

Test Statistics for Confirmatory Factor Analysis Under Non-Normality and Large p/n Ratios: A Comment on Muda and Yangxuan (2026)

Njål Foldnes^a and Steffen Grønneberg^b

^aUniversity of Stavanger; ^bBI Norwegian Business School

ABSTRACT

Muda and Yangxuan recommended a ridge-calibrated test statistic for controlling Type I error inflation in confirmatory factor analysis under non-normality. Although improving finite-sample inference is an important goal, their empirical evaluation is undermined by incorrectly implemented benchmark statistics, misattributed comparison methods, and a simulation design restricted to asymptotically robust conditions. Using corrected code, we reanalyze the original conditions and replace the misattributed benchmarks with the penalized eigenvalue procedures $pEBA4_{RLS}$ and $pOLS2_{RLS}$. Under the original conditions, the best-performing procedures are the ridge-calibrated T_{CsFCr} statistic and the penalized eigenvalue methods. However, under additional non-asymptotically robust conditions, including high-dimensional models, $pEBA4_{RLS}$ provides the most reliable Type I error control and outperforms the ridge-calibrated competitors.

KEYWORDS

Asymptotic robustness; CFA; non-normality; Satorra-Bentler; test statistics; Type I error

1. Introduction

The development of test statistics for Confirmatory Factor Analysis (CFA) under non-normal conditions remains a highly active area of quantitative research (see Foldnes et al., 2026 and references therein).

Recently, Muda and Yangxuan (2026) introduced a ridge-calibrated test statistic, T_{FCr} , designed to correct Type I error inflation when the ratio of variables to sample size (p/n) is relatively large. To demonstrate the superiority of their proposed method, the authors conducted a Monte Carlo simulation comparing T_{FCr} against several established robust baselines, claiming a reduction in Type I errors.

However, a closer inspection of their methodology, numerical illustrations, and theoretical assumptions reveals several fundamental errors. In this comment, we demonstrate that the baseline statistics used for comparison were calculated incorrectly. We also address anomalies in the simulation design and literature attribution that further undermine the validity of the original study. We extend the simulation conditions in Muda and Yangxuan (2026) in terms of distributions and dimensionality, and show that the apparent superiority of the proposed ridge-statistics disappears.

2. Incorrect Calculation of the RLS Test Statistic and the Satorra-Bentler Scaling Factor

Muda and Yangxuan (2026) compare ridge-calibrated statistics to the reweighted least squares (RLS) statistic T_{RLS} (Browne, 1974) and to the Satorra-Bentler (SB) scaled chi-

square T_{SB} (Satorra & Bentler, 1994), and report severe Type I error inflation for these baselines. For instance, they report a Type I error rate of 0.172 for the SB statistic and 0.222 for the RLS statistic (Muda & Yangxuan, 2026, Table 2).

However, both statistics were incorrectly implemented in the code provided by the authors. The RLS statistic computed by Muda and Yangxuan (2026) on line 134 is actually $(n - 1)T_{ML}/n$, where T_{ML} is the standard maximum likelihood statistic. So although the authors give the correct formula in the article, its implementation in the simulation code is incorrect.

The formula for the Satorra-Bentler test statistic T_{SB} is given correctly in the article, with the scaling factor being $C_s = r(UG)/df$. However, this is not what is implemented in the simulation code (line 49), where the scaling is typically given as $1 + 0.05 \cdot \text{kurt}$, where kurt is the average marginal kurtosis.

The miscalculation of the scaling factor also carries over to the proposed ridge-calibrated statistics in Muda and Yangxuan (2026) that rely on C_s , namely T_{CsF} , T_{CsFC} , and T_{CsFCr} . In our proposed corrected code these statistics are calculated with the correct C_s .

3. Misattribution of Test Statistics

Beyond the statistics T_{RLS} and T_{SB} , Muda and Yangxuan (2026) benchmark their method against two novel statistics they label “Foldnes F1” and “Foldnes F2.” The authors attribute these statistics to a 2024 paper by Foldnes and Grønneberg, purportedly published in the *British Journal of Mathematical and Statistical Psychology*. However, a search

Table 1. Type I error control for original AR and non-AR conditions.

Condition	T_{ML}	T_{RLS}	T_{SB}	$pEBA4_{RLS}$	$pOLS2_{RLS}$	T_{CSF}	T_{CSF_C}	$T_{CSF_C_J}$	T_F	T_{F_C}	$T_{F_C_J}$
Panel A: original AR conditions (C1–C12, omitting C5)											
C1	0.249	0.063	0.280	0.041	0.043	0.154	0.166	0.048	0.135	0.145	0.038
C2	0.254	0.062	0.285	0.041	0.043	0.153	0.165	0.046	0.132	0.143	0.036
C3	0.249	0.055	0.279	0.036	0.038	0.149	0.161	0.043	0.128	0.140	0.035
C4	0.250	0.057	0.278	0.037	0.038	0.149	0.163	0.043	0.128	0.141	0.036
C6	0.252	0.059	0.281	0.040	0.042	0.154	0.166	0.047	0.133	0.143	0.037
C7	0.249	0.061	0.279	0.040	0.041	0.150	0.160	0.047	0.129	0.139	0.038
C8	0.249	0.060	0.281	0.040	0.041	0.151	0.162	0.046	0.130	0.140	0.037
C9	0.251	0.059	0.279	0.038	0.040	0.153	0.164	0.046	0.130	0.143	0.039
C10	0.249	0.058	0.279	0.035	0.037	0.148	0.160	0.043	0.128	0.139	0.035
C11	0.252	0.059	0.280	0.034	0.036	0.152	0.165	0.042	0.133	0.146	0.036
C12	0.247	0.059	0.276	0.037	0.039	0.151	0.162	0.044	0.129	0.139	0.036
Panel B: non-AR conditions											
VM ₁	0.816	0.658	0.326	0.043	0.042	0.179	0.193	0.052	0.729	0.742	0.556
PL ₁	0.895	0.801	0.323	0.041	0.038	0.174	0.188	0.049	0.844	0.852	0.726
IG ₁	0.411	0.192	0.316	0.038	0.037	0.177	0.191	0.054	0.282	0.296	0.136
VM ₂	0.923	0.858	0.426	0.084	0.074	0.261	0.277	0.090	0.890	0.896	0.811
PL ₂	0.972	0.951	0.511	0.097	0.079	0.359	0.376	0.161	0.962	0.964	0.940
IG ₂	0.540	0.322	0.380	0.055	0.051	0.227	0.243	0.083	0.426	0.439	0.261
VITA	0.890	0.795	0.390	0.041	0.038	0.238	0.254	0.082	0.845	0.853	0.741

Notes: C1–C12 (omitting C5) are the original distributional conditions. IG: independent generator; VM: Vale-Maurelli; PL: piecewise linear; Subscript 1: moderate non-normality; subscript 2: severe non-normality; VITA: Vine copula with Clayton pair-copulae and $\chi(df = 1.2)$ margins. Values in bold satisfy $.025 \leq \alpha \leq .075$. Nominal $\alpha = .05$.

Table 2. Type I error control across seven distributions and three p/n ratios.

Distribution	T_{ML}	T_{RLS}	T_{SB}	$pEBA4_{RLS}$	$pOLS2_{RLS}$	T_{CSF}	T_{CSF_C}	$T_{CSF_C_J}$	T_F	T_{F_C}	$T_{F_C_J}$
Panel A: $p/n = 0.125$ ($p = 100, n = 800$)											
Norm	0.758	0.060	0.779	0.022	0.003	0.676	0.677	0.000	0.657	0.657	0.000
VM ₁	1.000	1.000	0.720	0.034	0.002	0.616	0.616	0.000	1.000	1.000	0.996
PL ₁	1.000	1.000	0.686	0.027	0.001	0.576	0.576	0.000	1.000	1.000	1.000
IG ₁	0.930	0.330	0.768	0.027	0.001	0.682	0.682	0.001	0.886	0.886	0.013
VM ₂	1.000	1.000	0.531	0.125	0.010	0.435	0.435	0.000	1.000	1.000	1.000
PL ₂	1.000	1.000	0.563	0.094	0.003	0.464	0.464	0.000	1.000	1.000	1.000
IG ₂	0.982	0.759	0.810	0.051	0.003	0.727	0.727	0.002	0.974	0.974	0.273
Panel B: $p/n = 0.150$ ($p = 60, n = 400$)											
Norm	0.547	0.050	0.574	0.020	0.013	0.438	0.441	0.005	0.407	0.410	0.005
VM ₁	0.999	0.979	0.568	0.036	0.010	0.429	0.433	0.006	0.999	0.999	0.832
PL ₁	1.000	1.000	0.534	0.026	0.007	0.396	0.401	0.002	1.000	1.000	0.984
IG ₁	0.817	0.322	0.612	0.028	0.008	0.469	0.473	0.009	0.745	0.746	0.090
VM ₂	1.000	1.000	0.575	0.106	0.031	0.446	0.450	0.008	1.000	1.000	1.000
PL ₂	1.000	1.000	0.566	0.089	0.019	0.431	0.432	0.005	1.000	1.000	1.000
IG ₂	0.927	0.658	0.657	0.066	0.019	0.552	0.556	0.020	0.894	0.895	0.396
Panel C: $p/n = 0.250$ ($p = 100, n = 400$)											
Norm	0.999	0.059	0.999	0.025	0.002	0.996	0.996	0.089	0.995	0.995	0.067
VM ₁	1.000	1.000	0.998	0.052	0.001	0.989	0.989	0.037	1.000	1.000	1.000
PL ₁	1.000	1.000	0.997	0.051	0.000	0.986	0.986	0.033	1.000	1.000	1.000
IG ₁	0.999	0.397	0.999	0.027	0.000	0.997	0.997	0.105	0.997	0.997	0.395
VM ₂	1.000	1.000	0.977	0.263	0.021	0.943	0.943	0.017	1.000	1.000	1.000
PL ₂	1.000	1.000	0.978	0.196	0.009	0.943	0.943	0.020	1.000	1.000	1.000
IG ₂	1.000	0.724	1.000	0.081	0.001	0.997	0.997	0.210	1.000	1.000	0.743

Notes: Nominal $\alpha = .05$. VM: Vale-Maurelli; PL: Piecewise Linear; IG: Independent Generator. Subscript 1: moderate non-normality ($s = 2, k = 7$); subscript 2: severe non-normality ($s = 3, k = 21$). Values in bold satisfy $(.025 \leq \alpha \leq .075)$.

of the literature confirms that no such article exists in that journal's records, nor do these statistics appear in actual published work by Foldnes and Grønneberg. These statistics (line 155 and 163) are calculated by dividing T_{ML} by sample statistics derived from the marginal skewness and kurtosis of the sample.

4. Simulation Issues

An important limitation in the simulation design employed by Muda and Yangxuan (2026) is that it imposes conditions of asymptotic robustness (AR) by generating independent, non-normally distributed common factors alongside mutually independent, normally distributed errors. Statistical

theory (Amemiya & Anderson, 1990; Browne & Shapiro, 1988) then guarantees the asymptotic validity of T_{ML} and T_{RLS} statistics. We confirmed this by simulating at sample size $N = 300,000$, where T_{ML} approaches the nominal $\alpha = 0.05$.

The simulation is additionally flawed for several non-normal conditions, which generated latent factors that were standardized using the sample-based `scale()` function in R, rather than standardizing via theoretical population moments. By standardizing using the empirical sample mean and sample variance (e.g., line 307), the generated observations within a sample are no longer independent and identically distributed. Furthermore, the scaling forces the sample variance of the factors to be exactly 1.0 in every

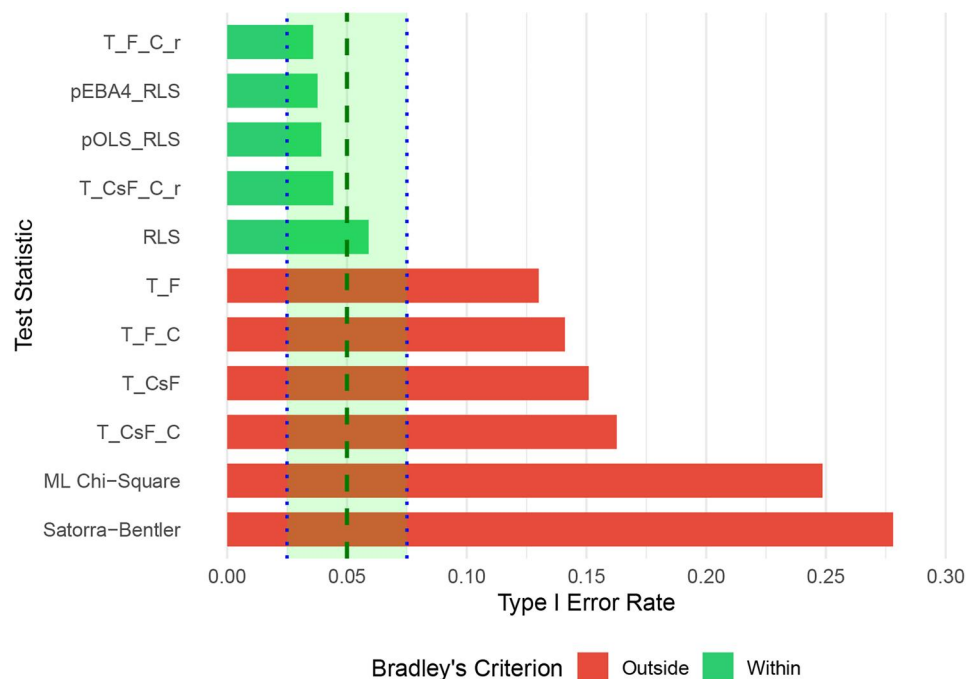


Figure 1. Type I error control aggregated over all original AR conditions.

single replication, removing the natural sampling variability from the sample covariance matrix.

5. Method

In order to correct the above shortcomings, we modified the code supplied by Muda and Yangxuan (2026). First, we corrected the flawed implementation of T_{SB} and T_{ML} . Second, we replaced the “Foldnes F1” and “Foldnes F2” test statistic by the new test statistics $pEBA4_{RLS}$ and $pOLS2_{RLS}$ proposed by Foldnes et al. (2025). These test statistics are available in the R package *semTests* (Moss, 2024). We also corrected the empirical scaling by using the exact theoretical scalings in several of the simulation conditions. Since the C3 and C5 conditions in Muda and Yangxuan (2026) were identical in terms of covariance matrix, we removed the C5 condition. The original and corrected code is available at <https://github.com/njaalf/mudacomment>.

We extended the simulation design to also include non-AR conditions with the same setup of nine model and sample size conditions as used by Muda and Yangxuan (2026). Non-AR was imposed by simulating data on the observed variables directly, using the following seven distributions: Vale-Maurelli (Vale & Maurelli, 1983) with moderate and severe non-normality; piecewise linear (Foldnes & Grønneberg, 2022) with moderate and severe non-normality; and independent generator (Foldnes & Olsson, 2016) with moderate and severe non-normality. Moderate non-normality was represented by a skewness of 2 and kurtosis of 7, while severe non-normality was defined by a skewness of 3 and kurtosis of 21. The last distribution (VITA) was a regular vine distribution (Grønneberg & Foldnes, 2017) based on Clayton pair-copulas and a chi-square distribution with 1.2 degrees of freedom on the marginals, which implies

marginal skewness and kurtosis equal to 2.58 and 10, respectively.

In addition, we investigated three high-dimensional non-AR conditions adapted from Foldnes et al. (2026), to assess whether the findings drawn from the low p ($p \leq 24$) conditions examined by Muda and Yangxuan (2026) were generalizable to higher-dimensional conditions. Specifically we investigated a five-factor CFA in the following (p, n) conditions: $p/n = .125 : p = 100, n = 800$, $p/n = .15 : p = 60, n = 400$, and $p/n = .25 : p = 100, n = 400$. In each (p, n) condition we generated data from seven distributions: normal, Vale-Maurelli with moderate and severe non-normality, piecewise linear with moderate and severe non-normality, and independent generator with moderate and severe non-normality.

Data generation was done using package *covsim* (Grønneberg et al., 2022). While Muda and Yangxuan (2026) used 1,000 replications, we use 5,000 replications in their original model and sample size conditions. For the three high-dimensional conditions we use 2,000 replications.

6. Results

6.1. The Original Conditions Where AR Holds

Let us first consider overall Type I error rate, aggregated over all simulation conditions studied by Muda and Yangxuan (2026). Figure 1 shows the same information as in Muda and Yangxuan (2026, Figure 1), though with the corrections described above. The differences are striking. First, Muda and Yangxuan (2026) reported a RLS Type I error rate of .22, while when properly implemented the error rate is 0.06, which is remarkably close to the nominal rate, indicating that the simulations are done under conditions that are not too far away from normality. Second, the

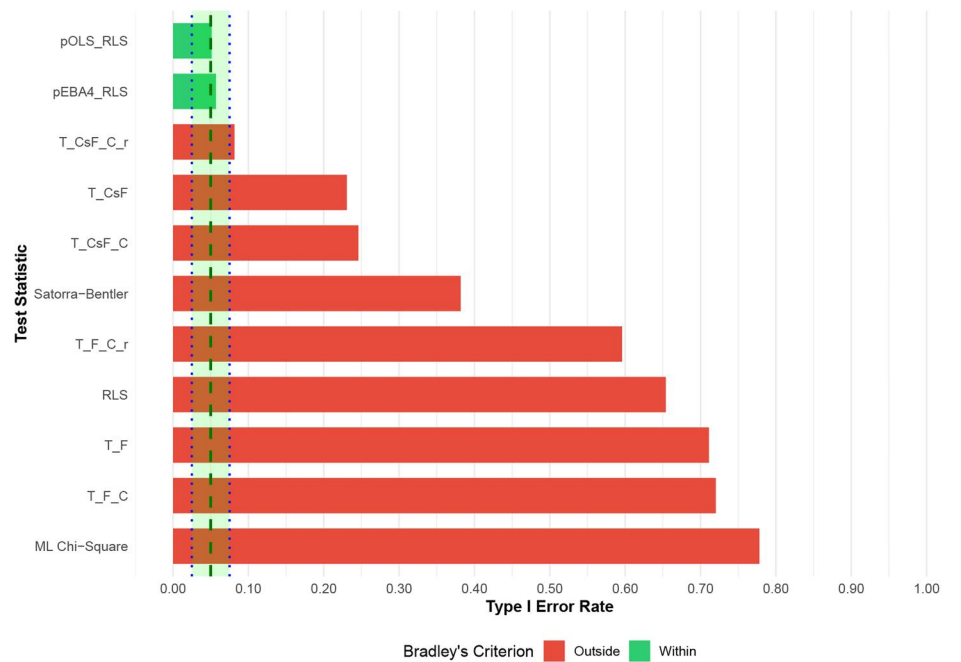


Figure 2. Type I error control aggregated over non-AR conditions.

correct implementation of the scaling factor C_s produces Type I error rates for T_{SB} , T_{CsF} , T_{CsF_C} and $T_{CsF_C_r}$ quite different from what is reported in Muda and Yangxuan (2026). For instance, $T_{CsF_C_r}$ when properly implemented has a Type I error of 0.045, compared to 0.02 in Muda and Yangxuan (2026), which we deem the best performing test statistic under AR conditions. The two runners-up are pEBA4_{RLS} and pOLS2_{RLS}: although their distance to the nominal level is comparable to that of T_{RLS} , they err on the conservative side whereas T_{RLS} over-rejects, inflating Type I error above the nominal level.

Panel A of Table 1 presents the Type I error rates per distributional condition, aggregated over the nine model and sample size configurations. For each test statistic, we observe minimal variation across distributional conditions, indicating that the selected scenarios do not capture much diversity in terms of test-performance.

6.2. Non-AR Conditions

Due to the remarkably low levels of variation observed in Panel A of Table 1, we also changed the distributions of the simulation configurations, keeping the model-configurations otherwise fixed. The aggregated (across model and sample size configurations) Type I errors per distributional class are shown in Panel B of Table 1. Fully aggregated Type I errors are shown in Figure 2.

Following the model configurations in Muda and Yangxuan (2026), but now violating the AR assumption, several shifts in performance emerge. First, RLS performs poorly, yielding a rejection rate of .65. Second, while pOLS2_{RLS} and pEBA4_{RLS} exhibit aggregated Type I error rates near nominal levels, this warrants caution: Panel B of Table 1 reveals a pattern of simultaneous over- and under-rejection that cancels out in the aggregate. Nonetheless,

pOLS2_{RLS} and pEBA4_{RLS} maintain a higher count of acceptable conditions according to Bradley's criterion than the ridge-based statistics.

6.3. New High-Dimensional Conditions

Rejection rates in the three high-dimensional model conditions, for each distribution, are given in Table 2. Under normality, T_{RLS} achieves remarkable Type I error control in all three panels, echoing earlier findings that T_{RLS} is clearly superior to T_{ML} in small-sample normal settings. Under non-normality T_{RLS} has inflated rejection rates, while the best performance is exhibited by pEBA4_{RLS}. Both pOLS2_{RLS} and $T_{CsF_C_r}$ markedly under-reject, while the remaining ridge tests severely over-reject.

For each test statistic, we observe that increasing skewness and kurtosis increases the rejection rates. Also, the VM distribution is associated with higher rejection rates than the IG distribution.

7. Discussion and Conclusion

The development of accurate test statistics for CFA with small samples is a valuable pursuit. However, new methodological proposals must be evaluated against correctly calculated and accurately sourced baselines. Because the findings in Muda and Yangxuan (2026) rely on incorrect baseline calculations for standard estimators (SB and RLS), and were only based on asymptotic robust conditions, their conclusions cannot be trusted.

Based on our corrected implementation of their simulation design, the recommendation to adopt the T_{FCr} statistic was not supported. We extended the simulation conditions of Muda and Yangxuan (2026) and found that the penalized eigenvalue method pEBA4_{RLS} proposed by Foldnes et al.

(2025) best controlled Type I error rates under non-asymptotic robust conditions, which we believe is the most realistic scenario in real-world applications of CFA.

References

- Amemiya, Y., & Anderson, T. W. (1990). Asymptotic chi-square tests for a large class of factor analysis models. *The Annals of Statistics*, 18, 1453–1476. <https://doi.org/10.1214/aos/1176347760>
- Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal*, 8, 1–24. https://doi.org/10.10520/AJA0038271X_175
- Browne, M. W., & Shapiro, A. (1988). Robustness of normal theory methods in the analysis of linear latent variate models. *British Journal of Mathematical and Statistical Psychology*, 41, 193–208. <https://doi.org/10.1111/j.2044-8317.1988.tb00896.x>
- Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Sage.
- Vale, C. D., & Maurelli, V. A. (1983). Simulating multi-variate nonnormal distributions. *Psychometrika*, 48, 465–471. <https://doi.org/10.1007/bf02293687>
- Muda, N., & Yangxuan, S. (2026). A bias-corrected test statistic for confirmatory factor analysis under non-normality. *Structural Equation Modeling: A Multidisciplinary Journal*. Advance online publication. <https://doi.org/10.1080/10705511.2025.2606043>
- Foldnes, N., & Grønneberg, S. (2022). Non-normal data simulation using piecewise linear transforms. *Structural Equation Modeling: A Multidisciplinary Journal*, 29, 36–46. <https://doi.org/10.1080/10705511.2021.1949323>
- Foldnes, N., Grønneberg, S., & Moss, J. (2026). Penalized eigenvalue block averaging: Extension to nested model comparison and Monte Carlo evaluations. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-026-02968-4>
- Foldnes, N., Moss, J., & Grønneberg, S. (2025). Improved goodness of fit procedures for structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal*, 32, 1–13. <https://doi.org/10.1080/10705511.2024.2372028>
- Foldnes, N., & Olsson, U. H. (2016). A simple simulation technique for nonnormal data with prespecified skewness, kurtosis, and covariance matrix. *Multivariate Behavioral Research*, 51, 207–219. <https://doi.org/10.1080/00273171.2015.1133274>
- Grønneberg, S., & Foldnes, N. (2017). Covariance model simulation using regular vines. *Psychometrika*, 82, 1035–1051. <https://doi.org/10.1007/s11336-017-9569-6>
- Grønneberg, S., Foldnes, N., & Marcoulides, K. M. (2022). covsim: An R package for simulating non-normal data for structural equation models using copulas. *Journal of Statistical Software*, 102, 1–45. <https://doi.org/10.18637/jss.v102.i03>
- Moss, J. (2024). *semTests: Goodness-of-fit testing for structural equation models* [R package version 0.6.0]. <https://cran.r-project.org/web/packages/semTests/>