Psychometrika (2026), **00**, 1–4
doi:10.1017/psy.2026.10122

PSYCHOMETRIC SOCIETY

THEORY & METHODS – COMMENTARY ON FOCUS ARTICLE

User-Friendly Software and Estimated Conditional Standard Errors of Measurement. A Commentary on Pfadt et al.

L. Andries van der Ark

Research Institute of Child Development and Education, University of Amsterdam, Netherlands
Email: l.a.vanderark@uva.nl

(Received 19 March 2026 UTC; revised 30 April 2026 UTC; accepted 1 May 2026 UTC)

Abstract

In this commentary on Pfadt et al. (2026, *Psychometrika*, 1–35), I first make the case for implementing psychometric methods, such as the conditional standard error of measurement (CSEM), in software that is user-friendly from a practitioner's perspective. Furthermore, I argue that bias and variance in CSEM estimates are still poorly understood and I report a small simulation study comparing the coverage rates of the CSEM estimate recommended by Pfadt et al. with those of the estimated (unconditional) standard error of measurement. The results point to possible directions for future research on CSEM estimation.

Keywords: conditional standard error of measurement; JASP; standard error of measurement; statistical software

1. User-friendly software

Over the past decades, rapid advances in psychometrics—which, in my view, primarily serves psychology and education—have widened the gap between psychometrics and these fields, making dissemination one of its most important tasks. So, I welcome the accessible tutorial in *Psychometrika* on the conditional standard error of measurement (CSEM) by Pfadt et al. (2026), and the CSEM estimates provided in JASP. Nowadays, most authors implement newly developed psychometric methods in R or Python, which may suggest that advocating their implementation in user-friendly software is no longer necessary. However, I disagree and argue that the notion of "user-friendly software" requires further qualification, and that making psychometric methods available in software such as JASP, alongside R and Python, merits continued support.

The majority of users of psychometric methods—particularly those outside academia—are unfamiliar with R and Python. In my institution, approximately 500 students graduate each year, of whom only about 30—those pursuing academic careers—have strong R skills. The remaining 470 follow content-oriented programs, often including clinical training, leaving no time to become proficient in R. Yet, many will apply psychometrics in practice. They will assess children in forensic youth care based on test results, monitor the progress of special-needs students, or conduct small-scale studies within the organization they work. For psychometric methods to have a societal impact, it is not sufficient that software is user-friendly for psychometricians, it must also be user-friendly for practitioners.

JASP may be on its way to becoming user-friendly for practitioners (see also Shepard & Richardson, 2024). Based on our institution's recent transition from SPSS to JASP, I learned that students find JASP

© 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.2026.10122 Published online by Cambridge University Press

intuitive. JASP also offers three key advantages over SPSS. First, as free software, JASP remains accessible after graduation, unlike SPSS, whose high licensing costs often prevent continued use. Second, JASP keeps pace with developments in psychometrics, as illustrated by the availability of CSEMs. Third, JASP provides R functions that enable further investigation of its methods through simulation studies, potentially improving the software or stimulating new research.

2. Bias and sampling variability in estimated CSEMs

In their Figure 2, Pfadt et al. illustrated the use of the *unconditional* standard error of measurement (SEM) and CSEMs. Its key message, in my view, is that individual test scores should be accompanied by error bars. However, whether these error bars should be based on CSEMs estimated using the ANOVA method (from here on referred to as $\widehat{\text{CSEM}}_{\text{ANOVA}}$), as advocated by Pfadt et al., remains open to debate, because bias and variance in CSEM estimates are not yet well understood and may obscure the additional accuracy they promise.

The bias of the estimated SEM in a sample ($\widehat{\text{SEM}}$) can be approximated by

$$\text{Bias}\left[\widehat{\text{SEM}}\right] \approx -\frac{\sigma_X}{2\sqrt{1-\rho}}\,\text{Bias}\left[\widehat{\rho}\right] \tag{1}$$

(Van der Ark et al., in press). As test-score standard deviation σ_X serves as a scaling parameter, $\text{Bias}\left[\widehat{\text{SEM}}\right]$ depends on reliability ρand the bias of its sample estimate $\text{Bias}\left[\widehat{\rho}\right]$. For lower bound reliability estimates, such as Cronbach's alpha, $\text{Bias}\left[\widehat{\text{SEM}}\right]$ is positive, yielding conservative SEM estimates. For example, with $\sigma_X = 1$, $\rho = 0.9$ and $\text{Bias}\left[\widehat{\rho}\right] = -0.05$, we obtain $\text{SEM} = 1\sqrt{1-0.9} \approx 0.316$, whereas $\mathbb{E}\left[\widehat{\text{SEM}}\right] \approx -\frac{1}{2\sqrt{1-0.9}}\left(-0.05\right) \approx 0.395$. As CSEM estimates also depend on ρ and $\text{Bias}\left[\widehat{\rho}\right]$, they are likely to be biased as well.

In addition to bias in the reliability estimator, binning (i.e., combining adjacent test-score groups to obtain sufficient sample sizes per group) may introduce additional bias. Binning is generally required in the ANOVA, Feldt, and Thorndike methods. Pfadt et al. recommended using groups of at least 25 respondents; in JASP, the argument `caseMin` controls the minimum group size. I examined the effect of `caseMin` on CSEM estimates using the ADD data—a dataset available in JASP containing the scores of 1,162 respondents on nine polytomous items, used by Pfadt et al. to illustrate estimated SEM and CSEMs. Figure 1 shows that `caseMin` affects both the absolute magnitude of $\widehat{\text{CSEM}}_{\text{ANOVA}}$ and its relative ordering across test scores. This was also observed for the Feldt and Thorndike methods (not shown). Furthermore, for the ANOVA method setting `caseMin` to its maximum did not yield the estimated SEM, whereas it did for the Feldt method.

Under a multinomial sampling scheme, Van der Ark et al. (in press) derived a standard error for $\widehat{\text{SEM}}$, and simulations showed that these standard errors are relatively small for $N > 200$. For the ADD data, $\widehat{\text{SEM}} = 2.182$, $\text{SE}\left(\widehat{\text{SEM}}\right) \approx 0.024$, and the relative standard error— $\widehat{\text{SEM}}/\text{SE}\left(\widehat{\text{SEM}}\right)$—is approximately 1.1%, indicating that sampling variability has little effect on $\widehat{\text{SEM}}$. However, the effect of sampling variability on estimated CSEMs is yet unknown.

In a short simulation study, I examined whether using $\widehat{\text{CSEM}}_{\text{ANOVA}}$, as recommended by Pfadt et al., improves individual-level decisions by comparing the coverage rates of 95% confidence intervals based on (i) the estimated SEM and (ii) $\widehat{\text{CSEM}}_{\text{ANOVA}}$. To mimic realistic psychological test data, a three-dimensional graded response model estimated from the ADD data using the R package *mirt* (Chalmers, 2012) was used as the data-generating model, which was the best fitting model using the BIC as a criterion. For seven true-score values, computed from the data-generating model (see Table 1), coverage rates based on 10,000 replications were computed using JASP's R functions for SEM and CSEM estimation. Details of the true-score computation and accompanying R code of the simulation study are available on Open Science Framework at https://osf.io/xyfsp.

https://doi.org/10.1017/psy.2026.10122 Published online by Cambridge University Press

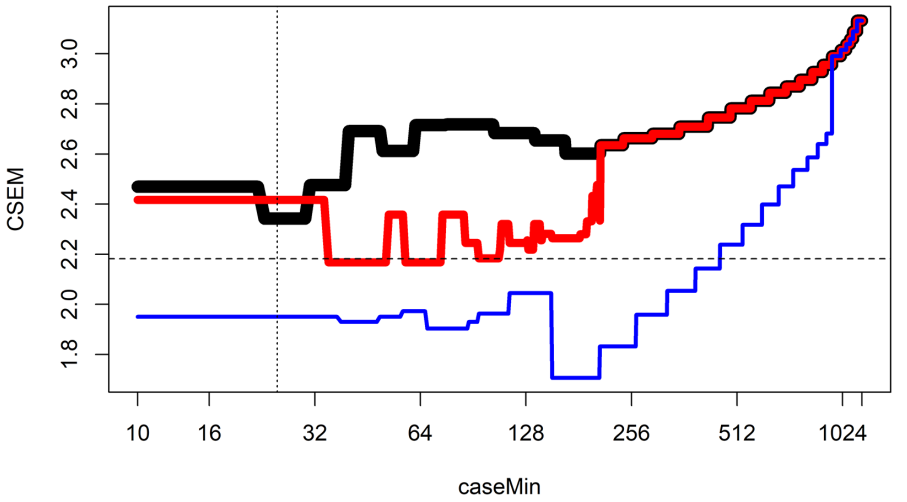


Figure 1. The estimated CSEM using the ANOVA method for test scores 6 (blue, smallest linewidth), 18 (red), and 34 (black, largest linewidth) on the ADD-data as a function of the minimum sample size used for binning (`caseMin`) expressed on a log2-scale. The dashed horizontal line represents the estimated SEM and the vertical dotted line represents `caseMin` = 25.

Table 1. Coverages rates for the 95% confidence intervals based on $\widehat{\text{SEM}}$ and $\widehat{\text{CSEM}}_{\text{ANOVA}}$, for seven true-score values

Estimator	True score						
	0.1152	0.6945	3.6952	11.2736	19.2848	26.6288	31.6043
$\widehat{\text{SEM}}$	**1.000**	**0.999**	**0.979**	0.945	0.945	**0.966**	**0.988**
$\widehat{\text{CSEM}}_{\text{ANOVA}}$	**0.592**	**0.750**	**0.898**	0.949	**0.967**	**0.973**	**0.989**

Note: T, true score; $\widehat{\text{SEM}}$, estimated standard error of measurement; $\widehat{\text{CSEM}}_{\text{ANOVA}}$, conditional standard errors estimated using the ANOVA-method. Coverage rates in bold fall outside the 95% Agresti–Coull confidence interval. Coverage rates are based on 10,000 replications.

Except for two true scores, confidence intervals based on the $\widehat{\text{SEM}}$ showed overcoverage, which means these confidence intervals were too large—they contained the true score more often than claimed. This can be expected because $\widehat{\text{SEM}}$ is positively biased (too large) if reliability is underestimated (Equation (1)). For the four highest true-scores, the confidence intervals based on $\widehat{\text{CSEM}}_{\text{ANOVA}}$ showed coverage rates similar to those of the $\widehat{\text{SEM}}$ but severe undercoverage for the three lowest true scores, which means that the confidence intervals were too optimistic—they contained the true score less often than claimed. Hence, in this study, $\widehat{\text{SEM}}$ provided better coverage than $\widehat{\text{CSEM}}_{\text{ANOVA}}$, which is not in line with the suggestion that CSEMs are more recommended. The simulation is too small to draw any conclusions about the appropriateness of using $\widehat{\text{CSEM}}_{\text{ANOVA}}$, but I hope its results and the considerations with respect to bias and variance may provide directions for future research.

3. Discussion

I have argued that the implementation of psychometric methods in software that is truly accessible to practitioners, such as JASP, is extremely important. In addition, awareness that individual test scores should have error bars is arguably one of the key insights for individual diagnostics. Therefore, the implementation of these error bars in JASP and the accompanying tutorial on CSEMs by Pfadt et al. is an important step forward in making such error bars accessible to practitioners. I then showed that,

https://doi.org/10.1017/psy.2026.10122 Published online by Cambridge University Press

in practice, CSEMs may be affected by estimation bias, which raises the question of whether error bars based on the recommended CSEMs are actually more accurate than those based on the more familiar unconditional SEM. Future research will need to address this question.

Acknowledgements. The author thanks Romke Rouw and Niels Smits for commenting on a previous version of this manuscript.

Data availability statement. All computer code is available from OSF and accessible via https://osf.io/xyfsp.

Competing interests. The author declares none.

Funding statement. This research received no specific grant funding form any funding agency, commercial, or not-for-profit sectors. Open access funding provided by University of Amsterdam.

References

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06.

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*, 1–35. https://doi.org/10.1017/psy.2026.10081.

Shepard, M. A., & Richardson, E. J. (2024). Opting for open-source? A review of free statistical software programs. *Teaching Statistics*, *46*, 53–63. https://doi.org/10.1111/test.12360.

Van der Ark, L. A., Emons, W. H. M., Ellis, J. L., & Sijtsma, K. (in press). Bias and precision in true-score estimation. British Journal of Mathematical and Statistical Psychology. https://doi.org/10.1111/bmsp.70058.

Cite this article: van der Ark, L. A. (2026). User-Friendly Software and Estimated Conditional Standard Errors of Measurement. A Commentary on Pfadt et al.. *Psychometrika*, 1–4. https://doi.org/10.1017/psy.2026.10122

https://doi.org/10.1017/psy.2026.10122 Published online by Cambridge University Press