Psychometrika (2026), **00**, 1–5
doi:10.1017/psy.2026.10116

PSYCHOMETRIC SOCIETY

THEORY & METHODS – COMMENTARY ON FOCUS ARTICLE

Conceptual Issues in CSEM Estimation and Interpretation: Commentary on Pfadt et al.

Won-Chan Lee

The University of Iowa, USA
Email: won-chan-lee@uiowa.edu

(Received 20 February 2026 UTC; revised 21 April 2026 UTC; accepted 27 April 2026 UTC)

Abstract

Pfadt et al. present accessible methods for estimating conditional standard errors of measurement (CSEMs) and implement them in the open-source software JASP. Their emphasis on individual-level precision represents an important contribution to applied measurement practice. This commentary discusses several conceptual issues that clarify and extend the authors' treatment of dimensionality, error definition, and score interpretation. The aim is to strengthen alignment between CSEM estimation and the interpretive purposes for which test scores are used.

Keywords: classical test theory; conditional standard error of measurement; generalizability theory; item response theory; reliability

Pfadt et al. (2026) present several methods for estimating conditional measurement precision, expressed as conditional standard errors of measurement (CSEMs), in a clear and accessible nontechnical format. The paper appropriately emphasizes the value of CSEMs relative to conventional reliability coefficients and overall SEM, which function primarily as group-level statistics, and it provides practical implementation of these methods in the open-access software package JASP (JASP Team, 2024).

I strongly concur with the authors' focus on individual-level measurement error. In many applied contexts, the central question is not how reliable a test is overall, but how uncertain a particular individual's score may be. Practitioners need to know how precisely a given score estimates an individual's true score. This perspective echoes Cronbach's (2004, p. 413) reflection on coefficient alpha, emphasizing the importance of reporting measurement error rather than relying solely on reliability coefficients.

The implementation of multiple CSEM methods within both classical test theory (CTT) and item response theory (IRT) frameworks in JASP is a valuable contribution, particularly given the accessibility of an open-source platform. While I view the overall contribution very positively, I offer three comments intended to clarify and extend several conceptual issues raised by the paper: (a) unidimensionality, subdomains, and composite scores in CSEM applications; (b) the distinction between relative and absolute error, and its implications for norm-referenced versus criterion-referenced interpretations, from a generalizability theory (GT) perspective; and (c) score metric in the interpretation and reporting of CSEMs.

© The Author(s), 2026. Published by Cambridge University Press on behalf of Psychometric Society.
This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (https://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/psy.2026.10116 Published online by Cambridge University Press

1. Unidimensionality, subdomains, and composite scores in CSEM applications

Pfadt et al. primarily illustrate CSEM methods in a single-domain context and emphasize unidimensionality. While this focus is appropriate for exposition, many psychological instruments are organized around multiple construct domains, where the score of interest is often a composite score formed from subdomain scores. Clarifying how CSEM methods operate across single- and multidomain contexts would strengthen the paper's practical guidance. In this section, *total scores* refer to scores obtained by summing item-level responses within a single domain, whereas *composite scores* refer to scores formed by combining subdomain-level scores across multiple domains.

Items are often grouped into subdomains reflecting distinct content areas or psychological constructs. The generalized Lord method discussed by Pfadt et al. accommodates such situations by grouping items into multiple "parts." These differ conceptually from the split-half parts used in the Thorndike method or the multiple parts in the Feldt method, which are better interpreted as parallel parts reflecting the idea of replication for reliability estimation. In contrast, the "parts" in the generalized Lord framework represent substantively distinct subdomains defined by content or construct differences.

When composite scores are formed from multiple subdomains, CSEMs can be obtained by aggregating domain-level error variances. Assuming uncorrelated domain errors, the composite CSEM is given by the square root of the weighted sum of subdomain error variances. If a reliability coefficient is desired, stratified alpha (Rajaratnam et al., 1965) provides a natural extension of coefficient alpha for multidomain instruments. These procedures illustrate that CTT-based approaches readily accommodate multidomain structures at the composite-score level.

The treatment of dimensionality, however, differs across measurement frameworks. In CTT, reliability is defined with respect to the observed total score and does not require a single latent dimension, even within a single domain. A set of items intended to measure one domain may still reflect multiple underlying processes, and CTT accommodates this by operating at the level of observed score decomposition ($X = T + E$). Reliability reflects the ratio of true-score to observed-score variance rather than assumptions about latent structure. Coefficient alpha likewise depends on covariance assumptions rather than unidimensionality. Consequently, even when a test is conceptually defined as a single domain, it may exhibit multidimensionality while still yielding a meaningful and reliable total score, provided the score represents a coherent and interpretable construct. Multidomain cases introduce an additional layer, for which the procedures described above for composite scores can be applied.

In IRT, by contrast, dimensionality is built directly into the model. Items within a given domain or scale are typically assumed to measure a single latent trait, and violations of this assumption affect model fit and parameter estimation. In multidomain settings, composite scores are, therefore, modeled either through multidimensional IRT or by combining results from separate unidimensional calibrations at the reporting stage. Thus, unlike CTT, IRT requires explicit specification of dimensional structure, both within and across domains.

It is also useful to distinguish multidomain structure from multidimensionality in a latent-variable sense. The discussion above concerns multidomain cases defined by content or subscales, where composite scores are formed across domains. By contrast, multidimensionality refers to the presence of multiple underlying latent traits. Pfadt et al. include unidimensional IRT methods, in which conditional precision is defined with respect to a single latent trait. In more general multidimensional settings, measurement precision is defined with respect to a vector of latent traits, and conditional precision is represented by a matrix rather than a single variance. In practice, these trait-level precision estimates are often integrated and expressed on the scale of summed scores. When viewed in this way, precision depends not only on the information within each dimension but also on the relationships among dimensions. Although such extensions are beyond the scope of the approaches discussed by Pfadt et al., this distinction highlights that multidimensional structure is conceptually distinct from multidomain structure and may require different approaches to CSEM estimation.

Against this background, the paper's suggestion that CTT-based CSEM methods require approximate unidimensionality warrants clarification. CTT, as originally formulated, does not assume

https://doi.org/10.1017/psy.2026.10116 Published online by Cambridge University Press

unidimensionality; as Cronbach (2004) noted, he explicitly rejected the assumption that all items measure a single type of individual difference. In IRT, violations of dimensionality assumptions lead to model misspecification, whereas in CTT, multidimensionality mainly affects how total scores are interpreted and how stable they are, rather than violating the theory itself.

For users of the authors' software, clearer distinctions among (a) single-domain scores, (b) multidomain composite scores, and (c) multidimensional modeling approaches would improve both conceptual clarity and practical guidance.

2. Absolute versus relative error: A generalizability theory perspective

Pfadt et al. correctly note that reliability does not directly indicate whether one individual's true score differs from another's or exceeds a cut score. This observation highlights an important distinction between norm-referenced and criterion-referenced interpretations. In norm-referenced settings, scores are interpreted relative to other individuals. For example, whether Mary scored higher than John or ranks in the top 20% of a group. In criterion-referenced settings, however, the focus is on classification relative to a fixed standard, such as whether an examinee passes a test based on a cut score. In such cases, the decision is not about relative standing but about whether an individual meets a predefined criterion.

These two types of interpretations involve different notions of measurement error. Norm-referenced interpretations depend primarily on the consistency of rank ordering across individuals, whereas criterion-referenced interpretations depend on the consistency of classification decisions relative to a fixed threshold.

Most CTT reliability coefficients are tied to relative decisions and reflect how well individuals can be distinguished within a population. This type of error is often referred to as *relative error*, as it affects comparisons among individuals. Criterion-referenced testing, however, requires attention to *absolute error*, which reflects how precisely an individual's observed score estimates their true score relative to a fixed standard. From this perspective, computing CSEMs directly from conventional CTT reliability coefficients may be of limited relevance, because the resulting error estimates primarily reflect inconsistency in relative rank ordering rather than the absolute precision required for classification decisions. This distinction is also reflected in classification-based indices, such as classification consistency and accuracy, which directly evaluate the stability of decisions relative to a cut score and therefore depend on absolute, rather than relative, measurement error.

While not the primary focus of Pfadt et al., GT provides a natural framework for clarifying this distinction by separating relative and absolute error variances. Absolute CSEMs derived from this framework are directly relevant for criterion-referenced interpretations. Notably, Lord's (1955, 1957) CSEM is closely related to absolute error in GT; in a simple $p \times i$ design, it is equivalent to the absolute CSEM derived from a GT formulation. Historically, because absolute error variance is larger than relative error variance, Lord's CSEM was sometimes criticized as an overestimate, and correction methods such as that proposed by Keats (1957) were preferred. From a GT perspective, however, this criticism largely reflects a failure to distinguish between absolute and relative error definitions rather than a deficiency of Lord's approach itself. Accordingly, Lord's CSEMs are conceptually more appropriate for criterion-referenced interpretations than relative CSEMs.

The central issue, therefore, is whether the definition of error matches the intended interpretation of scores. Relative CSEMs may be appropriate for norm-referenced interpretations, whereas criterion-referenced decisions require an absolute-error perspective.

3. Score scale and interpretation of CSEMs

A related issue concerns the score scale on which measurement precision is expressed. Reliability coefficients are scale-free, whereas SEMs and CSEMs are expressed on reported score scales. Although a scale-free index is often viewed as advantageous, it can obscure information that users ultimately need when interpreting scores.

https://doi.org/10.1017/psy.2026.10116 Published online by Cambridge University Press

The practical value of CSEMs lies in their interpretability at the individual level. When reported scores are linearly transformed from raw summed scores (e.g., standardized scores), CSEMs can be transformed straightforwardly by multiplying by the slope of the transformation. Many psychological instruments and educational testing programs, however, employ nonlinear transformations (e.g., percentile ranks) as final reported scales. In these cases, equal differences on the reporting scale do not correspond to equal differences in the raw-score scale; for example, a one-point increase in percentile rank may reflect very different changes in raw scores depending on where it occurs along the raw-score-to-percentile-rank transformation.

Under such nonlinear transformations, the relationship between raw-score precision and reporting-scale precision varies across the score range, and transforming raw-score CSEMs requires more careful treatment. As a result, the pattern of CSEMs over the score range may differ substantially depending on the reporting scale.

Given Pfadt et al.'s emphasis on practical implementation, clearer guidance on transforming and reporting CSEMs across score scales would further enhance the utility of the proposed methods. A broader discussion of these issues is provided in Lee and Harris (2025).

4. Concluding remarks

Pfadt et al. make an important contribution by emphasizing conditional measurement precision and implementing CSEM methods in an accessible, open-source platform. Their focus on individual-level uncertainty represents a meaningful step beyond reliance on global reliability indices.

The comments offered here highlight three areas that could further strengthen the discussion: clearer distinctions between single- and multidomain applications, explicit consideration of relative versus absolute error in norm- versus criterion-referenced contexts, and greater attention to score scale in the interpretation and reporting of CSEMs. Together, these points underscore that the usefulness of CSEMs depends not only on computational methods but also on alignment among error definitions, score structure, and intended score use.

Overall, the work by Pfadt et al. provides a strong foundation for broader adoption of CSEMs, and continued integration of conceptual clarity with practical implementation will further enhance their role in psychological and educational measurement.

Data Availability Statement. No new data were collected or analyzed for this commentary. Therefore, no data are available.

Funding statement. This research received no specific grant funding from any funding agency, commercial, or not-for-profit sectors.

Competing interests. The author declares none.

Disclosure of artificial intelligence (AI) use. The author used ChatGPT-5.3 solely for language editing purposes, including checking grammar, improving clarity, and enhancing readability of the manuscript. The tool was not used for generation of scientific content. The author takes full responsibility for the content of this manuscript.

References

Cronbach, L. J. (2004). My current thoughts on coefficient alpha and successor procedures. *Educational and Psychological Measurement*, *64*, 391–418.

JASP Team. (2024). *JASP* (Version 0.19.3) [Computer software]. https://jasp-stats.org.

Keats, J. A. (1957). Estimation of error variances of test scores. *Psychometrika*, *22*, 29–41.

Lee, W., & Harris, D. J. (2025). Reliability in educational measurement. In L. L. Cook & M. J. Pitoniak (Eds.), *Educational measurement* (5th ed., pp. 277–381). Oxford University Press.

Lord, F. M. (1955). Estimating test reliability. *Educational and Psychological Measurement*, *15*, 325–336.

Lord, F. M. (1957). Do tests of the same length have the same standard error of measurement? *Educational and Psychological Measurement*, *17*, 510–521.

https://doi.org/10.1017/psy.2026.10116 Published online by Cambridge University Press

Pfadt, J. M., Molenaar, D., Hurks, P., & Sijtsma, K. (2026). A tutorial on estimating the precision of individual test scores for anyone constructing and using psychological tests. *Psychometrika*, *2026*, 1–35. https://doi.org/10.1017/psy.2026.10081.
Rajaratnam, N., Cronbach, L. J., & Gleser, G. C. (1965). Generalizability of stratified-parallel tests. *Psychometrika*, *30*, 39–56.

Cite this article: Lee, W.-C. (2026). Conceptual Issues in CSEM Estimation and Interpretation: Commentary on Pfadt et al. *Psychometrika*, 1–5. https://doi.org/10.1017/psy.2026.10116

https://doi.org/10.1017/psy.2026.10116 Published online by Cambridge University Press