introduce more error than others, but they do not necessarily reflect the actual reliability of the composite, unless one of the facets is expected to be fixed (i.e., identical in future samples). For the same reason, we do not generally recommend (hypothetical) reliability coefficients for latent components (e.g., ICC_L or Geldhof et al.'s (2014) level-specific coefficients, in the bottom rows of Table 1) because they lack the key characteristic of reflecting the degree[22] of an actual composite variable's attenuated effect.

Researchers often report reliability estimates from previous research. We recommend providing both point and interval estimates for any reported ICCs, to transparently reflect uncertainty about the true reliability. Because specific goals may not always match those of previous research, researchers should strive to provide as much information as possible to enable readers to calculate any particular ICC of interest. Thus, we recommend that researchers always report all variance-component estimates from a fitted LMM—or all parameters from a multilevel CFA—so that readers can calculate point estimates of any ICCs of interest for their particular goals. For readers to also obtain interval estimates for their ICCs, it would be necessary to share the raw data (preferably with software scripts), which we encourage whenever it would not risk the privacy of participants who provided their data.

7.2 | Limitations and future directions

For unbalanced designs, the harmonic mean ($\tilde{n}$, Equation 13) can be used as the sample size for that component in ICC formulas (Putka et al., 2008). This is a general solution for incomplete data (Ten Hove et al., 2024), enabling ICC estimation with planned missing data. For example, using the three-form design (Graham et al., 1996) with our motivating example, each teacher would have responded to only 3 out of 4 indicators. Thus, this 4-item design would yield composites calculated with $n_i = 3$ indicators, so $n_i = 3$ should be used in reliability formulas.

This is straightforward with parallel items, but congeneric items have varying reliability. Thus, reliability will differ between subsamples of participants who completed the three different forms. In this case, it may be preferable to calculate reliability for each composite, by applying Equations 35–39 with weights $w_i = 0$ for items excluded from a composite. When there are additional unplanned missing values, the harmonic mean $\tilde{n}$ could be used within each subsample. However, future research may reveal how to calculate an overall reliability coefficient for composites calculated from varying subsets of congeneric items.

[22]Note that the exact degree of attenuation involves more complicated formulae for partial-correlation or multiple-regression coefficients, or for coefficients in multilevel models (e.g., Cho & Preacher, 2016).

8 | CONCLUSION

GT provides a flexible framework for ICCs to be defined for complex measurement designs, expressing reliability across various dimensions of measurement error. When measuring group-level constructs with responses of group members to multiple-item scales, ICCs can be reported to express inter-item reliability, IRR, or generalizability across both items and raters. This allows researchers to evaluate measurement procedures in fine-grain detail and predict the most efficient way to improve the measurement tool or protocol, to attain sufficiently reliable measurements for various purposes.

AUTHOR CONTRIBUTIONS

Terrence D. Jorgensen: conceptualization; methodology; writing – original draft; software; formal analysis; visualization; data curation. **Suzanne Jak:** writing – review and editing; funding acquisition.

ACKNOWLEDGEMENTS

We acknowledge Ye et al. (2024) for providing the motivating example in our paper, using TIMSS data available on their Open Science Framework (OSF) project associated with their open-access publication (Ye et al., 2022). Likewise, our OSF project provides the (transformed) data and R syntax files to reproduce all reported analyses in this paper: Terrence D. Jorgensen was responsible for the conceptualization, data curation, formal analysis, methodology, software, visualization, and writing (original draft preparation). Suzanne Jak was responsible for writing (reviewing and editing) and acquired funding (Dutch Research Council (NWO) project number VI.Vidi.201.009).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

Our OSF project provides the (transformed) data and R syntax files to reproduce all reported analyses in this paper: https://osf.io/vg5uq/.

DISCLOSURE OF THE USE OF ARTIFICIAL INTELLIGENCE-GENERATED CONTENT (AIGC) TOOLS

No generative AI tools were used in the creation of this article.

ORCID

Terrence D. Jorgensen https://orcid.org/0000-0001-5111-6773
Suzanne Jak https://orcid.org/0000-0002-2223-5594

REFERENCES

Bartko, J. J. (1966). The intraclass correlation coefficient as a measure of reliability. *Psychological Reports*, *19*(1), 3–11. https://doi.org/10.2466/pr0.1966.19.1.3

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bentler, P. (1968). Alpha-maximized factor analysis (alphamax): Its relation to alpha and canonical factor analysis. *Psychometrika*, *33*(3), 335–345. https://doi.org/10.1007/BF02289328

Billiet, J. B., & McClendon, M. J. (2000). Modeling acquiescence in measurement models for two balanced sets of items. *Structural Equation Modeling*, *7*(4), 608–628. https://doi.org/10.1207/S15328007SEM0704_5

Bliese, P. D. (2000). Within-group agreement, non-independence, and reliability: Implications for data aggregation and analysis. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations* (pp. 349–381). Jossey-Bass/Wiley.

Bollen, K. A. (1989). *Structural equations with latent variables*. Wiley. https://doi.org/10.1002/9781118619179

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics*, *27*(4), 335–340. https://doi.org/10.3102/10769986027004335

Bonett, D. G., & Wright, T. A. (2015). Cronbach's alpha reliability: Interval estimation, hypothesis testing, and sample size planning. *Journal of Organizational Behavior*, *36*(1), 3–15. https://doi.org/10.1002/job.1960

Brennan, R. L. (2001). *Generalizability theory*. Springer. https://doi.org/10.1007/978-1-4757-3456-0

Brennan, R. L. (2010). Generalizability theory and classical test theory. *Applied Measurement in Education*, *24*(1), 1–21. https://doi.org/10.1080/08957347.2011.532417

Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods*, *19*(4), 651–682. https://doi.org/10.1177/1094428116656239

Cho, E. (2021). Neither Cronbach's alpha nor McDonald's omega: A commentary on Sijtsma and Pfadt. *Psychometrika*, *86*(4), 877–886. https://doi.org/10.1007/s11336-021-09801-1

Cho, S.-J., & Preacher, K. J. (2016). Measurement error correction formula for cluster-level group differences in cluster randomized and observational studies. *Educational and Psychological Measurement*, *76*(5), 771–786. https://doi.org/10.1177/0013164415612255

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

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*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, *19*(1), 72–91. https://doi.org/10.1037/a0032138

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research*, *31*(2), 197–218. https://doi.org/10.1207/s15327906mbr3102_3

Graham, J. W., Taylor, B. J., Olchowski, A. E., & Cumsille, P. E. (2006). Planned missing data designs in psychological research. *Psychological Methods*, *11*(4), 323–343. https://doi.org/10.1037/1082-989X.11.4.323

Gwet, K. L. (2021). *Handbook of inter-rater reliability, volume 2: Analysis of quantitative ratings* (5th ed.). Advanced Analytics, LLC.

Heise, D. R., & Bohrnstedt, G. W. (1970). Validity, invalidity, and reliability. *Sociological Methodology*, *2*, 104–129. https://doi.org/10.2307/270785

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2018). *Multilevel analysis: Techniques and applications* (3rd ed.). Routledge.

Jak, S., & Jorgensen, T. D. (2017). Relating measurement invariance, cross-level invariance, and multilevel reliability. *Frontiers in Psychology*, *8*(1640), 1640. https://doi.org/10.3389/fpsyg.2017.01640

Jak, S., Jorgensen, T. D., ten Hove, D., & Nevicka, B. (2023). Modeling cluster-level constructs measured by individual responses: Configuring a shared approach. *Advances in Methods and Practices in Psychological Science*, *6*(3), 1–18. https://doi.org/10.1177/25152459231182319

Jak, S., Oort, F. J., & Dolan, C. V. (2013). A test for cluster bias: Detecting violations of measurement invariance across clusters in multilevel data. *Structural Equation Modeling*, *20*(2), 265–282. https://doi.org/10.1080/10705511.2013.769392

Jiang, J. (2007). *Linear and generalized linear mixed models and their applications*. Springer. https://doi.org/10.1007/978-0-387-47946-0

Jiang, Z. (2018). Using the linear mixed-effect model framework to estimate generalizability variance components in R: A lme4 package application. *Methodology*, *14*(3), 133–142. https://doi.org/10.1027/1614-2241/a000149

Jiang, Z., Raymond, M., DiStefano, C., Shi, D., Liu, R., & Sun, J. (2022). A Monte Carlo study of confidence interval methods for generalizability coefficient. *Educational and Psychological Measurement*, *82*(4), 705–718. https://doi.org/10.1177/0013164421 1033899

Jöreskog, K. G. (1971). Statistical analysis of sets of congeneric tests. *Psychometrika*, *36*(2), 109–133. https://doi.org/10.1007/BF02291393

Jorgensen, T. D. (2021). How to estimate absolute-error components in structural equation models of generalizability theory. *Psych*, *3*(2), 113–133. https://doi.org/10.3390/psych3020011

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2026). semTools: Useful tools for structural equation modeling [Computer software manual]. (R package version 0.5-9). https://doi.org/10.32614/CRAN.package.semTools

Kush, J. M., Konold, T. R., & Bradshaw, C. P. (2022). The sampling ratio in multilevel structural equation models: Considerations to inform study design. *Educational and Psychological Measurement*, *82*(3), 409–443. https://doi.org/10.1177/0013164421 1020112

Lai, M. H. (2021). Composite reliability of multilevel data: It's about observed scores and construct meanings. *Psychological Methods*, *26*(1), 90–102. https://doi.org/10.1037/met0000287

Li, H. (1997). A unifying expression for the maximal reliability of a linear composite. *Psychometrika*, *62*(2), 245–249. https://doi.org/10.1007/BF02295278

Lin, T.-Y., Tuerlinckx, F., & Vanbelle, S. (2025). Reliability for multilevel data: A correlation approach. *Psychological Methods*. https://doi.org/10.1037/met0000738

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Addison-Welsley.

Marcoulides, G. A. (1996). Estimating variance components in generalizability theory: The covariance structure analysis approach. *Structural Equation Modeling*, *3*(3), 290–299. https://doi.org/10.1080/10705519609540045

Maydeu-Olivares, A., & Coffman, D. L. (2006). Random intercept item factor analysis. *Psychological Methods*, *11*(4), 344–362. https://doi.org/10.1037/1082-989x.11.4.344

McDonald, R. P. (1970). The theoretical foundations of principal factor analysis, canonical factor analysis, and alpha factor analysis. *British Journal of Mathematical and Statistical Psychology*, *23*(1), 1–21. https://doi.org/10.1111/j.2044-8317.1970.tb00432.x

McDonald, R. P. (1999). *Test theory: A unified treatment*. Psychology Press. https://doi.org/10.4324/9781410601087

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

Molenberghs, G., & Verbeke, G. (1997). *Linear mixed models for longitudinal data*. Springer. https://doi.org/10.1007/978-1-4419-0300-6

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019: International results in mathematics and science*. (Tech. Rep.). Lynch School of Education and Human Development, Boston College and International Association for the Evaluation of Educational Achievement (IEA). https://www.iea.nl/sites/default/files/2021-01/TIMSS%202019-International-Results-in-Mathematics-and-Science.pdf

Nicewander, W. A. (2018). Conditional reliability coefficients for test scores. *Psychological Methods*, *23*(2), 351–362. https://doi.org/10.1037/met0000132

Preacher, K. J., & Selig, J. P. (2012). Advantages of Monte Carlo confidence intervals for indirect effects. *Communication Methods and Measures*, *6*(2), 77–98. https://doi.org/10.1080/19312458.2012.679848

Putka, D. J., Le, H., McCloy, R. A., & Diaz, T. (2008). Ill-structured measurement designs in organizational research: Implications for estimating interrater reliability. *Journal of Applied Psychology*, *93*(5), 959–981. https://doi.org/10.1037/0021-9010.93.5.959

R Core Team. (2026). R: A language and environment for statistical computing. (version 4.6.1) [Computer software manual]. Vienna, Austria. https://www.R-project.org.

Raykov, T., & Marcoulides, G. A. (2006). Estimation of generalizability coefficients via a structural equation modeling approach to scale reliability evaluation. *International Journal of Testing*, *6*(1), 81–95. https://doi.org/10.1207/s15327574ijt0601_5

Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge. https://doi.org/10.4324/9780203841624

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling and more. *Journal of Statistical Software*, *48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Rosseel, Y., Jorgensen, T. D., & De Wilde, L. (2026). lavaan: Latent variable analysis [Computer software manual]. (R package version 0.7-2). https://doi.org/10.32614/CRAN.package.lavaan

Savalei, V., & Falk, C. F. (2014). Recovering substantive factor loadings in the presence of acquiescence bias: A comparison of three approaches. *Multivariate Behavioral Research*, *49*(5), 407–424. https://doi.org/10.1080/00273171.2014.931800

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

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Sijtsma, K., & Pfadt, J. M. (2021). Part II: On the use, the misuse, and the very limited usefulness of Cronbach's alpha: Discussing lower bounds and correlated errors. *Psychometrika*, *86*(4), 843–860. https://doi.org/10.1007/s11336-021-09789-8

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford University Press.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Stapleton, L. M., & Johnson, T. L. (2019). Models to examine the validity of cluster-level factor structure using individual-level data. *Advances in Methods and Practices in Psychological Science*, *2*(3), 312–329. https://doi.org/10.1177/2515245919855039

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, *41*(5), 481–520. https://doi.org/10.3102/1076998616646200

Ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2022). Interrater reliability for multilevel data: A generalizability theory approach. *Psychological Methods*, *27*(4), 650–666. https://doi.org/10.1037/met0000391

Ten Hove, D., Jorgensen, T. D., & van der Ark, L. A. (2024). Updated guidelines on selecting an intraclass correlation coefficient for interrater reliability, with applications to incomplete observational designs. *Psychological Methods*, *29*(5), 967–979. https://doi.org/10.1037/met0000516

Ten Hove, D., Jorgensen, T. D., & Van der Ark, L. A. (2025). How to estimate intraclass correlation coefficients for interrater reliability from planned incomplete data. *Multivariate Behavioral Research*, *60*(5), 1042–1061. https://doi.org/10.1080/00273171.2025.2507745

Vangeneugden, T., Laenen, A., Geys, H., Renard, D., & Molenberghs, G. (2005). Applying concepts of generalizability theory on clinical trial data to investigate sources of variation and their impact on reliability. *Biometrics*, *61*(1), 295–304. https://doi.org/10.1111/j.0006-341X.2005.031040.x

Vispoel, W. P., Hong, H., Lee, H., & Jorgensen, T. D. (2023). Analyzing complete generalizability theory designs using structural equation models. *Applied Measurement in Education*, *36*(4), 372–393. https://doi.org/10.1080/08957347.2023.2274573

Vispoel, W. P., Lee, H., Xu, G., & Hong, H. (2022). Expanding bifactor models of psychological traits to account for multiple sources of measurement error. *Psychological Assessment*, *34*(12), 1093–1111. https://doi.org/10.1037/pas0001170

Vispoel, W. P., Morris, C. A., & Kilinc, M. (2018). Applications of generalizability theory and their relations to classical test theory and structural equation modeling. *Psychological Methods*, *23*(1), 1–26. https://doi.org/10.1037/met0000107

Vispoel, W. P., Xu, G., & Kilinc, M. (2021). Expanding G-theory models to incorporate congeneric relationships: Illustrations using the Big Five Inventory. *Journal of Personality Assessment*, *103*(4), 429–442. https://doi.org/10.1080/00223891.2020.1808474

Wang, T., & Merkle, E. C. (2018). merDeriv: Derivative computations for linear mixed effects models with application to robust standard errors. *Journal of Statistical Software*, *87*(1), 1–16. https://doi.org/10.18637/jss.v087.c01

Ye, W., Scherer, R., & Blömeke, S. (2022). SEAS in TIMSS 2019. (Project available from the Open Science Framework (OSF)). https://doi.org/10.17605/OSF.IO/YB9HV

Ye, W., Scherer, R., & Blömeke, S. (2024). Teachers' and principals' perceptions of school emphasis on academic success: Measurement invariance, agreement, and relations to student achievement. *Large-Scale Assessments in Education*, *12*(19), 1–19. https://doi.org/10.1186/s40536-024-00207-w

Zinbarg, R. E., Yovel, I., Revelle, W., & McDonald, R. P. (2006). Estimating generalizability to a latent variable common to all of a scale's indicators: A comparison of estimators for ω_h. *Applied Psychological Measurement*, *30*(2), 121–144. https://doi.org/10.1177/0146621605278814

How to cite this article: Jorgensen, T. D., & Jak, S. (2026). A taxonomy of reliability indices for cluster-level constructs measured by multiple individual-level indicators. *British Journal of Mathematical and Statistical Psychology*, *79*, 513–546. https://doi.org/10.1111/bmsp.70074