

Field Trial Sample Size for Scales (FiTS³): Requirements for Questionnaire Scale Analysis

Andrés Christiansen

Ana María Mejía-Rodríguez

Rolf Strietholt

Abstract

Besides assessing student achievement, IEA studies collect valuable information through context questionnaires. These questionnaires are evaluated during a field trial (FT), intended to test the functioning of context scales in each country and to provide empirical evidence to revise questionnaire material. However, the usefulness of the FT depends on the quality of the evidence it can provide, and a key issue in this regard is the sample size per country.

While previous studies have explored the topic of minimum sample size for scale construction and evaluation, such studies are theoretical or done with simulated data. Whether these findings apply to real data and, particularly, IEA studies remains unclear. To investigate this matter, we used existing student and teacher questionnaire data from TIMSS 2019 and ICILS 2018, with available data for up to 65 educational systems and 70 scales.

We analyzed how different sample size minimums influence the distribution and precision of the indicators obtained from a confirmatory factor analysis with real data scenarios. We focused on how different goodness-of-fit indices (CFI, TLI, RMSEA, SRMR) change between sample size minimums and how this may impact scale adequacy decisions based on common rule-of-thumb thresholds.

Our analysis confirms that bias decreases as sample size increases and supports the $N \geq 200$ guideline, where bias remains minimal. However, our findings also indicate that samples of $N \geq 100$ observations may still provide reliable evidence, with only a slightly higher bias compared to having 200 observations. Therefore, while $N \geq 200$ remains the preferred threshold, sample sizes in the range of 100-200 observations may be a practical and valid alternative in certain contexts.

1 Introduction

This report provides insights and recommendations to the IEA and its technical expert group on sample size requirements for field trial (FT) analyses. The findings are based on the Research and Development Fund project *Field Trial Sample Size for Scales (FiTS³): Requirements for Questionnaire Scale Analysis*.

The overarching goal of IEA studies is to describe educational opportunities and outcomes, such as student achievement and attitudes, and to analyze factors that influence these outcomes, especially those that can be manipulated to improve education systems worldwide. Alongside student achievement, the IEA studies collect non-cognitive data about attitudes and other contextual factors by administering questionnaires to students, teachers, school principals and, in some cases, parents and specialized school staff. This project focuses on data from these questionnaires, particularly responses summarized into measurement scales which are designed to provide more reliable and valid information about the topics under study than analyzing individual item responses (Martin, Rust, and Adams 1999).

The performance of questionnaire scales is evaluated at two different stages: during the FT phase and the main survey (MS) phase. It should be noted that an FT and MS have entirely different purposes; an MS aims to estimate population parameters with sufficient precision, while an FT focuses on testing the functioning of survey instruments, and the scales derived from them, to guide the selection of MS material based on empirical evidence. The usefulness of the FT analysis depends on the quality of the evidence it can provide, and a key issue in this regard is the sample size per country. While the MS samples are large-scale, covering approximately 150 schools and 4000 students per country, the FT samples are significantly smaller. For context, the FT of the

We would like to thank Leslie Rutkowski for their valuable feedback and insightful comments on an earlier draft of this paper. Here suggestions have helped to refine our analysis and improve the clarity of our arguments.

International Computer and Information Literacy Study (ICILS) from 2023 had sample sizes ranging from 229 to 1043 students (divided into two questionnaire forms), 51 to 610 teachers (also divided into two questionnaire forms), and 6 to 42 school principals.

The technical reports of IEA studies outline sample size requirements based on power analyses to ensure sufficient statistical power (e.g., Martin, Rust, and Adams 1999; Fraillon et al. 2020; Martin, Davier, and Mullis 2020; Davier et al. 2023). In contrast, the sample sizes for FT are less thoroughly documented, particularly regarding the necessary sample size required to evaluate questionnaire scales. According to the technical standards for IEA studies, FT sample sizes can be much smaller than in the MS, provided they are still large enough to provide helpful statistics for item selection and refinement ahead of the MS (Martin, Rust, and Adams 1999). However, which statistics are helpful and what constitutes a “large enough” sample size remains unclear.

The technical standards for IEA studies do not explicitly prescribe a minimum sample size for trustworthy item statistics but hint at one through an example from the FT of Trends in International Mathematics and Science Study (TIMSS) of 1999, which aimed for at least 200 student observations for each test and questionnaire item. The standards further describe that, although the school sample size can be smaller and selected with a less rigorous procedure, procedures for within-school samples should follow the MS procedures as closely as possible as this “will ensure that the test-item characteristics, the questionnaire-item characteristics, and the operations procedures are realistically evaluated in all participating countries.” (Martin, Rust, and Adams 1999, 47). However, unlike the guidelines for MS, which are based on power analyses, there is no information explaining why the chosen number is 200 and not, for instance, 100 or 500. Furthermore, while, generally speaking, the chosen number does ensure that the student questionnaire is realistically evaluated, having much smaller school sample sizes may pose significant challenges for achieving realistic and accurate evaluations of teacher and, particularly, school questionnaires.

In efforts to have precise measures, sample size minimums have been established in recent FT scaling analyses for the most recent ICILS cycle. Following the mentioned reference to TIMSS 1999 and in line with general guidelines commonly found in the literature, a minimum sample size of 200 was initially set for all respondent groups (i.e., students, teachers, and principals). However, the guidelines found in the literature are somewhat inconsistent and do not always specify precisely 200 as the required number (e.g., Kline 2016; Hoogland and Boomsma 1998; Bandalos 2014; Forero, Maydeu-Olivares, and Gallardo-Pujol 2009; Wolf et al. 2013). Furthermore, defining a clear threshold posed challenges, as not only do some countries have too small FT sample sizes to begin with, but there is also the use of different questionnaire forms (a common practice in FTs due to the aim of evaluating extensive questionnaire material and sometimes conduct experiments testing alternative versions of a question). Moreover, there is always the potential for high rates of non-response. All this combined not only limits the scope of the FT analysis but could also influence the conclusions derived from it, underscoring the need for careful consideration of FT sample size requirements.

1.1 Sample size and questionnaire scaling

Sample size is a critical factor in research, including when using latent variable modeling techniques like confirmatory factor analysis (CFA), commonly used for scaling questionnaire data in IEA studies. CFA is often called a “large sample technique” (Gagne and Hancock 2006), reflecting its reliance on sufficient sample sizes to ensure precision, replicability, and reliable parameter estimates. Some issues of insufficient sample sizes are that they can lead to convergence failures, improper solutions, bias in parameter estimates and their standard errors, and lack of statistical power to detect model misfits (Koran 2020).

During an FT scale evaluation, the model fit of CFA models is routinely evaluated using approximate goodness-of-fit indices (AFI, McNeish, An, and Hancock 2017). It is known that each AFI behaves differently on a range of other factors, such as the number of items, the type of items (continuous or ordinal), the distribution of responses, the selected estimator, and, importantly, the sample size (e.g., Brosseau-Liard and Savalei 2014; Chen et al. 2008; Fan and Sivo 2007; Laar and Braeken 2021).

Despite widespread agreement on the importance of sample size, determining an appropriate—or “large enough”—sample size has been historically challenging. A substantial body of research has focused on providing applied researchers with specific sample size guidelines for CFA (e.g., Gagne and Hancock 2006; Kyriazos 2018; MacCallum et al. 1999; Tanaka 1987). However, these guidelines are often varied and sometimes even conflicting. Unfortunately, as Jackson (2003) observes, no precise number can be substituted for the phrase “large enough.” Increasingly, researchers emphasize that appropriate sample sizes depend on multiple factors, such as model complexity, the number of indicators per factor, and data characteristics (Kyriazos 2018).

Kyriazos (2018) provides a comprehensive account of the various guidelines or “rules of thumb” proposed in the CFA literature. These guidelines can generally be categorized into unconditional rules that involve establishing an absolute minimum sample size and conditional rules that depend on model characteristics, such as the number of variables or estimated parameters. Kyriazos further notes that these “strict” rules have gradually been supplanted by guidelines derived from Monte Carlo simulations, which offer more tailored recommendations based on specific modeling conditions. However, despite these advancements, no consensus has emerged, and opinions on appropriate sample size thresholds continue to differ. Some approaches also incorporate power analysis to determine sample size. For a detailed overview of these methods and their evolution, see Kyriazos (2018).

1.1.1 Unconditional guidelines: absolute N

Sample size recommendations for CFA often begin with unconditional guidelines that provide an absolute minimum number of observations. Some researchers recommend a minimum of 200 cases (Singh, Junnarkar, and Kaur 2016; Comrey 1988; Kline 2016), while others suggest at least 300 cases for more complex models (Tabachnick and Fidell 2013). In the context of multi-group CFA, a standard guideline is to have at least 100 cases per group (Kline 2016; Wang and Wang 2012). Comrey and Lee (2013) propose a general rating for sample size minimums: 50 = very poor, 100 = poor, 200 = fair, 300 = good, 500 = very good, and 1000 = excellent. While useful as a starting point, these unconditional guidelines have significant limitations, as they do not consider the CFA model’s complexity or the study’s specific objectives.

1.1.2 Conditional guidelines: Ratio of number of observations to number of variables ($N:p$)

Recognizing the limitations of unconditional recommendations, researchers have proposed conditional guidelines, such as sample size ratios (N) to the number of measured variables (p). These ratios acknowledge the influence of model complexity on sample size needs. Nunnally and Bernstein (1993) and Gorsuch (2014), as cited in Dimitrov (2014), recommend an $N:p$ ratio of at least 5. Everitt (1975) suggests a stricter requirement of $N:p \geq 10$, while Tinsley and Tinsley (1987) offer a range between 5 and 10, particularly when the sample size exceeds 300. For unidimensional scale development, DeVellis and Thorpe (2021) recommends a minimum of 300 cases for a 20-item scale. While these ratios account for the number of measured variables, they still do not address other important factors, such as the estimator used, the type of data, or the desired precision of parameter estimates.

1.1.3 Conditional guidelines: Ratio of number of observations to number of estimated parameters ($N:q$)

Another set of conditional guidelines focuses on the sample size (N) ratio to the number of estimated parameters (q). These recommendations provide more tailored guidance by directly linking sample size to model complexity. For example, Bentler and Chou (1987) and Bollen (1989) recommend an $N:q$ ratio of 5 to 10, while Hoogland and Boomsma (1998) advocate for ratios greater than 10 for data with high kurtosis. Jackson (2003) suggests a minimum ratio of 20, with 10 as the absolute lower limit. These guidelines reflect that more complex models with more estimated parameters require larger sample sizes for reliable results. However, as with $N:p$ ratios, these recommendations depend on the specific model characteristics and evaluation criteria.

1.1.4 Evidence from simulation studies

Simulation studies provide empirical insights into sample size requirements, highlighting the interplay between model characteristics, data properties, and study objectives. These studies typically adopt one of three strategies: (1) setting a minimum sample size based on the estimator used, (2) establishing a minimum ratio between the sample size and the number of parameters or items, or (3) determining sample size based on desired statistical power. For example, Anderson and Gerbing (1988) and Ding, Velicer, and Harlow (1995) suggest a minimum of 100 to 150 cases, while others propose ranges like 100-200 cases MacCallum et al. (1999) or at least 200 cases for most models (Hoogland and Boomsma 1998; Boomsma and Hoogland 2001). Simulations using ordinal CFA models, as applied in IEA Hamburg’s scaling procedures, loosely point to a minimum of 200 cases (Bandalos 2014; Forero, Maydeu-Olivares, and Gallardo-Pujol 2009; Wolf et al. 2013).

Despite their utility, simulation studies have limitations. They often examine narrow conditions, such as continuous CFA models with homogeneous loadings or simplified sample structures. These constraints mean the

findings may not generalize to all contexts, particularly when studying categorical data or using more complex models. Consequently, while simulation studies provide valuable empirical evidence, their results must be interpreted within the specific conditions they address.

While unconditional guidelines offer a baseline for determining sample size, they are limited by their generalizability. Conditional guidelines, such as $N:p$ and $N:q$ ratios, provide more tailored recommendations by incorporating model complexity but still depend on study-specific factors. Simulation studies offer empirical support, yet their findings often reflect narrow conditions and cannot be universally applied. Ultimately, the sample size required for CFA depends on two key considerations: (a) the specific characteristics of the model, including data type, estimator, and complexity, and (b) the evaluation criteria, such as desired power, precision of parameter estimates, and fit indices. Given these complexities, it is difficult to establish universal rules that work across all conditions and purposes. Instead, researchers must carefully consider their study's unique requirements when determining an appropriate sample size.

1.2 What is “large enough” sample size for FTs in IEA studies?

Given the diversity of recommendations and the limitations of existing studies, determining an appropriate sample size in the context of IEA studies remains a complex challenge. While simulation studies provide valuable insights, they also have notable limitations, as even the most comprehensive studies tend to consider only a small subset of models encountered in practice (Tanaka 1987), let alone the real-world data encountered in IEA studies and, in particular, in FT phases.

Opting for more flexible approaches, such as fixing a minimum ratio based on several parameters (or items) or making sample size requirements scale-dependent by fixing a desired statistical power, may not be feasible in the context of IEA studies. These approaches complicate result clarity and may hinder the effective communication of findings in large-scale studies. When choosing an approach, it is essential to consider that FT analysis results must be communicated to diverse stakeholders, including national research coordinators and questionnaire development experts. Clear and effective communication is critical. With dozens of scales evaluated, selecting a ratio between sample size and parameters or fixing a statistical power, although possible, will mean using a scale-dependent solution, i.e., different rules for different scales, resulting in an unclear rule to communicate for national research coordinators and other stakeholders.

A more practical approach would involve establishing a uniform minimum sample size guideline for all scales. The $N \geq 200$ recommendation from simulation studies based on ordinal CFA models should be interpreted as a general guideline rather than as a definitive rule to be used in the analysis of IEA data, as these studies still do not resemble the wide variety of scales or items found in IEA questionnaires. Specifically, they fail to coincide with a) the number of items per scale, b) the number of response categories, and c) the distribution of responses. Furthermore, this guideline has posed challenges in recent FT analyses conducted by the IEA. As previously mentioned, some countries cannot meet this requirement by design. Depending on the amount of questionnaire material tested, the number of questionnaire forms used, and the extent of item non-response, some countries may be excluded from the analysis with a $N \geq 200$ requirement. This requirement represents a problem for smaller respondent groups like school principals or, in some instances, teachers.

1.3 Towards a practical, evidence-based solution

Therefore, exploring the impact of different sample size minimums on scale evaluation in simulations with a small range of parameters and using actual IEA data is crucial. Could the rule of $N \geq 200$ be relaxed without compromising the quality of the results? If so, what would be a minimum sample size that would still yield realistic and accurate scale evaluations? Or is such a sample too small to provide meaningful evidence? These questions lie at the heart of the FiTS³ project.

The main goal of FiTS³ is to analyze the effects of selecting different sample size minimums on the accuracy of approximate goodness-of-fit indices, or AFIs, and on overall scale quality assessment, which guides decisions to retain, modify (e.g., by improving or removing items), or drop scales ahead of the MS stage. While larger sample sizes might ensure more robust evaluations, smaller sample size requirements, if supported by evidence, could reduce the burden on data collection and analysis.

Drawing inspiration from the IEA standard for developing a sampling plan, we believe the aim should be to achieve samples that are as small as possible and as cost-effective as possible while maintaining an acceptable level of accuracy (Martin, Rust, and Adams 1999). The findings from this project will inform future FT designs and provide evidence-based recommendations for optimizing sample size requirements for scaling analyses.

2 Method

2.1 Data

The main goal of our study was to evaluate the impact of different sample sizes on the performance of questionnaire scales. Traditionally, this evaluation would be conducted via a simulation study. However, we aimed to analyze scenarios using real data from IEA studies. Specifically, we used data from ICILS 2018 and TIMSS 2019 student and teacher questionnaires. The data included 65 countries, 578953 students, and 69864 teachers (see Table 1). From these data, we selected all unidimensional scales with more than three items, comprising 70 scales (43 student scales and 27 teacher scales).

Table 1: Number of participants and countries

Study	Countries	Students		Teachers	
		N	Scales	N	Scales
ICILS 2018	13	43709	11	24295	13
TIMSS 2019 G4	56	300622	9	20713	5
TIMSS 2019 G8	37	234622	23	24856	9

2.2 Sample size and sample selection

We tested 14 different sample sizes ($N = 25, 50, 75, 100, 125, 150, 175, 200, 225, 250, 300, 350, 400, 500$). For each sample size, we selected 500 random samples. In large-scale assessments, cluster sampling is typically used instead of simple random sampling. In the first step, schools are selected, and in the second step, students and teachers are chosen. In the present study, we mimicked this approach by randomly selecting schools within each country until we reached the desired N . If the selection resulted in more cases than required, we randomly removed cases from the last selected school to achieve the exact N desired.

This procedure was designed to achieve target sample sizes (N) while respecting the two-stage clustered structure of IEA studies. However, it does not replicate the country-specific within-school sampling rules used in an actual FT. This is a necessary simplification that focuses our investigation on the effect of sample size while partially keeping the effect of the sample design.

2.3 Modeling approach

We assessed the psychometric adequacy of scales by estimating unidimensional categorical CFA models using the diagonally weighted least squares (DWLS) estimator. The modeling process was conducted in two stages: (1) estimating single-group CFA models for each country and each scale, and (2) estimating multi-group CFA (MGCFA) for each scale to evaluate its comparability across countries. All models were estimated using the lavaan library for R (Rosseel 2012).

2.3.1 Single-group study

For the single-group study, each country-scale combination was analyzed independently, resulting in a total of 1953 country-scale combinations. For each of these combinations, we first used the full country sample to calculate the true values of our models. These values were then compared to those obtained in each of the 500 random samples per sample size, resulting in a total of 13671000 comparisons.

2.3.2 Multi-group study

For the multiple-group study, we planned to use the same random samples used for the single-group analysis, pooling the data of all countries to perform a multiple-group CFA (MGCFA) on each scale. Therefore, we would only need one model per scale and sample size rather than one model per country, scale combination and sample size. However, as detailed in the results section, it was not feasible to proceed with this part of the project.

2.4 Evaluation Metrics

To evaluate scale performance, we focused on four approximate goodness-of-fit indices (AFI): the comparative fit index (CFI), the Tucker–Lewis index (TLI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). Each sample was compared to their respective true value using the following metrics: a) the distance between them; b) the change in precision by each step of N ; and c) the proportion of “correct” conclusions derived from the cut-offs points, i.e., if the same conclusion was derived from the total sample than from each random sample.

2.4.1 Distance

The distance between the true values and the estimated ones was calculated via three statistics: the mean error (ME), the mean absolute error (MAE), and the median absolute deviation (MAD). That is, for a scale s in country c we would have:

$$\text{ME}_j = \frac{\sum_{i=1}^{500} \hat{x}_{isc} - x_{sc}}{n},$$

$$\text{MAE}_j = \frac{\sum_{i=1}^{500} |\hat{x}_{isc} - x_{sc}|}{n},$$

$$\text{MAD}_j = \text{median}(|\hat{x}_{isc} - x_{sc}|),$$

where i is the number of a random sample, and j is a given sample size, e.g., 25, 50, 75, etc.

2.4.2 Change in precision

The change in precision by sample size was calculated using the difference of MAE and MAD between sample sizes, i.e.:

$$\Delta\text{MAE}_j = \text{MAE}_j - \text{MAE}_{j-1},$$

$$\Delta\text{MAD}_j = \text{MAD}_j - \text{MAD}_{j-1}.$$

2.4.3 Proportion of correct conclusions

In model evaluation, observed AFI are often compared to threshold values to determine whether a model fits the observed data. Following this idea, another method to examine accuracy is the proportion of “correct” or converging conclusions between the whole and random samples.

In the single-group study, we used two cut-off points for each index: for CFI and TLI, ≥ 0.95 indicated a good fit and ≥ 0.90 an acceptable fit; and for RMSEA and SRMR, ≤ 0.05 indicated a good fit and ≤ 0.10 an acceptable fit. Thus, we calculated the proportion of agreement between the whole sample analysis and each random sample.

In Figure 1, we illustrate the distribution of the whole sample data for each AFI, i.e., the distribution of the true values across all scales and countries. In general, CFI, TLI, and SRMR have an acceptable fit for more than 80% of the cases; however, for RMSEA, less than 40% of the cases present an acceptable fit.

In instances where AFI values exceeded 1, we converted them to a maximum value of 1.

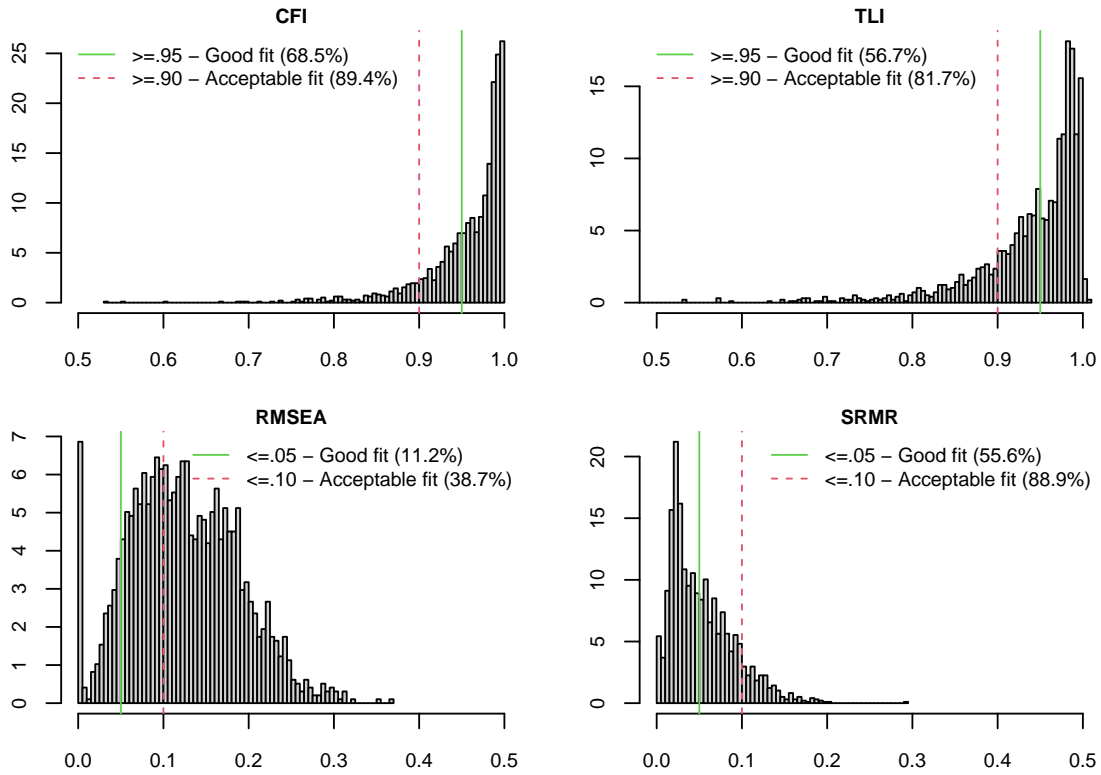


Figure 1: Distribution of AFI for the whole sample

3 Results

3.1 Single-group study

3.1.1 Distance

The average values across all estimations of the three indicators used to measure distance (or error) are presented in Table 2 and Figure 2.

The first measure, mean error (ME), is a raw error statistic. Since ME averages both negative and positive values, it does not capture the magnitude of the error but instead highlights its direction. Thus, ME is useful for identifying whether a particular AFI tends to be systematically under or overestimated at certain sample sizes. Across all AFI, SRMR exhibits the most systematic error, with a tendency to overestimate the true values. This overestimation occurs across all sample sizes tested but is more pronounced in smaller samples. Similarly, RMSEA tends to overestimate smaller sample sizes, though to a lesser extent. In contrast, TLI demonstrates a slight tendency toward underestimation in smaller sample sizes.

When considering absolute indicators of error, the mean absolute error (MAE) and the median absolute deviation (MAD) both demonstrate a consistent decline as sample size increases (see Figure 2). Notably, CFI errors are minimal even at the smallest sample size ($N = 25$). Across the four AFI, a clear stabilization pattern exists in both MAE and MAD within the 100–200 sample size range.

Additional results based on scale characteristics are available in the Appendix. We examined whether similar patterns emerged based on the number of items (4, 5, 6, 7, 8, 9, or 10+), the number of response categories (3, 4, or 5), the wording direction (fully positive, fully negative, or mixed), and the type of response categories (agreement, competence, extent, frequency, or intensity). The Appendix results align closely with the main findings, with only minor differences. Variations include small systematic under- or overestimation in fit indices. These differences are more pronounced in smaller sample sizes but do not substantially differ from the main results.

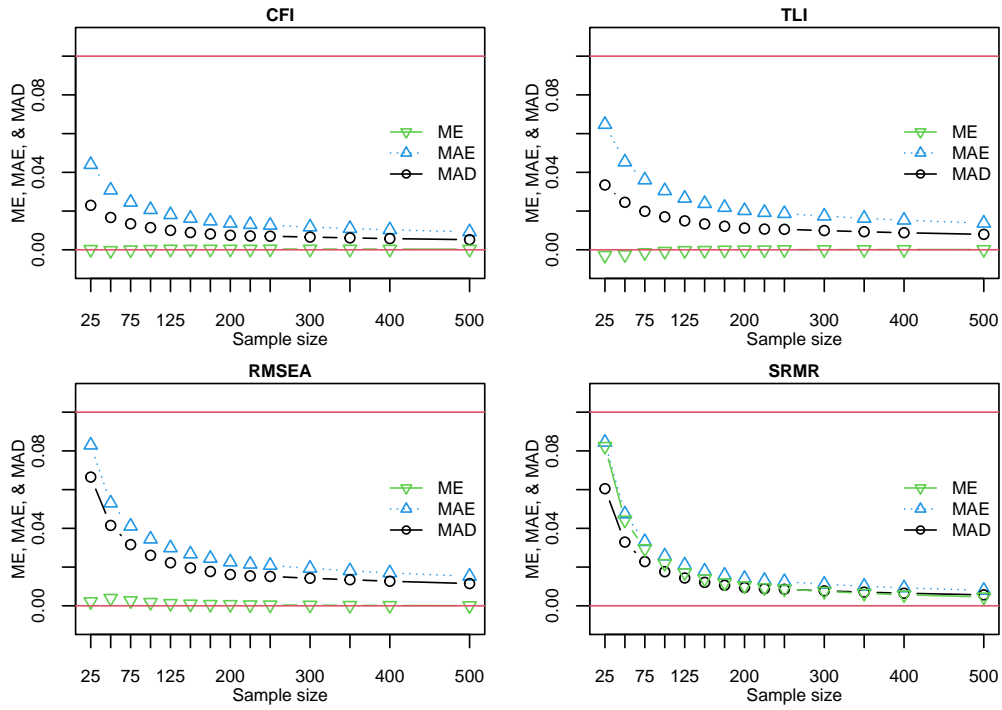


Figure 2: ME, MAE, and MAD for All scales (70)

Table 2: ME, MAE, and MAD for All scales (70)

N	CFI			TLI			RMSEA			SRMR		
	ME	MAE	MAD	ME	MAE	MAD	ME	MAE	MAD	ME	MAE	MAD
25	.0002	.0440	.0230	-.0029	.0647	.0335	.0021	.0830	.0665	.0823	.0844	.0605
50	-.0004	.0308	.0167	-.0025	.0453	.0245	.0039	.0531	.0415	.0443	.0474	.0329
75	.0000	.0245	.0134	-.0015	.0361	.0199	.0027	.0411	.0316	.0295	.0331	.0228
100	.0002	.0207	.0115	-.0008	.0305	.0170	.0018	.0344	.0261	.0219	.0257	.0176
125	.0004	.0181	.0100	-.0004	.0266	.0149	.0012	.0299	.0222	.0171	.0211	.0144
150	.0003	.0162	.0089	-.0003	.0239	.0133	.0009	.0268	.0196	.0140	.0179	.0121
175	.0003	.0149	.0082	-.0002	.0219	.0122	.0007	.0245	.0177	.0119	.0158	.0106
200	.0004	.0137	.0075	-.0001	.0202	.0112	.0007	.0226	.0162	.0103	.0141	.0094
225	.0003	.0131	.0071	-.0001	.0193	.0107	.0005	.0215	.0154	.0094	.0131	.0088
250	.0004	.0127	.0070	.0001	.0187	.0106	.0005	.0209	.0152	.0087	.0125	.0085
300	.0004	.0118	.0066	.0001	.0174	.0099	.0003	.0194	.0142	.0075	.0111	.0077
350	.0004	.0110	.0062	.0001	.0163	.0093	.0002	.0181	.0135	.0065	.0101	.0071
400	.0003	.0103	.0058	.0001	.0153	.0088	.0002	.0170	.0127	.0057	.0092	.0065
500	.0003	.0093	.0052	.0001	.0137	.0080	.0001	.0153	.0115	.0046	.0080	.0057

3.1.2 Increase in precision

Besides focusing on the distance or magnitude of the error, we can also examine the gains in precision achieved as sample size increases. Figure 3 shows that precision gains are more substantial when moving from smaller sample sizes, such as 50 to 75 or 75 to 100. However, these gains become negligible when sample sizes exceed 200. Precision gains stabilize at different sample sizes for specific AFI: around 175–200 for CFI and TLI, 100 for RMSEA, and 125–150 for SRMR.

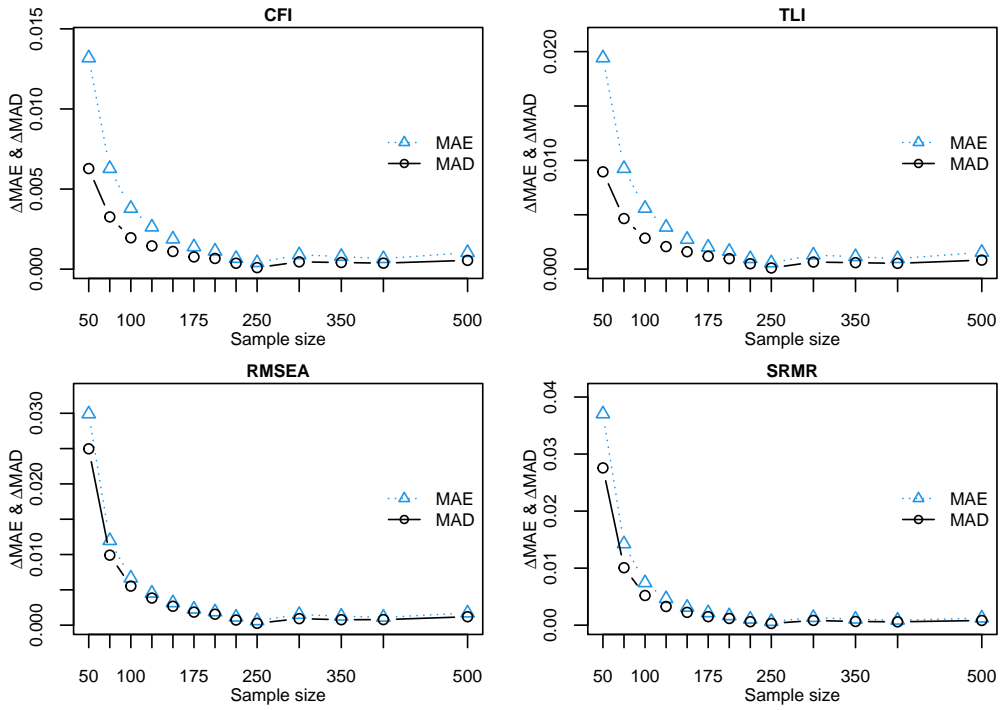


Figure 3: Δ MAD and Δ MAE for All scales (70)

3.1.3 Proportion of correct conclusions

While measuring the error is important, it is equally crucial to assess the accuracy of our evaluations of scale quality based on the cut-off points for good and acceptable model fit. The results depicted in Figure 4 offer complementary insights, showing the agreement between true and estimated values regarding scale quality.

The upper panels show the proportion of consistent conclusions, i.e., cases where the estimated and true fit evaluations align, counting both conclusions of good (or acceptable) fit and of not good (or not acceptable) fit. For acceptable fit, agreement exceeds 0.90 at around $N = 125$ for most AFI, except RMSEA, which reached this value at $N = 350$. For good fit, RMSEA reaches a high agreement (≥ 0.90) at around $N = 125$, CFI at $N = 200$, while SRMR and TLI reach this closer to $N = 400$. Nevertheless, all AFI show at least 0.80 agreement by $N = 125$.

The middle panels depict the proportion of true positives, i.e., cases where the samples accurately assessed a good or acceptable fit. For acceptable fit, true positives exceed 0.90 by $N = 100$ for all AFI, except RMSEA, which reaches 0.90 at $N = 350$ (but achieves 0.80 by $N = 125$). For good fit, true positives reach a 0.8 agreement proportion at $N = 100$ for most AFI. It is important to mention that RMSEA is more sensitive because of the distribution of its true values.

The bottom panels display the proportion of true negatives, i.e., cases where the samples accurately assessed a not good or not acceptable fit. For acceptable fit, true negatives exceed 0.80 agreement at an N of 200–250; for good fit, this threshold is reached at around $N = 125$.

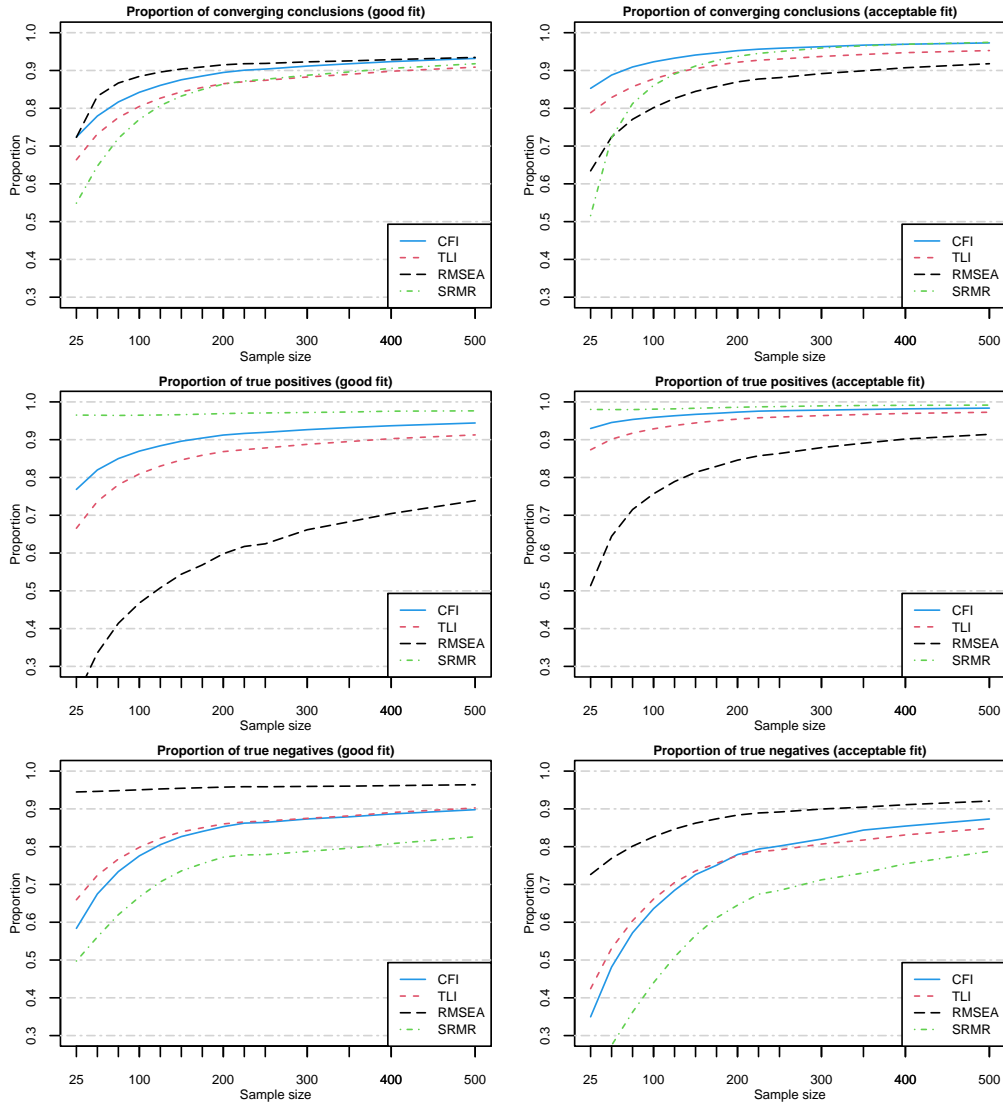


Figure 4: Proportion of converging conclusions

3.2 Multiple-group study

For the multiple-group study, our plan was to pool the sample from the single-group study and run an MGCFA per scale. This approach would, in theory, yield a total of 980 true values, resulting from the combination of N and the number of scales.

However, we faced a significant obstacle: since we are running a categorical MGCFA using the DWLS estimator, all groups (i.e., countries) in each sample must have the same number of categories for all items (Rutkowski, Svetina, and Liaw 2019). Unfortunately, not all countries met this requirement in every iteration, causing some countries to be dropped from the model. Consequently, across the 500 samples, we encountered different sets of countries in each iteration, which greatly increased the number of true values to test. As shown in Table 3, this number exceeds the already large number of combinations from the single-group study and becomes rapidly unmanageable when we decrease the sample size. For instance, in the most extreme cases, particularly with smaller sample sizes, we observed over 300 unique country-combinations after the 500 iterations. This variability made it exceedingly difficult to perform meaningful comparisons and, ultimately, rendered the analysis of unfeasible.

Table 3: Average country combinations, % of samples with all countries, and % of removed countries per N

N	True values	Average country combinations	% of samples with all countries	% removed countries
25	23600	337.14	0.90	69.6
50	24802	354.31	3.13	43.2
75	20779	296.84	9.34	29.5
100	17077	243.96	17.57	21.5
125	13646	194.94	25.29	16.5
150	10629	151.84	32.49	13.1
175	8404	120.06	41.29	10.6
200	6718	95.97	47.89	9.0
225	5396	77.09	53.15	7.8
250	4516	64.51	58.76	6.8
300	3274	46.77	67.42	4.9
350	2469	35.27	74.26	3.7
400	1817	25.96	78.65	3.1
500	1018	14.54	84.34	2.0

4 Discussion

The main goal of our project was to analyze how varying sample sizes impact the accuracy of categorical CFA model fit statistics and scale quality evaluations. We approached this from the perspective of FT questionnaire scaling in IEA studies, a crucial step that informs the selection of MS material but is constrained by small sample sizes by country. Although the technical standards for IEA studies suggest a minimum of 200 student responses per questionnaire item, achieving this for other respondents, such as teachers or principals, can be challenging. Furthermore, there is no clear rationale for the chosen sample size of 200 and not, for example, 100 or 500. While some literature supports the 200 minimum, it often does not reflect the nature of IEA data and scaling procedures. More evidence for an appropriate minimum sample size is needed, as this information can significantly impact FT scaling procedures and sampling design.

Using actual IEA data from TIMSS 2019 and ICILS 2018, we conducted a simulation-like analysis to evaluate scale performance under various sample size minimums ($N = 25, 50, 75, 100, 125, 150, 175, 200, 225, 250, 300, 350, 400, 500$). Based on standard analytical procedures employed in FT scale evaluation, our analysis aimed to include two approaches: (1) single-group analysis, in which each country-scale combination is examined independently, and (2) multi-group analysis, which assesses the level of measurement invariance across countries for each scale.

Our findings from the single-group CFA analyses show that larger sample sizes reduce errors in AFI, but precision begins to plateau after a certain threshold. We found no clear evidence to justify sample size requirements above $N = 200$, as precision gains beyond this point are minimal and may not outweigh the logistical challenges and costs associated with achieving larger FT sample sizes. Instead, our findings suggest that sample sizes between 100 and 200 may be sufficient for achieving acceptable accuracy in FT analyses. These findings remain consistent when examining results based on different scale characteristics, such as the number of items and the number and type of response categories.

The exact minimum sample size depends on factors such as the acceptable trade-off between precision and feasibility and the amount of data one is willing to exclude from the analysis, which could mean losing all available data from a single country. While $N = 200$ is a reasonable guideline for a minimum sample size, an N of 100 can still yield valuable information if sample size constraints arise. However, although we recognize the potential utility of 100 as a minimum threshold, we prefer not to establish it as a baseline recommendation due to the advantages of aiming for larger sample sizes and the influence of factors such as item non-response.

Unfortunately, the MGCFA component of our project could not be completed due to technical challenges encountered during the sample selection. Specifically, issues arose from empty cells across countries, especially in the lower sample size ranges ($N < 200$). This is a known issue when running categorical MGCFA using the DWLS estimator (Rutkowski, Svetina, and Liaw 2019; Svetina, Rutkowski, and Rutkowski 2019), which became unmanageable for the scope of our study, with a large number of countries, scales, sample size minimums, and iterations. As random samples of a given N were drawn 500 times, not only did the number of empty cells

vary across iterations, but the countries affected by these empty cells also changed, resulting in hundreds of unique country combinations. Although the issue was less pronounced at larger sample sizes, conducting the analysis solely with these larger sizes would have been uninformative, as it would not address the region of interest identified in the single-group analysis, i.e., sample sizes between 100 and 200. We considered other approaches, such as collapsing response categories (Rutkowski, Svetina, and Liaw 2019) or switching from categorical to continuous models. However, the first would still not fully solve the issue and the second would have undermined the study’s core purpose of evaluating data and models directly relevant to IEA studies. Another approach could be to design a simulation study to avoid empty cells entirely; however, such an approach may not accurately reflect the challenges encountered in real-life FT (and MS) phases. This finding, though disappointing, underscores the need for methodological innovations in handling MGCFA-based measurement invariance analyses. Future research could explore alternative methods for measurement invariance testing that are less reliant on large sample sizes and better equipped to handle diverse data structures. Moreover, future research must focus on strategies for reducing the appearance of empty cells, given that cross-country comparison is one of the most important results of a FT. These strategies may include, for example, determining a minimum N (for reducing empty cells not just for the accuracy of CFA estimators); or renaming, decreasing or collapsing Likert categories by design.

4.1 Limitations

Some limitations should be acknowledged. First, this project’s scope also required narrowing our focus to TIMSS 2019 and ICILS 2018. This decision was made to ensure depth of analysis within the available timeframe. However, we acknowledge the need to expand future research to additional IEA studies, such as PIRLS or ICCS, to validate and extend our findings. This expansion would likely pertain to scale content rather than characteristics, as many scale attributes are common across IEA studies.

Second, while we provide additional results across different scale characteristics (i.e., number of items, number of response categories, type of response categories, type of construct, direction of wording), interpretation of these results should be approached with caution due to the limited variation in scale characteristics within IEA studies. However, this limitation reflects the nature of IEA scales, which predominantly feature four response categories of an agreement or frequency type and fully positive phrasing. While broader variation would enrich methodological research, the characteristics examined here represent most IEA scales, ensuring the relevance of our findings for practical applications.

Third, our study simplified the within-school sampling design. In a FT, countries may have different rules for selecting classes (i.e., one or more classes per school). Therefore, results may be unbalanced for countries with schools that have larger samples. Future research should address this issue by improving the mimicking of a FT sample.

Finally, the inability to complete the multi-group analysis limits our understanding of how sample size influences measurement invariance evaluations. Addressing this gap will require carefully designed simulation studies or alternative methodological approaches to mitigate challenges associated with categorical MGCFA, including empty cells. Moreover, the lack of a multi-group analysis influences the cross-cultural validity of the questionnaire constructs. Consequently, the observed variance in responses could be influenced not only by actual differences in the constructs being measured but also by country-specific factors. These include nuances in conceptual understanding of the constructs within different contexts or subtle discrepancies introduced through translation.

4.2 Conclusions

This study provides valuable insights into the sample size requirements for FT analyses in IEA studies. Our results suggest that smaller sample sizes (below the current $N \geq 200$ guideline) can yield acceptable levels of accuracy while being more practical. When determining the sample size for an FT, each study must weigh increases in precision against additional administrative and financial costs. In general, based on our findings, we recommend avoiding samples below 100. However, a sample size minimum within the range of 100–200 observations per country may serve as a practical guideline to maximize the use of available data without compromising the reliability of scale evaluations.

Recent experiences from the ICILS 2023 FT indicate that even this suggested sample size range may be too high for school questionnaires. For example, ICILS 2023 FT school samples ranged from six to 56 schools per country, with an average of only 25 schools (and principals). Moreover, in studies such as TIMSS and PIRLS, teacher samples are also limited, given that they are the subject teachers of the assessed students. Therefore,

the total number of teachers would be similar to the number of sampled schools. If the goal is to analyze scales from school or teacher questionnaires to the same extent as student, i.e., using single-country CFA models and examining measurement invariance across countries, FT sampling designs would need to adapt accordingly. This would require careful consideration of the administrative and financial costs associated with obtaining larger FT school samples in relation to the benefits of achieving more accurate scale evaluations. Alternatively, pooled CFA models could be estimated, gathering information on how well the scale performs in the study's overall population, but without information about individual countries.

References

- Anderson, James C., and David W. Gerbing. 1988. "Structural Equation Modeling in Practice: A Review and Recommended Two-Step Approach." *Psychological Bulletin* 103 (3): 411–23. <https://doi.org/10.1037/0033-2909.103.3.411>.
- Bandalos, Deborah L. 2014. "Relative Performance of Categorical Diagonally Weighted Least Squares and Robust Maximum Likelihood Estimation." *Structural Equation Modeling: A Multidisciplinary Journal* 21 (1): 102–16. <https://doi.org/10.1080/10705511.2014.859510>.
- Bentler, P. M., and Chih-Ping Chou. 1987. "Practical Issues in Structural Modeling." *Sociological Methods & Research* 16 (1): 78–117. <https://doi.org/10.1177/0049124187016001004>.
- Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. Wiley. <https://doi.org/10.1002/9781118619179>.
- Boomsma, A., and J. J. Hoogland. 2001. "The Robustness of LISREL Modeling Revisited." In *Structural Equation Models: Present and Future. A Festschrift in Honor of Karl Jöreskog*, edited by R. Cudeck, S. du Toit, and D. Sörbom, 139–68. Chicago: Scientific Software International.
- Brosseau-Liard, Patricia E., and Victoria Savalei. 2014. "Adjusting Incremental Fit Indices for Nonnormality." *Multivariate Behavioral Research* 49 (5): 460–70. <https://doi.org/10.1080/00273171.2014.933697>.
- Chen, Feinian, Patrick J. Curran, Kenneth A. Bollen, James Kirby, and Pamela Paxton. 2008. "An Empirical Evaluation of the Use of Fixed Cutoff Points in RMSEA Test Statistic in Structural Equation Models." *Sociological Methods & Research* 36 (4): 462–94. <https://doi.org/10.1177/0049124108314720>.
- Comrey, Andrew L. 1988. "Factor-Analytic Methods of Scale Development in Personality and Clinical Psychology." *Journal of Consulting and Clinical Psychology* 56 (5): 754–61. <https://doi.org/10.1037/0022-006x.56.5.754>.
- Comrey, Andrew L., and Howard B. Lee. 2013. *A First Course in Factor Analysis*. Psychology Press. <https://doi.org/10.4324/9781315827506>.
- Davier, Matthias von, Ina V. S. Mullis, Bethany Fishbein, and Pierre Foy, eds. 2023. *Methods and Procedures: PIRLS 2021 Technical Report*. Boston College, TIMSS & PIRLS International Study Center.
- DeVellis, Robert F., and Carolyn T Thorpe. 2021. *Scale Development*. 5th ed. Applied Social Research Methods. Thousand Oaks, CA: SAGE Publications.
- Dimitrov, Dimitar M. 2014. *Statistical Methods for Validation of Assessment Scale Data in Counseling and Related Fields*. American Counseling Association.
- Ding, Lin, Wayne F. Velicer, and Lisa L. Harlow. 1995. "Effects of Estimation Methods, Number of Indicators Per Factor, and Improper Solutions on Structural Equation Modeling Fit Indices." *Structural Equation Modeling: A Multidisciplinary Journal* 2 (2): 119–43. <https://doi.org/10.1080/10705519509540000>.
- Everitt, B. S. 1975. "Multivariate Analysis: The Need for Data, and Other Problems." *British Journal of Psychiatry* 126 (3): 237–40. <https://doi.org/10.1192/bjp.126.3.237>.
- Fan, Xitao, and Stephen A. Sivo. 2007. "Sensitivity of Fit Indices to Model Misspecification and Model Types." *Multivariate Behavioral Research* 42 (3): 509–29. <https://doi.org/10.1080/00273170701382864>.
- Forero, Carlos G., Alberto Maydeu-Olivares, and David Gallardo-Pujol. 2009. "Factor Analysis with Ordinal Indicators: A Monte Carlo Study Comparing DWLS and ULS Estimation." *Structural Equation Modeling: A Multidisciplinary Journal* 16 (4): 625–41. <https://doi.org/10.1080/10705510903203573>.
- Fraillon, Julian, John Ainley, Wolfram Schulz, Tim Friedman, and Daniel Duckworth, eds. 2020. *IEA International Computer and Information Literacy Study 2018. Technical Report*. International Association for the Evaluation of Educational Achievement.
- Gagne, Phill, and Gregory R. Hancock. 2006. "Measurement Model Quality, Sample Size, and Solution Propriety in Confirmatory Factor Models." *Multivariate Behavioral Research* 41 (1): 65–83. https://doi.org/10.1207/s15327906mbr4101_5.
- Gorsuch, Richard L. 2014. *Factor Analysis: Classic Edition*. Routledge. <https://doi.org/10.4324/9781315735740>.
- Hoogland, Jeffrey J., and Anne Boomsma. 1998. "Robustness Studies in Covariance Structure Modeling: An Overview and a Meta-Analysis." *Sociological Methods & Research* 26 (3): 329–67. <https://doi.org/10.1177/0049124198026003003>.

- Jackson, Dennis L. 2003. "Revisiting Sample Size and Number of Parameter Estimates: Some Support for the n:q Hypothesis." *Structural Equation Modeling: A Multidisciplinary Journal* 10 (1): 128–41. https://doi.org/10.1207/s15328007sem1001_6.
- Kline, Rex B. 2016. *Principles and Practice of Structural Equation Modeling, Fourth Edition*. 4th ed. Methodology in the Social Sciences. New York, NY: Guilford Publications.
- Koran, Jennifer. 2020. "Indicators Per Factor in Confirmatory Factor Analysis: More Is Not Always Better." *Structural Equation Modeling: A Multidisciplinary Journal* 27 (5): 765–72. <https://doi.org/10.1080/10705511.2019.1706527>.
- Kyriazos, Theodoros A. 2018. "Applied Psychometrics: Sample Size and Sample Power Considerations in Factor Analysis (EFA, CFA) and SEM in General." *Psychology* 09 (08): 2207–30. <https://doi.org/10.4236/psych.2018.98126>.
- Laar, Saskia van, and Johan Braeken. 2021. "Understanding the Comparative Fit Index: It's All about the Base!" *Practical Assessment, Research, and Evaluation*. <https://doi.org/10.7275/23663996>.
- MacCallum, Robert C., Keith F. Widaman, Shaobo Zhang, and Sehee Hong. 1999. "Sample Size in Factor Analysis." *Psychological Methods* 4 (1): 84–99. <https://doi.org/10.1037/1082-989x.4.1.84>.
- Martin, Michael O., Matthias von Davier, and Ina V. S. Mullis, eds. 2020. *Methods and Procedures: TIMSS 2019 Technical Report*. Boston College, TIMSS & PIRLS International Study Center.
- Martin, Michael O., Keith Rust, and Raymond J Adams, eds. 1999. *Technical Standards for IEA Studies*. International Association for the Evaluation of Educational Achievement.
- McNeish, Daniel, Ji An, and Gregory R. Hancock. 2017. "The Thorny Relation Between Measurement Quality and Fit Index Cutoffs in Latent Variable Models." *Journal of Personality Assessment* 100 (1): 43–52. <https://doi.org/10.1080/00223891.2017.1281286>.
- Nunnally, Jum C, and Ira Bernstein. 1993. *Psychometric Theory*. 3rd ed. McGraw-Hill Series in Psychology. Maidenhead, England: McGraw Hill Higher Education.
- Rosseel, Yves. 2012. "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software* 48 (2): 1–36. <https://doi.org/10.18637/jss.v048.i02>.
- Rutkowski, Leslie, Dubravka Svetina, and Yuan-Ling Liaw. 2019. "Collapsing Categorical Variables and Measurement Invariance." *Structural Equation Modeling: A Multidisciplinary Journal* 26 (5): 790–802. <https://doi.org/10.1080/10705511.2018.1547640>.
- Singh, Kamlesh, Mohita Junnarkar, and Jasleen Kaur. 2016. *Measures of Positive Psychology: Development and Validation*. Springer India. <https://doi.org/10.1007/978-81-322-3631-3>.
- Svetina, Dubravka, Leslie Rutkowski, and David Rutkowski. 2019. "Multiple-Group Invariance with Categorical Outcomes Using Updated Guidelines: An Illustration Using Mplus and the Lavaan/semTools Packages." *Structural Equation Modeling: A Multidisciplinary Journal* 27 (1): 111–30. <https://doi.org/10.1080/10705511.2019.1602776>.
- Tabachnick, Barbara G, and Linda S Fidell. 2013. *Using Multivariate Statistics: Pearson New International Edition*. Pearson Higher Ed.
- Tanaka, J. S. 1987. "How Big Is Big Enough?: Sample Size and Goodness of Fit in Structural Equation Models with Latent Variables." *Child Development* 58 (1): 134. <https://doi.org/10.2307/1130296>.
- Tinsley, Howard E. A., and Diane J. Tinsley. 1987. "Uses of Factor Analysis in Counseling Psychology Research." *Journal of Counseling Psychology* 34 (4): 414–24. <https://doi.org/10.1037/0022-0167.34.4.414>.
- Wang, Jichuan, and Xiaoqian Wang. 2012. *Structural Equation Modeling: Applications Using Mplus*. Wiley Series in Probability and Statistics. Wiley. <https://doi.org/10.1002/9781118356258>.
- Wolf, Erika J., Kelly M. Harrington, Shaunna L. Clark, and Mark W. Miller. 2013. "Sample Size Requirements for Structural Equation Models: An Evaluation of Power, Bias, and Solution Propriety." *Educational and Psychological Measurement* 73 (6): 913–34. <https://doi.org/10.1177/0013164413495237>.