Challenging Conventional Wisdom for Multivariate Statistical Models With Small Samples

Article *in* Review of Educational Research · July 2017

DOI: 10.3102/0034654317727727

Daniel McNeish
Arizona State University

Challenging Conventional Wisdom for Multivariate Statistical Models with Small Samples

Daniel McNeish
Arizona State University

CONTACT INFORMATION:

Daniel McNeish, Department of Psychology, Arizona State University, P.O. Box 871104, Tempe, AZ 85287. Email: dmcneish@asu.edu.

Abstract

In educational research, small samples are common because of financial limitations, logistical challenges, or exploratory studies. With small samples, statistical principles upon which researchers rely do not hold, leading to trust issues with model estimates and possible replication issues when scaling up. Researchers are generally aware of such pitfalls and often attempt to tailor their analyses to accommodate small samples. Despite well-intentioned efforts, conventional statistical axioms do not always translate to best practice with small samples and common recommendations can, counter-intuitively, exacerbate small-sample problems. In this paper, we overview the landscape of small-sample techniques and note why conventionally-recommended approaches can fail with small samples while also suggesting lesser-known alternatives that tend to perform better in statistical research but are not widely adopted in educational research. Topics include bootstrapping, latent variable model fit, and Bayesian methods for multilevel, latent variable, and growth models. Simulated and real data examples are interspersed throughout.

KEYWORDS: multilevel model, model fit, Bayes, growth model, small sample

Challenging Conventional Wisdom for Multivariate Statistical Models with Small Samples

In educational research, small sample datasets are common for a variety of reasons. For instance,

1. Researchers may not have the necessary funds to amass large samples or to sample from a wide enough geographical area to recruit enough participants.
2. Researchers may be using secondary data or a subsample of secondary data for which it is not possible to augment the sample size.
3. The population of interest may be small by definition and it may therefore be challenging to amass large samples.
4. Studies may also be exploratory in nature and may not be intended to have large sample sizes.

Regardless of the rationale or the aim of a study, it is a ubiquitous interest in quantitative research that statistical methods be appropriate and yield trustworthy results. As advanced statistical methods continue to permeate into the educational literature, multivariate techniques such as multilevel models, growth models, factor analysis, and structural equation models are increasingly common, even for studies with smaller samples. For example, multilevel models are increasingly popular to account for dependencies that may arise for clustered data (e.g., students within schools; Raudenbush & Bryk, 2002; Hox, 2010), growth models are commonly applied to assess change over time (e.g., Singer & Willett, 2003), and factor analysis models are used to assess properties of scales used to capture directly unobservable constructs (e.g., Kline, 2013).

Based on recent meta-analytic evidence, each of these multivariate statistical methods are commonly applied to small sample data in education or closely aligned behavioral science disciplines. As examples, 18% of structural equation model (SEM) studies featured samples below 100 (MacCullum & Austin, 2000), 40% of exploratory factor analysis studies in personality psychology had samples less than 100 and over 60% had samples below 200 (Russell, 2002), 50% of educational meta-analyses had samples less than 40 (Ahn, Ames, & Myers, 2012), 33% of growth models had samples below 100 (Roberts & del Vecchio, 2000; Roberts, Walton, & Viechtbauer, 2006), and 21% of multilevel model studies had fewer than 30 higher-level units (Dedrick et al., 2009). Though obviously not ideal statistically, small sample sizes are quite common and studies based on small samples are routinely conducted, reported, and published.

As many responsible researchers are aware, small samples often require alternative statistical methods or corrective procedures in order to yield trustworthy results and many suggestions permeate the literature. However, despite the intuitive appeal of some potential remedies, routine small sample suggestions are not necessarily wise choices and do not produce estimates with reliable statistical properties. Results obtained via methods with poor statistical properties may contribute to the ongoing replication crisis in behavioral sciences as statistical instability undoubtedly produces results that are difficult to replicate.

In this paper, we overview some of the common small sample suggestions that tend to be recommended even though they do not always perform well and better alternatives exist. Much of the small sample research that appears in methodological journals is fairly technical, is spread across many different disciplines, and often deals with very narrow problems. As such, empirical researchers who are analyzing small sample data may not possess the statistical background or interest to fully follow all the arguments being made or may, more importantly, be unable to locate relevant studies on the topic at hand because they are spread across a vast number of fields, often outside of the educational journals that researchers are accustomed to reading.

Additionally, the educational research community currently does not demand coherence of small sample methods to small sample data as strictly as adjacent fields like psychology or public health where many small sample methods are developed and are more commonly employed. Therefore, some researchers may not be aware that they possess small sample data, may not be aware that there is a growing statistical literature dedicated to small sample analyses, or may not be aware that specialized methods are required to more appropriately model such data.

To present these findings in the most digestible format, this paper will deviate from the format that is generally seen in articles appearing in *Review of Educational Research.* The focus of this paper is methodological in nature and reviews of methodological studies tend to be quite dense and not much more coherent to empirical researchers than the original studies themselves because they tend to recap aspects of simulation studies like design conditions, relative bias, or coverage intervals that have little overlap with how empirical researchers interact with statistical methods in applied settings. Instead, this paper is formatted as a “how to” guide such that the review of methodological studies is embedded in the presentation of why certain methods are or are not desirable with small samples.

Specifically, the format of the paper is as follows: first, we overview methodological studies to provide a rough outline of sample sizes that are considered to fall into the “small sample” realm for different types of models. Then, we list a common statistical approach that is either specifically related to small samples or a recommendation that may be accurate generally but that fails with small sample data. We provide some background context on the issue and why traditional methods tend to breakdown at smaller sample sizes (in a similar flavor as recent papers on similarly muddled topics such as Curran, Obeidat, & Losardo, 2010 for growth models; Depaoli & van de Schoot, 2016 for Bayesian analyses; van de Schoot, Lugtig, & Hox, 2012 for conducting measurement invariance studies). We proceed by reviewing the support and intuition behind common recommendations. Then, we discuss the issues associated with such approaches in small sample contexts, discuss why intuitively sound suggestions are not always sound, and provide a demonstration with a real data example or a small illustrative simulation to support the use of lesser-known techniques.[1] We end each section with a thorough set of recommendations for best practice with multivariate analyses and small samples.

What Classifies as a "Small Sample"?

Before delving into nuances of modeling small sample data, readers may wonder at what point their data become a "small sample problem" and when the following discussion may be pertinent to their data. Unfortunately, it is difficult to arrive at a strict cut-off point because the answer ultimately depends on a few important facets such as (a) model type, (b) sample size at each level if data are multilevel, (c) model complexity, and (d) measurement scale of the outcome (e.g., dichotomous, ordinal, continuous), among other criteria. Plainly, small sample issues boil down to the data not containing sufficient information for the model being fit. The declaration that data contain "sufficient" information is not categorical but rather a spectrum that depends of several factors pertaining to aspects of the data and model complexity.

Though specifics are difficult to pin-point, we can speak in rough generalities based on the literature to provide some guidance. In this paper, we focus on three general classes of models: (a) multilevel models for cross-sectionally clustered data, (b) growth models for repeated measures data, and (c) latent variable models such as confirmatory factor analysis. We also focus on continuous outcomes to keep the presentation as streamlined as possible.

[1] These examples are not intended to be comprehensive and are included merely to contextualize the issues that arise with small samples. Readers should refer to cited studies for more complete arguments as to why and when certain methods breakdown with small samples.

With multilevel models for cross-sectional data, assuming an intraclass correlation around 0.20 as is common in educational research (Hedges & Hedberg, 2007), a small to moderate number of predictors (about 4 to 8), no missing data, and 2 or fewer cluster-level random effects, data with 40 or fewer higher-level units are at risk for small samples problems and studies with 20 or fewer higher-level units should not be modeled with standard methods (see McNeish and Stapleton, 2016a for a review). In the classic example of students within schools, the “higher-level units” would be schools. With continuous outcomes, sample sizes within each cluster are not too worrisome provided that one has about 5 observations per cluster on average (Clarke, 2008). If data are sparsely clustered such that many clusters have 1 or 2 observations, researchers may wish to consider alternatives to multilevel models such as generalized estimating equations (McNeish, 2014).

For growth models, assuming one has a moderate number of observations per person (e.g., between 4 and 8), random intercepts and slopes, linear growth, no time-varying covariates, and a small number of time-invariant covariates (e.g., 4 or less), longitudinal data with 100 or fewer participants are at risk for small sample issues and data with 50 or fewer participants should avoid standard statistical procedures (e.g., Curran et al., 2010; McNeish, 2016a). Sample size requirements tend to be slightly higher for growth models compared to multilevel models generally even though the models are sometimes mathematically interchangeable (e.g., Curran, 2003) because more information tends to be desired from covariance structures in growth models (which are harder to estimate with small samples) compared to multilevel models and there are often fewer overall data points (i.e., there are fewer repeated measures than the average number of students in a school or classroom). More complex models such as those featuring non-linear growth, second-order growth, or latent classes will have higher sample size requirements.

Covariance structure models are much more difficult to generalize because models can vary in complexity much more drastically than multilevel or growth models. Additionally, a primary driving force of model stability with covariance structure models is actually the magnitude of the standardized factor loadings rather than the absolute sample size (de Winter & Doudou, 2012; de Winter, Dodou, & Wieringa, 2009; McNeish, 2017a; Wolf, Harrington, Clark, & Miller, 2013). With standardized factor loadings near .90, samples as small as 30 can be sufficient (de Winter et al., 2009). With standardized loadings near .40, sample sizes closer to 400 are necessary (e.g., MacCallum, Widaman, Zhang, & Hong, 1999; Wolf et al., 2013). When considering the adequacy of data-model fit, sufficient sample sizes for use of standard methods is about 150-200 for moderate models (e.g., 4 latent variables with 5 indicators each) and increases as a function of model complexity (Herzog & Boomsma, 2009).

Overview of the Paper

The remainder of this paper takes the following form: Section 1 focuses on bootstrapping standard errors for multilevel data, Section 2 focuses on model fit in latent variable models, and Section 3 focuses on using Bayesian estimation for growth models. Table 1 provides a general condensed summary of the three issues we cover.

Section 1: Bootstrapped Standard Errors for Small Samples

The Issue

In frequentist analyses, each parameter in the model (e.g., regression coefficient, factor loading) is estimated with a single point estimate. The interest of many statistical analyses, however, is to make inferences about whether parameters are non-null in the population.[2] These

[2] By non-null, we refer to the common goal of statistical tests to determine if the parameter of interest is equal to 0 in the population. For instance, testing whether a regression coefficient is non-null is essentially asking if the relation between the predictor and the outcome is zero, holding all other variables in the model constant.

hypotheses are typically tested (in a frequentist framework) with a test statistic (e.g., *t*, *F*, *Z*, χ^2) and an associated *p*-value to determine if the point estimate is statistically different from zero in the population. For instance, the *t* statistic for a regression coefficient is equal to the value of the coefficient divided by its standard error ($t = \hat{\beta} / SE_{\hat{\beta}}$). The standard error is a measure of the *sampling variability* of the parameter which assesses how precisely the parameter was estimated.[3] In other words, it represents how much the point estimate would change across different random samples from the population of interest. A large standard error (relative to the size of the point estimate) indicates that there is much uncertainty in the estimate whereas a small standard error (relative to the size of the point estimate) indicates that the estimate is more precise.

In the frequentist framework, a single point estimate is provided for each parameter in the model. For instance, in the software output for a regression model, there is a single coefficient estimate for each parameter. But, to inferentially test each parameter, some information about the sampling variability is required. Ideally, the study would be run over and over with unique random samples from the population of interest and then an empirical sampling distribution of values could be created. In empirical contexts, this approach is pure fantasy and studies are often conducted only once. So, how can the information about sampling variability be obtained from a single point estimate?

The solution in frequentist analyses is to make assumptions about the *sampling distribution* of the estimated parameter. The point estimate and null hypothesis can inform where the sampling distribution should be centered, but information about the shape of the distribution

[3] Standard error estimates are taken from the square root of the sampling variability estimate, so these two terms denote a simple transformation of the same quantity within the frequentist framework. We use both in this paper depending on which is more appropriate in context.

and information about its variability are still needed. To obtain this information, it can be shown mathematically that, as the sample size increases to infinity, the sampling distribution will more closely approach a normal distribution (often referred to as the *central limit theorem*). With maximum likelihood estimation, if the sampling distribution is assumed to be normal, then mathematical arguments can show that the curvature of the likelihood function (referred to as Fisher Information) can closely approximate the sampling variability as sample size approaches infinity.

However, notice that the mathematical definitions presuppose that the sample size is approaching infinity (referred to as *asymptotic* in mathematical terms). This means that for larger samples, sophisticated mathematics can accurately estimate the sampling variability from a single point estimate (without having to repeat the study over and over). However, with smaller samples the central limit does not yet apply and Fisher Information is a poor approximation of the sampling variability (Efron & Hinkley, 1978). This typically leads to sampling variability estimates that are too small which inflates the operating Type-I error rate.[4]

Common Supporting Argument

Instead of relying on the central limit theorem and assuming that the sampling distribution is normal with smaller samples, it would be more appropriate to allow the sampling distribution to take whatever form necessary. This is the idea behind *bootstrapping* (e.g., Efron, 1979, 1986). Bootstrapping does not rely on the central limit theorem and instead uses resampling to determine the shape of the sampling distribution and to determine sampling

[4] The operating Type-I error rate refers to the proportion of times that an effect would be declared non-null when it is in fact null in the population. Researchers typically set this value to .05 in empirical studies (referred to as the nominal Type-I error rate); however, there is no guarantee that statistical methods will accurately adhere to the .05 value. Statistical research often studies contexts in which the properties of statistical methods breakdown and the nominal and operating Type-I error rates diverge. Divergence is highly undesirable because decision errors become more probable and are characterized inaccurately.

variability. Bootstrapping takes many *n*-sized replicated samples of the original data (usually 1,000 or more), *with replacement*, where *n* is the total sample size. That is, random samples of the original data are taken so that different combinations of the original data form "pseudo-datasets" that share the properties of the original data but that are not exact copies. In essence, this mimics the ideal situation of conducting the study over and over but only requires a single set of data. The standard resampling bootstrapping algorithm is depicted in Figure 1.

Sampling with replacement means that some observations may be selected multiple times while others may not be selected at all for a particular pseudo-dataset. This ensures that the replicated datasets are equal in size but that each replication is composed of different combinations of observations from the original data. The model is then estimated for each of the replicated pseudo-datasets and the point estimate for each replication is saved for each parameter. At the end of the process, researchers have thousands of possible estimates for each parameter. Instead of having a single point estimate, the replicate point estimates can be used to form an empirical sampling distribution, similar to the ideal case whereby researchers could run their study infinitely many times. As a result, researchers need not assume the distribution is normal because the estimates from each replicate pseudo-dataset allow the distribution to have any shape possible. Calculating the sampling variability (or a confidence interval) with so many values is quite easy and no longer depends on large sample assumptions (e.g., a 95% confidence interval is taken from the 2.5 and 97.5 percentiles of the empirical distribution of estimates from each of the pseudo-datasets).

Counter Argument

Bootstrapping is theoretically appealing and widely thought of as a panacea that solves many problems that plague statistical inference. However, bootstrapping encounters several difficulties with smaller samples, particularly for multilevel data structures (students within schools, repeated measures within students) that are ubiquitous in educational research and are often responsible for reduced sample sizes (Cameron, Gelbach, & Miller, 2008; Flynn & Peters, 2004; Sherman & Cessie, 1997). Because of the frequency with which multilevel datasets are responsible for small sample issues in educational research, this section focuses on issues with bootstrapping in this context.

One large issue stems from the sampling of observations with multilevel data. In single-level data, resampling observations to create pseudo-data is relatively straight forward and is performed with a simple random sample (with replacement) to produce a replicated sample size equal to the sample size in the original data. This issue becomes much more complex when faced with clustered data (Flynn & Peters, 2004; Preacher & Selig, 2012; Sherman & Cessie, 1997).

The common approach to bootstrapping with clustered data is the cluster bootstrap (Cameron et al., 2008; Kolenikov, 2010). In the standard cluster bootstrap, pseudo-data are created by selecting clusters rather than individuals as in the standard bootstrap. For example, if School 1 is sampled, then all students in School 1 are included in the pseudo-data for that replication. This approach tends to work well with larger samples (Field & Welsh, 2007; Ren et al., 2010), but it falters with smaller samples. With multilevel data, the dependent structure of the data is meaningful analytically, otherwise it would simply be ignorable or the multilevel structure would be treated as a nuisance with a method like cluster-robust errors rather than a multilevel model (McNeish, Stapleton, & Silverman, 2017). With smaller sample sizes and resampling, the dependent data structure is much less likely to be retained within the pseudo-datasets. As a result, key assumption of bootstrapping, namely that each pseudo-dataset is equally representative of the population as the original data, is more tenuous. As seen in simulation studies devoted to the topic of bootstrapping clustered data with small sample sizes (for full details, see Cameron et al., 2008), the result is that the sampling variability remains underestimated – no longer because of assumptions about asymptotics, but because resampling does not adequately capture the appropriate amount of sampling variability at smaller sample sizes.

As shown in a recent simulation by Huang (2017), the issues associated with the standard cluster bootstrap yield worse performance compared to a multilevel model with small sample methods such as restricted maximum likelihood estimation and a Kenward-Roger correction (these methods are discussed below) and tends to meaningfully underestimate the standard errors of regression coefficients in clustered data. Cameron et al. (2008), Flynn and Peters (2004), and Sherman and Cessie (1997) similarly found that the cluster bootstrap had inflated Type-I error rates for higher-level sample sizes as large as 30. Flynn and Peters (2004) note that higher-level sample sizes of 30 far exceed the number of clusters that are present in most clustered randomized trials, a common situation in educational research that leads to small higher-level sample sizes.

What to Do Instead

Difficulties with bootstrapping multilevel data are recognized by some software such as M*plus* which does not permit the use of bootstrap methods with clustered data (Muthén & Muthén, 2012). Though bootstrapping is sometimes seen as a blanket method that makes problems disappear by relaxing assumptions, with the type of small sample data commonly seen in educational research, the complexities of the data structure tend to make the standard cluster bootstrap less appealing, less advantageous, and less effective for small samples. This is especially true when one considers alternative, small sample specific methods that exist in the multilevel modeling framework which have been shown to provide estimates with highly desirable statistical properties even when the number of higher-level units is in the single digits (Ferron et al., 2009; McNeish & Stapleton, 2016b). As succinctly noted in the discussion section of Cameron et al. (2008), “The usual way that the bootstrap is used, to obtain an estimate of the standard error, does not lead to improved inference with few clusters as it does not provide an asymptotic refinement” (pp. 424).

Instead, we advise that empirical researchers take one of two approaches. First, researchers can simply avoid bootstrapping with small sample multilevel data in favor of restricted maximum likelihood estimation and a Kenward-Roger correction (Kenward & Roger, 1997), if possible. This combination of methods has been shown to yield more appropriate inferential decisions with small samples in many simulation studies (Baldwin & Fellingham, 2013; Bell et al., 2014; Bryan & Jenkins, 2016; Ferron et al., 2009; McNeish & Stapleton, 2016a, 2016b). Restricted maximum likelihood is an estimation method that is equivalent to traditional maximum likelihood with large samples but takes additional precautions to improve small sample properties of estimates. Without getting into technical details, the difference between maximum likelihood and restricted maximum likelihood is akin to formulas for computing population variance (which uses a denominator of *n*) and for computing sample variance (which uses a denominator of *n* – 1): there is no difference at large samples but the population version yields biased estimates at smaller sample sizes.

The Kenward-Roger correction is a two-step procedure that improves sampling variability estimates of multilevel models with smaller samples. Though quite mathematically intensive, the first step of Kenward-Roger is a finite-sample size correction that attempts to correct the sampling variability estimates to account for the fact that asymptotic assumptions are not preserved and that Fisher Information will underestimate sampling variability with small samples. The second step is a sophisticated procedure to more accurately calculate degrees of freedom which further helps to refine *p*-values with smaller samples. These methods are available in SAS, Stata, and R (in the pbkrtest package) but, at present, cannot be implemented in HLM or SPSS. One possible downside with these methods is that they can be difficult to generalize, so the statistical literature has yet to adapt them broadly. Therefore, they are available for straightforward models but may not be available for complex models. For a full article-length non-technical explanation of the differences between restricted and full maximum likelihood as well as the Kenward-Roger correction, readers are referred to McNeish (2017b).

Alternatively, if restricted maximum likelihood and the Kenward-Roger correction are not available for the desired analysis or if there are additional issues with the data that necessitate bootstrapping with small samples, the standard clustered bootstrap should generally be avoided in favor of alternative versions of the bootstrap. Cameron et al. (2008) found that the *wild cluster bootstrap* (available in Stata or the clusterSEs R package) performs best with small sample multilevel data and has desirable properties with higher-level sample sizes as small as 6 with less complex models (similar to the Kenward-Roger correction which has been found to perform well with higher-level samples as small as 7 with less complex models; McNeish & Stapleton, 2016b). The general algorithm for the wild bootstrap is presented in Figure 2. The basic idea is that pseudo-datasets are created by randomly perturbing residual terms rather than by resampling the data as in the standard cluster bootstrap. This is advantageous for multilevel data because the data structure is left completely intact and need not be resampled (Wu, 1986), alleviating the main drawback of the standard cluster bootstrap.

Real-data example. To demonstrate the issues present with the cluster bootstrap for clustered data with small samples, we compare the performance of five different methods for data from a clustered randomized trial from Stapleton, Pituch, and Dion (2015). The methods include:

- Maximum Likelihood
- Restricted Maximum Likelihood
- The Cluster Bootstrap
- The Wild Cluster Bootstrap
- Restricted Maximum Likelihood with a Kenward-Roger Correction

The study is interested in a new socioemotional curriculum for 84 children in 14 Head Start sites to determine if the intervention (the new curriculum) affects children's behavior. The Head Start sites are balanced such that 6 children are clustered within each site. This example is fairly standard for clustered randomized trials conducted in educational settings: the overall sample size reasonable but the higher-level sample size is quite small because it is difficult to attract a large number of sites.

Before continuing, we would like to point out that the What Works Clearinghouse (WWC) standards only require 14 clusters in cluster randomized trials for the extent of evidence to be considered "medium to large" (see Table IV.4 of the WWC Procedures and Standards Handbook, Version 3.0). Though the number of children per center in this example would need to be larger for the whole study to meet these requirements (the guidelines call for 25 observations per higher-level unit), issues related to estimation and power with small sample sizes in multilevel models are almost exclusively tied to the higher-level sample size (Scherbaum & Ferreter, 2009; Snijders & Bosker, 1993). Lower-level sample sizes tend to affect models with continuous outcomes only when the number of observations per cluster is less than 5, on average (Clarke, 2008; McNeish, 2014) and increasing lower-level sample sizes tends to minimally augment statistical power (Scherbaum & Ferreter, 2009). Thus, it is very possible to meet the WWC standards for "medium to large" extent of evidence and have the model be stricken with serious statistical concerns simultaneously if appropriate precautions are not taken to accommodate reduced higher-level sample sizes.[5]

Returning our attention to the example data, the outcome variable is a continuous measure of behavior (*M* = 36.48, *SD* = 4.38) and is predicted by a group-mean centered measure of the child's knowledge of socioemotional concepts, the treatment group (assigned at the site level), and the site average for knowledge of socioemotional concepts (i.e., the model uses a between-within specification common for assessing contextual effects; Enders & Tofighi, 2007). The data are intended to be modeled with a mediation model though we simplify the model such that only the dependent variable is modeled (i.e., the "b" path of the mediation model).

Figure 3 shows treatment effect *p*-values (the focal predictor) for each estimation type. The point estimates of the treatment effect between methods were equal to the second decimal point at 2.76 (i.e., children in the treatment sites score 2.76 points higher on the behavior measure than children in the control sites). As has been noted in many studies, sampling variability is highly underestimated with maximum likelihood (Browne & Draper, 2006;

[5] The full WWC requirements to have a medium or large extent of evidence are (a) the domain includes more than one study, (b) the domain includes more than one setting, and (c) the domain findings are based on a total sample size of 350 OR 14 classrooms of 25 students each (What Work Clearinghouse, p. 30). Sample sizes for multilevel studies are not mentioned elsewhere in the document. The recommendation for these sample sizes are stated to be related to having adequate statistical power. However, it is well-known in the multilevel model literature that power, especially for higher-level coefficients where an intervention effect would be located, is primarily related to the number of higher-level units, not the overall sample size. Additionally, the value for the intraclass correlation has a high impact on power in multilevel models and is nearly as relevant as the overall sample size (Bliese & Halverson, 1998; Hedges & Hedberg, 2007) because higher intraclass correlations reduce the amount of unique information within each site and reduce the effect sample size (Kish, 1965).

McNeish & Stapleton, 2016a); therefore, the *p*-value for the maximum likelihood treatment effect is the lowest of all methods (underestimated sampling variability means that the precision of estimates is overestimated, leading to artificially low *p*-values). The restricted maximum likelihood *p*-value is noticeably higher than maximum likelihood, although restricted maximum likelihood is not completely free from asymptotic assumptions. The Kenward-Roger correction reduces these asymptotic assumptions and, as a result, it can be seen that the *p*-value is again higher compared to restricted maximum likelihood because the standard errors are further adjusted and the degrees of freedom are refined. Next, compare the two different versions of bootstrapping to the various likelihood-based *p*-values. Even though the maximum likelihood *p*-value is known to be deflated, the cluster bootstrap *p*-value is quite close to the maximum likelihood *p*-value, indicating that the sampling variability is likely underestimated. Conversely, the wild cluster bootstrap *p*-value is most closely aligned to the Kenward-Roger *p*-value which best accounts for nuances of small sample multilevel data.

Though this is only one example and methodologically advanced readers will be quick to note that the results do not generalize as they would if obtained via a comprehensive simulation, these data demonstrate consistent patterns that have been found previously in such comprehensive simulations (Browne & Draper, 2006; de Winter, 2013; Hesterberg, Moore, Monaghan, Clipson, & Epstein, 2005; Huang, 2017; McNeish & Stapleton, 2016a, 2016b; Preacher & Selig, 2012; Sherman & Cessie, 1997, van der Leeden, Meijer, & Busing, 2008; Wang, Carpenter, & Kepler, 2006).

Section 2: The Maximum Likelihood Test Statistic for Latent Variable Model Fit with Small Samples

The Issue

Latent variable models propose that unobserved (and directly unmeasurable) variables theoretically exist and their presence is implied through relations with observed variables. The network of these relations is commonly referred to as a *measurement model* where latent variables "load" on a set of observed variables. When a latent variable model consists solely of a measurement model to empirically test the existence of the directly unobservable construct(s) of interest, then the model is a *confirmatory factor analysis* (CFA).[6] If causal relations are specified between latent variables (in addition to the measurement model), then the model is a *structural equation model* (SEM). If the hypothesized model is advanced as an explanation for the phenomenon of interest, it is necessary (but not sufficient) to demonstrate that the specified theoretical model reproduces relations between the observed variables (Kline, 2013). That is, there exists some relation between the individual observed variables in the data and a CFA model posits that these relations arise because multiple observed variables are indicators for a broader, unmeasured latent variable. To ensure the plausibility of this theoretical model, it is important to verify that the relations implied by the CFA model (where relations in observed variables are caused by the presence of an unmeasured variable) are reasonably close to the raw empirical relations between the observed data. Though conceptually clear, the appropriate way to demonstrate good or poor data-model fit is not clear and arguments for and against existing methods have been rather contentious (Barrett, 2007; Bentler, 2007; Hu & Bentler, 1999; Marsh, Hau, & Wen, 2004).

[6] In educational measurement contexts, when the observed variables are discrete as is common with item responses, a measurement model is often referred to as an *item response theory* (IRT) model. There is much overlap between IRT models and CFA models for discrete data (Glockner-Rist & Hoijtink, 2003) and many IRT models (e.g., the Rasch model and the 2PL model) can be equivalently specified as a CFA model (Curran et al., 2014). However, other IRT models such as the 3PL model are less adaptable as CFA models. Given that we focus on continuous outcomes in this paper, we stick to CFA terminology but this distinction should be noted.

As an overview of the debate, some argue that the maximum likelihood test statistic ($T_{\mathrm{ML};}$ often referred to as "the χ^2 test") is the most appropriate because it is a true statistical test of model fit: it sets up a null hypothesis that the relations implied by the hypothesized model exactly reproduce the relations of the observed data and the test reports a *p*-value for the hypothesis test. Conversely, others support approximate goodness of fit indices such as the standardized root mean square residual (SRMR; Jöreskog & Sörbom, 1981), root mean square error of approximation (RMSEA; Steiger & Lind, 1980), or comparative fit index (CFI; Bentler, 1990) on the grounds that the null hypotheses of T_{ML} is not appropriate and rejection of the T_{ML} null hypothesis does not indicate that the hypothesized model is poor. A common support for approximate goodness of fit indices comes from the adage that all models are wrong, but some are useful (paraphrase of Box, 1976 p. 792) and that a model that is not exactly correct (as is being tested by the T_{ML} null hypothesis) should not necessarily be dismissed (Hooper, Coughlan, & Mullen, 2008). Thus, fit indices act more like effect size measures whereby they attempt to quantify the degree of potential misspecifications rather than testing a particular null hypothesis.

Common Supporting Argument

When comparing T_{ML} to fit indices, a common argument for fit indices is that T_{ML} tends to be overpowered as sample sizes grow larger (Jöreskog & Sörbom, 1993). The null hypothesis advanced by T_{ML} is that the model-implied relations and the observed variable relations are *exactly* equal. As the sample size grows larger, just as in any statistical analysis, the probability of rejecting the null hypothesis when it is false increases (i.e., power increases). Therefore, with large samples, researchers have very high power to detect even trivial differences between the model-implied and observed variable relations. As alluded to above, this is not necessarily desirable because trivial differences between the model and the data are not usually interesting (Steiger, 2007).

However, with smaller samples, the "overpowered T_{ML}" rationale does not apply – with a few hundred or fewer observations, there is no immediate danger of T_{ML} being overpowered (Markland, 2007).[7] Furthermore, recent studies have noted that the meaning of fit indices is often inconsistent (Hancock & Mueller, 2011; Heene, Hilbert, Draxler, Ziegler, & Bühner, 2011; Kang, McNeish, & Hancock, 2016; Miles & Shevlin, 2007; Saris, Satorra, & van der Veld, 2009; Savalei, 2012). Because these indices are not true statistics, they often do not have known distributions and therefore are not capable of providing inferential information (e.g., there are no hypotheses or *p*-values). Researchers reporting these indices must rely on recommended cut-off values which are often derived via simulation designs. Despite an existing set of commonly cited cut-offs provided by Hu and Bentler (1999) that have proved quite popular, recent studies have shown via demonstration (Hancock & Mueller, 2011; Miles & Shevlin, 2007) and analytical proofs (Heene et al., 2011) that these cut-offs are anything but fixed and that their meaning changes depending on aspects of the model and data such as reliability of the latent variables, model complexity, sample size, and model type (Fan & Sivo, 2005, 2007; Kang et al., 2016; Kim, 2005; Marsh et al., 2004). Thus, T_{ML} is considered to be more appropriate with small or moderate sample sizes because it has a clearer meaning and interpretation while also allaying the associated worries of being unnecessarily overpowered (e.g., Iacobucci, 2010; Markland, 2007).

Counter Argument

[7] Also note that as sample sizes fall toward the small end of the spectrum, T_{ML} can also become underpowered to detect meaningful differences between the model-implied and observed relations between variables (Tomarken & Waller, 2003). Before conducting a small sample latent variable analysis, it may be prudent to conduct a small Monte Carlo study where data are generated from one model and fit with misspeicified models (e.g., Nevitt & Hancock, 2004). This would allow researchers to provide evidence that, if the model were misfitting, then T_{ML} would have a reasonable chance of detecting such an effect. Interested readers are referred to Muthén & Muthén (2002) for details on how to conduct such a Monte Carlo analysis.

Despite the improved clarity of interpretation of T_{ML} in the absence of large samples, small sample issues exist within properties of the T_{ML} statistic itself (Curran et al., 2002). Namely, T_{ML} is χ^2 distributed asymptotically with degrees of freedom equal to the degrees of freedom for the latent variable model in question. The keyword here, as in Section 1, is "asymptotically".

Going back 65 years, statistical researchers were aware that T_{ML} performs poorly with small sample sizes (where "poor performance" refers to inaccurate *p*-values, not insufficient power to detect differences). Bartlett (1950) demonstrated this issue in addition to proposing a small sample correction in the case of exploratory factor analysis. Since then, many other methodological studies found similar results in that T_{ML} does not follow the appropriate χ^2 distribution with smaller sample sizes – typically less than 200 observations – although model complexity is important to consider (Bentler & Yuan, 1999; Herzog & Boomsma, 2009; Kenny & McCoach, 2003; Nevitt & Hancock, 2004; Savalei, 2010, Yuan, 2005). Importantly, T_{ML} with small samples exceeds the expected value of the associated χ^2 distribution, meaning that T_{ML} tends to *over-reject* models and tends to make models appear worse-fitting than they actually are. Statistical arguments have shown that it is also not possible to mathematically transform T_{ML} such that is follows the appropriate χ^2 distribution (for details, see Fujikoshi, 2000 or Yuan, Tian, & Yanagihara, 2015).

Illustrative simulation. To visually demonstrate the issue with T_{ML} and small samples, we generated data from a three factor CFA model where each factor loaded on 5 items. Figure 4 shows the path diagram from which the data were generated. We generated 1,000 unique datasets based on this model and then fit the *exact same* model to the data. For readers not well-versed in the idea behind simulations, given that we generated the data ourselves, we know the true population model (unlike real data analyses). This allows us to have insight as to whether the model should fit well or fit poorly so that we can assess the performance of model fit statistics. The data generation model and the fitted model are identical in this demonstration, so the fit of the model should be quite good. If T_{ML} is accurate, we should reject the model in 5% of the replications by chance (the nominal Type-I error rate of the test). We generated data with sample sizes of 25, 50, 75, 100, 150, 200, 250, and 500 to show how the adherence of T_{ML} to the theoretical χ^2 distribution changes with sample size. Results are shown graphically in Figure 5. Notice how the rejection rates of T_{ML} do not correspond to the nominal 5% rate until somewhere between 250 and 500 observations are present in the model. Even with 200 people, about 10% of models are rejected and considered poorly fitting, even though the model should fit *perfectly* and the rejection rate should be around 5% strictly from chance alone. For smaller samples, the rejection rates continue to climb as T_{ML} deviates from a χ^2 distribution, ultimately reaching 86% in the (unrealistic) condition of only 25 observations (That is, 86% of perfectly specified models are reported to fit poorly according to T_{ML} with a sample of 25 people). As with the example from Section 1, these results are not intended to generalize to all situations and are merely a demonstration. Also like Section 1, the general idea that T_{ML} does not follow the appropriate χ^2 distribution and over-rejects models that fit well with small samples has been repeatedly shown in the literature (e.g., Herzog & Boomsma, 2009; Herzog, Boomsma, & Reinecke, 2007; Kenny & McCoach, 2003; Nevitt & Hancock, 2004).

What to Do Instead

Even though T_{ML} performs poorly with small samples and cannot be directly transformed to follow the appropriate χ^2 distribution, methodological research has devised two broad options for more appropriately assessing data-model fit with small samples: (1) heuristic small sample corrections to T_{ML} or (2) reporting alternative fit statistics that have more desirable small sample properties.

Heuristic small sample corrections to $\boldsymbol{T}_{\mathbf{ML}}$. In addition to the Bartlett correction for exploratory models provided in Bartlett (1950), two other more general heuristic small sample corrections have appeared in the literature: Swain (1975) and Yuan (2005). These corrections are quite easy to apply – researchers only need to estimate the model as they normally would (typically with maximum likelihood estimation). The T_{ML} statistic output by the software is then manipulated based on the number of latent variables, number of observed variables, sample size, and/or the degrees of freedom to correct it so that it more closely follows the appropriate χ^2 distribution. As noted previously, it is not possible to perfectly transform T_{ML} so that it follows the appropriate χ^2 distribution but these heuristic corrections can yield values that are much closer to the appropriate distribution than raw T_{ML} values. We will not report the exact formulas here, but they can readily be found in Herzog & Boomsma (2009), Savalei (2010), or McNeish and Harring (2017), the last of which also provides a link to an Excel spreadsheet for calculating these values. General latent variable software programs do not provide these corrections nor options to implement them, although the Swain correction can be automated in the `swain` or `FAiR` R packages using software output. Though there is not much software support, the corrections are straightforward enough to compute manually with a hand calculator or with a spreadsheet. The corrected test statistics are then compared to a χ^2 distribution with degrees of freedom equal to the degrees of freedom for the latent variable model of interest that would normally be used with T_{ML}.

To demonstrate the utility of these corrections, we re-analyzed the simulation data but applied the Swain, Yuan, and Bartlett corrections to T_{ML}. The results are shown in Figure 6. The changes between the corrected and uncorrected versions Type-I error rates are stark. The only circumstances in which the Type-I error rate is not reasonably close to the nominal rate are Swain correction with sample sizes of 25 or 50. Otherwise, rejection rates are well-behaved across the range of sample sizes. These results replicate patterns that have been found with these corrections in simulation studies. Namely, that the Swain correction tends to provide the smallest correction but provides researchers with the most power to detect whether their model fits poorly while the Bartlett correction tends to provide the harshest correction and best controls the Type-I error rate but has the lowest power to detect misfit. A handful of simulation studies explored these corrections in the context of CFA models, SEMs, and growth models when there are no missing data (Fouladi, 2000; Herzog & Boosmsa, 2009; Nevitt & Hancock, 2004). The corrections are also not limited only to T_{ML} and can be applied to robust estimators like the Satorra-Bentler test statistic which accounts for violations of normality assumptions (Satorra & Bentler, 2001; Savalei, 2010).

One issue with these corrections is that they use the sample size in the correction formulas. If missing data are present, the value to use for sample size in the formulas is not clear (counting each individual with missing data as a full observation is too liberal, counting each individual with any missing data as completely missing is too conservative). McNeish and Harring (2017) address the issue of how to calculate sample size in these corrections with missing data and present a post-hoc solution in the context of growth models, though there is much room for improvement in this area in future research.

Alternative fit statistics. Recent research by Yuan, Bentler, and colleagues has developed many alternative statistics to assess the fit of latent variable models in the context of small samples (e.g., Bentler & Yuan, 1999; Yuan & Bentler, 1997, 1999; see the simulation in

Bentler & Yuan, 1999 for a comparison of these methods under many conditions). One aspect worth considering with these statistics is that they are based on asymptotic distribution free (ADF) estimation methods. With ADF methods, the sample size must be at least equal to $p(p+1)/2$ where p is the number of observed variables in the model. This can be problematic for reasonably complex models and small samples, though the benefit of ADF methods is augmented robustness to non-normality compared to maximum likelihood. Many of these statistics can be obtained in the EQS software program but they tend to be too complex for manual computation outside of specialized software.

Section 3: Bayesian Methods with Small Samples

The Issue

As a brief recap of the frequentist framework, the parameters are considered fixed quantities that exist in the population (and the data are considered to be random), which means that each parameter is estimated by a single value. However, when performing inferential tests to determine whether parameters are non-null in the population, some estimate of variability is required to gauge the uncertainty of the estimates. Although only a single point estimate is obtained in the frequentist framework, sampling variability can still be discerned via methods that make asymptotic assumptions like the central limit theorem and Fisher Information as discussed in Section 1.

Though reasonable with large samples, this assumption is problematic with small samples and the viability of the assumption tends to be poor (van de Schoot et al., 2014). This leads to poor variability estimates, usually such that standard error estimates are too small, which overstates precision and inflates the operating Type-I error. Bootstrapping is one proposed method to relax these assumptions; Bayesian methods are another method suggested to obtain better sampling variability estimates with improved statistical properties with small samples. Many recent studies have noted the small sample advantages of Bayesian methods and have advocated for their use over frequentist methods (Baldwin & Fellingham, 2013; Hox, van de Schoot, & Matthijsse, 2012; Lee & Song, 2004; Muthén & Asparouhov, 2012; Stegmueller, 2013)

Common Supporting Argument

Bayesian methods alter the conception of the parameters and the data – in the Bayesian framework, the data are viewed as fixed and parameters are viewed as random (Kruschke, 2015). This means that each parameter in a Bayesian analysis is estimated with a distribution rather than being estimated with a single point estimate. As a brief overview, Bayesian methods are concerned with combining information and interpret the definition of probability to be about updating knowledge (Gelman, Carlin, Stern, & Rubin, 2014). This is opposed to the frequentist framework which defines probability in terms of long-run frequencies (Spiegelhalter, Myles, Jones, & Abrams, 1999). Maximum likelihood estimation bases its estimation strictly on one quantity: the likelihood. The Bayesian interpretation of probability results in estimates being based on three quantities: the prior distribution, the likelihood (the same likelihood from maximum likelihood), and the posterior distribution.

Bayesian probability is about combining and updating information, so it follows that there must be some initial baseline from which to begin updating. In Bayesian analyses, this is the prior distribution. Before considering any data, researchers must specify a *prior distribution* for each parameter in the model of interest. Prior distributions can be *informative* if researchers have a strong inclination regarding plausible values for each parameter or, more commonly, they can be *diffuse* such that the distribution covers a wide range of possible values if researchers are not sure (a prior distribution is placed on each parameter in the model so it is also possible for some priors to be informative while allowing others to be diffuse). When the data are subsequently modeled, the likelihood is computed. The likelihood contains information strictly from the data and is not affected by the prior distribution. A weighted combination of the prior distribution and the likelihood form the posterior distribution, which is conceptually similar to the sampling distribution in a frequentist framework. When using the common Markov Chain Monte Carlo (MCMC) method to estimate Bayesian models, the posterior distribution is formed by several thousand replications of possible parameter estimates. The central tendency of the posterior distribution is the Bayesian conceptual equivalent of a frequentist point estimate and the standard deviation of the posterior distribution is the Bayesian conceptual equivalent to the frequentist standard error. Figure 7 shows a conceptual diagram of the basic tenets of how the posterior distribution is formed in a Bayesian analysis.

Because researchers obtain an actual empirical distribution of values via the posterior distribution, they need not rely on the central limit theorem or Fisher Information (Lee & Song, 2004). Without these assumptions, it follows that Bayesian methods can produce estimates with more desirable statistical properties at smaller samples because they yield more accurate estimates of sampling variability compared to frequentist methods, which consequently leads to more appropriate statistical inferences (Depaoli & van de Schoot, 2016; Dunson, 2001; Lee & Song, 2004; Muthén & Asparouhov, 2012). A review of empirical studies using Bayesian methods by van de Schoot et al. (2017) found that about 15% of studies resort to Bayesian methods because of sample size related issues, demonstrating that empirical researchers often turn to Bayesian methods to combat small sample data.

Counter Argument

The general logic behind Bayesian methods being more appropriate for small samples is sound. *Theoretically,* the freedom from asymptotic assumptions gained by making inferences from a Bayesian posterior distribution does give Bayesian methods the *potential* to be more appropriate and more trustworthy than frequentist methods with small samples. Additionally, prior distributions *can* be cleverly specified to augment the amount of information available so that one's ability to detect effects is increased. However, the *implementation* of Bayesian methods can be complex and certain modeling choices can result in Bayesian methods being equal to or even worse than frequentist methods if caution is not taken (McNeish, 2016a).

The difficulty with small samples stems from the specification of the prior distribution. Researchers often specify diffuse prior distributions to avoid unduly intervening in the analysis or to let the data (via the likelihood) be the driving force in formation of the posterior – the review by van de Schoot et al. (2017) reports that about 73% of papers utilized prior distributions that were non-informative. With moderate or large samples, diffuse prior distributions tend to be rather innocuous and minimally impact the posterior distributions. This is due to the relative dominance of the likelihood in such cases: with larger samples, the likelihood contains much information (because it contains contributions from many observations included in the data). Even though the posterior distribution is a combination of the prior and the likelihood, the contribution is not necessarily equal. When the likelihood contains a large amount of information as with large samples and the prior is diffuse, the likelihood essentially drowns out the prior and the posterior is based almost entirely on the likelihood.

With small samples, however, the likelihood does not contain as much information because the data are not as informative. The likelihood has much less relative weight in forming the posterior distribution and the prior distribution (which, again, is set by the researcher and does not come from the data) plays a much larger relative role in the posterior distribution with smaller samples. Given the dominant preference in empirical studies to avoid informative priors (van de Schoot et al., 2017), this can unintentionally result in poor statistical properties (van de Schoot et al., 2015). Due to the increased relative importance of the prior, *any* prior distribution becomes informative with small samples and impacts the posterior distribution. So, utilizing the default diffuse prior in software or specifying a similarly diffuse prior distribution in more Bayesian specific software can adversely affect the resulting estimates (McNeish 2016a). The result of specifying diffuse priors with small samples is artificially inflated sampling variance which quickly reduces the model's ability to detect non-null effects (the Bayesian equivalent of reducing power). That is, the uncertainty contained with the diffuse prior is propagated into the posterior distribution, which unnecessarily increases sampling variability. The differential effect of diffuse priors for large samples and small samples is shown in a conceptual diagram in Figure 8.

Illustrative simulation example. Similar to Section 2, we conduct a small illustrative simulation to highlight some of the possible issues that have been discussed with Bayesian methods and small samples in the methodological literature. Sections 1 and 2 focused on multilevel models and latent variable models, respectively, so we focus on growth models in this section (though the issues are not unique to growth models).

We generate data from a linear growth model with 4 time-points and two binary time-invariant predictor variables. Four time-points were selected because this number of repeated measures is common because it balances financial costs of following participants longitudinally while also ensuring sufficient time-points to model possible non-linear trajectories (Curran et al, 2010; Vickers, 2003). Two time-invariant covariates are included to (a) make the model more realistic and (b) to track the ability of the model to detect non-null effects.[8] Standardized coefficients were set to 0.10 and 0.30. We explore sample sizes of 20, 30, 50, 75, and 100. Four estimation methods will be used for each of 500 generated datasets and the fitted models will contain no misspecifications. The four methods are:

- Diffuse priors on all parameters including an improper Inverse Wishart prior for covariance of the subject-specific growth factors (i.e., the M*plus* defaults).
- Diffuse priors on all parameters including an identity scaled Inverse Wishart prior for covariance of the subject-specific growth factors.
- Maximum likelihood as if the model were fit in general SEM software packages
- Restricted maximum likelihood with a Kenward-Roger correction

We selected the M*plus* defaults as one condition because, in their review, van de Schoot et al. (2017) note that M*plus* is the most common software reported for Bayesian analyses in behavioral sciences and its popularity has increased steadily over time. The identity scaled inverse Wishart was selected because this is a commonly recommended diffuse prior for researchers working in specialized Bayesian software like Stan, WinBUGS, or JAGS (Schuurman, Grasman, & Hamaker, 2016). A distinction is made between priors for the subject specific growth factors because these parameters tend to be the most difficult to estimate with smaller samples and also have a notable role in computing sampling variability (McNeish, 2016b). The two likelihood methods were included to serve as a comparison of the performance that researchers are avoiding when they choose Bayesian methods over frequentist methods with small sample sizes.

[8] In the frequentist framework, this would be referred to as power. However, in the Bayesian framework, there are no null hypotheses, so the technical definition of power (rejecting the null when the null is false) does not apply. We use the broader definition of “detecting a non-null effect” but readers may conceptually think of this as power if it helps reduce the technicality of the terminology.

Given the illustrative purpose of this simulation, we report on the relative bias of one problematic parameter with small sample growth models (the variance of the subject-specific slopes) and on the ability of each method to detect that the path of the time-invariant predictor with the 0.30 standardized effect is non-null. To overview relative bias for those not familiar with simulation studies, through generating the data ourselves, we know the true population values. If the statistical properties of the model are sound, on average, the model should estimate parameters to be reasonably close to the population values we set. Within ±10% is a common threshold (Hoogland & Boomsma, 1998) such that bias between -10% and +10% of the population value is considered acceptably close and bias less than -10% or greater than +10% is considered problematic. Figure 9a shows the relative bias of the subject-specific slope variance and Figure 9b shows the probability that each method was able to detect that the time-invariant predictor was non-null.

In Figure 9a, we first point out the bias of the (black, dashed) maximum likelihood line. The maximum likelihood estimates in this simulation are highly downward biased with sample sizes of 75 or fewer; a known issue that researchers hope to avoid by switching to a Bayesian framework. However, note the bias of the two Bayesian (solid) lines – the direction is reversed such that the estimates are upwardly biased but the magnitude of the bias generally exceeds the magnitude of bias from maximum likelihood and continues to exceed 10% beyond a sample size of 75. Thus, researchers are trying to avoid the bias associated with maximum likelihood by choosing Bayesian methods, but if the common strategy of using diffuse priors is employed, the bias can be *worse* than maximum likelihood. As noted previously, this behavior occurs precisely because the likelihood is not dominant enough to overtake the supposedly diffuse priors with smaller samples. So, the posterior contains a notable amount of noise from the prior. Put another way, though intended to be non-informative, at smaller samples, the diffuse prior becomes an informative prior (and not a particularly good one at that) because the likelihood carries relatively less weight in the posterior with smaller samples. Thus, the posterior distribution is adversely affected. Note, however, that frequentist methods with appropriate small sample corrections yield acceptable estimates (grey, dashed line), even in the smallest sample size condition.

Nonetheless, with smaller samples, researchers may be willing to live with some bias in the estimates provided that the model can detect effects when they are present. However, consider Figure 9b which shows the proportion of models that correctly identified that the time-invariant predictor with a 0.30 standardized coefficient is non-null (i.e., this plot shows the power of each method if one adopts frequentist terminology). Because maximum likelihood severely underestimates sampling variability with smaller samples and consequently has inflated Type-I error rates, it is not included in Figure 9b because inflated Type-I error rates distort interpretations of power. As expected, the detection probabilities are rather small at lower sample sizes and increase proportionate to sample size. However, rates for Bayesian methods are noticeably depressed compared to the restricted maximum likelihood estimates with the Kenward-Roger correction until the sample size reaches 100 (the point at which small sample bias dissipates). These reduced rates are at least partially attributable to the inflated variance of the posterior distribution, a side effect of diffuse priors with small samples.

What to Do Instead

There is no inherent flaw in resorting to Bayesian methods with small samples, generally speaking. The theoretical backing for Bayesian methods with small samples is sound; it is the *implementation* of Bayesian methods with small samples that presents issues. If one switches to

Bayesian methods merely to avoid small sample issues present within a frequentist framework without thinking through how to properly specify prior distributions, then researchers are not escaping the problem. In fact, as illustrated in our small simulation here (and more comprehensively elsewhere in the literature), employing Bayesian methods with small samples and diffuse priors can *increase* bias in parameter estimates relative to frequentist methods and *decrease* one's ability to detect effects when they are present compared to small-sample-corrected methods in the frequentist framework – both of which are highly undesirable. Bayesian methods certainly have promise to outperform frequentist methods with smaller samples because they are not reliant on the central limit theorem and because prior information can be included to complement the data (van de Schoot et al., 2014); however, researchers essentially trade issues related to the central limit theorem for issues related to specifying prior distributions (McNeish, 2016a).

If prior information is available via expert opinions or meta-analytic evidence, then the issue of specifying a meaningful prior is not daunting and Bayesian methods are indeed a viable option that can lead to higher probabilities of detecting non-null effects (McNeish, 2016a; van de Schoot et al., 2015). This property is heavily dependent on having accurate prior knowledge and, if the informative priors are set but they are inaccurate, the posteriors may also be adversely affected (Depaoli, 2014; Depaoli & van de Schoot, 2016). Alternatively, one may use methods like empirical Bayes or data-dependent priors to use the data to help inform prior distributions, though some may view this strategy as inappropriate because it "double dips" the data (see, e.g., McNeish, 2016b). Gelman (2006) also advanced a half-Cauchy prior that has shown promise as a fairly non-informative prior with small samples, although it can be difficult or impossible to specify in user-friendly software preferred by some researchers like M*plus* and it may not perform as well for complex models with small samples (McNeish & Stapleton, 2016b). The half-Cauchy prior distribution also requires researchers to specify a scale parameter, in which small changes can affect results. At any rate, using wide, diffuse priors with small samples to avoid intervening in the analysis does more harm than good with small samples and researchers hoping to follow this strategy may be better off staying within the frequentist framework because, with small samples, setting diffuse priors *is* intervening.

Instead, for many straightforward models with continuous outcomes that are commonly applied to educational data, frequentist small sample corrections suffice, although these methods are not always available in the context of more complex models. More succinctly, neither framework makes the small sample issue go away or resolves it entirely – researchers must either use small-sample-specific frequentist methods or must provide informative priors in a Bayesian context. Either is acceptable and it is true that, with strongly informative priors, Bayesian methods minimize bias and maximize the ability to detect effects compared to frequentist methods (McNeish, 2016a; van de Schoot et al., 2015). However, switching to Bayesian methods without strong prior information does not immediately void small sample problems and researchers must do their due diligence to capitalize on the favorable small sample properties of Bayesian methods.

Conclusions

Small sample sizes are never an optimal outcome of an empirical study but they are indeed a fact of life in the realm of educational research, especially for multilevel data, repeated measures data, or pilot studies for novel scales. Despite their prevalence in the empirical literature, there seem to be some misperceptions regarding how small samples should be accommodated in the statistical analysis. Though intuitively logical and appealing, simple fixes like “Bayes solves that” or “just use bootstrapping” are rarely adequate and can often lead to no discernible improvement in the performance or can ironically result in *worse* performance than taking no action. Though these suggestions are not completely without merit, they do not provide a well-formed analytic strategy without closer scrutiny. Bootstrapping can be useful, but only particular types of bootstraps in certain contexts can be helpful; Bayesian methods are better-suited to handle smaller samples but researchers must have strong prior information to take advantage of these properties.

It is important to note that improved statistical methods with smaller samples address purely statistical issues in estimation of models, trustworthiness in parameter estimates, and to ensure that Type-I error rates are properly controlled. This should not be taken to imply that use of these methods solves research design issues that historically plague studies with smaller samples. For example, use of advanced small sample statistical methods will do very little to help to overcome issues with statistical power that trouble studies with smaller samples. Additionally, no statistical method, no matter how advanced, can remedy sampling issues with smaller samples such as data that do not faithfully represent key properties of the population. Nonetheless, small sample studies are sometimes unavoidable in certain contexts such as studying underrepresented groups, studying students with less common disabilities, or in school-randomized interventions where the administration costs can be quite high. In these circumstances, some small sample issues are beyond the researchers’ control once data are collected. However, researchers can always take steps to assure that the properties of their statistical models are satisfactory.

Additionally, the small sample methods we suggest will not impact analyses on larger samples. That is, if one utilizes the Kenward-Roger correction or the Swain correction with larger datasets, the results will not be appreciably different than if no corrections were used because these corrections converge to maximum likelihood estimates asymptotically. So, if researchers are unsure whether their sample size classifies as a “small sample”, use of these corrective procedures will not harm the analysis. At best, they may yield estimates with more desirable statistical properties, at worst the results will come back identical to those if no correction were used. We hope that this paper has helped clarify when use of small sample methods is required and how to responsibly implement them.

References

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research*, *82*, 436-476. doi:10.3102/0034654312458162

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods*, *18*, 151-164. doi:10.1037/a0030642

Barrett, P. (2007). Structural equation modelling: Adjudging model fit. *Personality and Individual Differences*, *42*, 815-824. doi:10.1016/j.paid.2006.09.018

Bartlett, M. S. (1950). Tests of significance in factor analysis. *British Journal of Statistical Psychology*, *3*, 77-85. doi:10.1111/j.2044-8317.1950.tb00285.x

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., & Ferron, J. M. (2014). How low can you go?. *Methodology*, *10*, 1-11. doi:10.1027/1614-2241/a000062

Bentler, P. M. (2007). On tests and indices for evaluating structural models. *Personality and Individual Differences*, *42*, 825-829. doi:10.1016/j.paid.2006.09.024

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin*, *107*, 238-246. doi:10.1037/0033-2909.107.2.238

Bentler, P. M., & Yuan, K. H. (1999). Structural equation modeling with small samples: Test statistics. *Multivariate Behavioral Research*, *34*, 181-197. doi:10.1207/S15327906Mb340203

Bliese, P. D., & Halverson, R. R. (1998). Group size and measures of group-level properties: An examination of eta-squared and ICC values. *Journal of Management*, *24*, 157-172. doi:10.1177/014920639802400202

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*, 791-799. doi:10.1080/01621459.1976.10480949

Bradley, J. V. (1978). Robustness?. *British Journal of Mathematical and Statistical*

Psychology, *31*, 144-152. doi:10.1111/j.2044-8317.1978.tb00581.x

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis*, *1*, 473-514.

Bryan, M. L., & Jenkins, S. P. (2016). Multilevel modelling of country effects: A cautionary tale. *European Sociological Review*, *32*, 3-22. doi:10.1093/esr/jcv059

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. *The Review of Economics and Statistics*, *90*, 414-427. doi:10.1162/rest.90.3.414

Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology and Community Health*, *62*, 752-758. doi:10.1136/jech.2007.060798

Curran, P. J. (2003). Have multilevel models been structural equation models all along?. *Multivariate Behavioral Research*, *38*, 529-569. doi:10.1207/s15327906mbr3804_5

Curran, P. J., McGinley, J. S., Bauer, D. J., Hussong, A. M., Burns, A., Chassin, L., ... & Zucker, R. (2014). A moderated nonlinear factor model for the development of commensurate measures in integrative data analysis. *Multivariate Behavioral Research, 49*, 214-231.

Curran, P. J., Bollen, K. A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi square distribution in misspecified structural equation models: Finite sample results from a Monte Carlo simulation. *Multivariate Behavioral Research*, *37*, 1-36. doi:10.1080/00273171.2014.889594

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development*, *11*, 121-136. doi:10.1080/15248371003699969

de Winter, J. C. (2013). Using the Student's t-test with extremely small sample sizes. *Practical Assessment, Research & Evaluation, 18*, 1-12.

de Winter, J. C., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics*, *39*, 695-710. doi:10.1080/02664763.2011.610445

de Winter, J. D., Dodou, D. & Wieringa, P. A. (2009). Exploratory factor analysis with small sample sizes. *Multivariate Behavioral Research*, *44*, 147-181. doi:10.1080/00273170902794206

Dedrick, R. F., Ferron, J. M., Hess, M. R., Hogarty, K. Y., Kromrey, J. D., Lang, T. R., ... & Lee, R. S. (2009). Multilevel modeling: A review of methodological issues and applications. *Review of Educational Research*, *79*, 69-102. doi:10.3102/0034654308325581

Depaoli, S., & van de Schoot, R. (2015). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods,* Advance online publication. doi:10.1037/met0000065

Depaoli, S. (2014). The impact of inaccurate "informative" priors for growth parameters in Bayesian growth mixture modeling. *Structural Equation Modeling*, *21*, 239-252. doi:10.1080/10705511.2014.882686

Dunson, D. B. (2001). Commentary: practical advantages of Bayesian analysis of epidemiologic data. *American Journal of Epidemiology*, *153*, 1222-1226. doi:10.1093/aje/153.12.1222

Efron, B., & Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, *1*, 54-75. doi:10.1214/ss/1177013815

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7,* 1-26. doi:10.1214/aos/1176344552

Efron, B., & Hinkley, D. V. (1978). Assessing the accuracy of the maximum likelihood estimator: Observed versus expected Fisher information. *Biometrika*, *65,* 457-482.

doi:10.1093/biomet/65.3.457

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: a new look at an old issue. *Psychological Methods*, *12*, 121-138. doi:10.1037/1082-989X.12.2.121

Fan, X., & Sivo, S. A. (2007). Sensitivity of fit indices to model misspecification and model types. *Multivariate Behavioral Research*, *42*, 509-529. doi:10.1080/00273170701382864

Fan, X., & Sivo, S. A. (2005). Sensitivity of fit indexes to misspecified structural or measurement model components: Rationale of two-index strategy revisited. *Structural Equation Modeling*, *12*, 343-367. doi:10.1207/s15328007sem1203_1

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods*, *41*, 372-384. doi:10.3758/BRM.41.2.372

Field, C. A., & Welsh, A. H. (2007). Bootstrapping clustered data. *Journal of the Royal Statistical Society: Series B*, *69*, 369-390. doi:10.1111/j.1467-9868.2007.00593.x

Flynn, T. N., & Peters, T. J. (2004). Use of the bootstrap in analysing cost data from cluster randomised trials: some simulation results. *BMC Health Services Research*, *4*, doi:10.1186/1472-6963-4-33

Fouladi, R. T. (2000). Performance of modified test statistics in covariance and correlation structure analysis under conditions of multivariate nonnormality. *Structural Equation Modeling*, *7*, 356-410. doi:10.1207/S15328007SEM0703_2

Fujikoshi, Y. (2000). Transformations with improved chi-squared approximations. *Journal of Multivariate Analysis*, *72*, 249-263. doi:10.1006/jmva.1999.1854

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). *Bayesian data analysis*. Boca Raton, FL: Chapman & Hall.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*, 515-534.

Glockner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, *10*, 544-565. doi:10.1207/S15328007SEM1004_4

Hancock, G. R., & Mueller, R. O. (2011). The reliability paradox in assessing structural relations within covariance structure models. *Educational and Psychological Measurement*, *71*, 306-324. doi:10.1177/0013164410384856

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*, 60-87. doi:10.3102/0162373707299706

Heene, M., Hilbert, S., Draxler, C., Ziegler, M., & Bühner, M. (2011). Masking misfit in confirmatory factor analysis by increasing unique variances: a cautionary note on the usefulness of cutoff values of fit indices. *Psychological Methods*, *16*, 319-336. doi:10.1037/a0024917

Herzog, W., & Boomsma, A. (2009). Small-sample robust estimators of noncentrality-based and incremental model fit. *Structural Equation Modeling*, *16*, 1-27. doi:10.1080/10705510802561279

Herzog, W., Boomsma, A., & Reinecke, S. (2007). The model-size effect on traditional and modified tests of covariance structures. *Structural Equation Modeling*, *14*, 361-390. doi:10.1080/10705510701301602

Hesterberg, T., Moore, D. S., Monaghan, S., Clipson, A., & Epstein, R. (2005). Bootstrap methods and permutation tests. *Introduction to the Practice of Statistics*, *5*, 1-70.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling an overview and a meta-analysis. *Sociological Methods & Research*, *26*, 329-367.

doi:10.1177/0049124198026003003

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modeling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*, 53–60.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Hox, J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods*, *6*, 87-93.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1-55. doi:10.1080/10705519909540118

Huang, F. L. (2017). Using cluster bootstrapping to analyze nested data with a few clusters. *Educational and Psychological Measurement*, advance online publication, doi:10.1177/0013164416678980.

Iacobucci, D. (2010). Structural equations modeling: Fit indices, sample size, and advanced topics. *Journal of Consumer Psychology, 20*, 90-98. doi:10.1016/j.jcps.2009.09.003

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Hillsdale, NJ: Erlbaum

Jöreskog, K. G., & Sörbom, D. (1981). *LISREL V: Analysis of linear structural relationships by the method of maximum likelihood*. Chicago: National Educational Resources

Kang, Y., McNeish, D., & Hancock, G. R. (2016). The role of measurement quality on practical guidelines for assessing measurement and structural invariance. *Educational and Psychological Measurement*, *76*, 533-561. doi:10.1177/0013164415603764

Kenny, D. A., & McCoach, D. B. (2003). Effect of the number of variables on measures of fit in structural equation modeling. *Structural Equation Modeling*, *10*, 333-351. doi:10.1207/S15328007SEM1003_1

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics, 53*, 983-997. doi:10.2307/2533558

Kim, K. H. (2005). The relation among fit indexes, power, and sample size in structural equation modeling. *Structural Equation Modeling*, *12,* 368-390. doi:10.1207/s15328007sem1203_2

Kish, L. (1965). *Survey Sampling*. New York: Wiley.

Kline, R. B. (2013). *Principles and practice of structural equation modeling*. New York, NY: Guilford.

Kolenikov, S. (2010). Resampling variance estimation for complex survey data. *Stata Journal*, *10,* 165-199.

Kruschke J.K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Burlington, MA: Academic Press

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, *39,* 653-686. doi:10.1207/s15327906mbr3904_4

MacCallum, R. C., & Austin, J. T. (2000). Applications of structural equation modeling in psychological research. *Annual Review of Psychology*, *51*, 201-226. doi:10.1146/annurev.psych.51.1.201

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research*, *36,* 611-637. doi:10.1207/S15327906MBR3604_06

Markland, D. (2007). The golden rule is that there are no golden rules: A commentary on Paul Barrett's recommendations for reporting model fit in structural equation modelling. *Personality and Individual Differences*, *42,* 851-858. doi:10.1016/j.paid.2006.09.023

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cutoff values for fit indexes and dangers in overgeneralizing Hu and Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320-341. doi:10.1207/s15328007sem1103_2

McNeish, D. (2017a). Exploratory factor analysis with small samples and missing data. *Journal of Personality Assessment,* Advance online publication, doi:10.1080/00223891.2016.1252382

McNeish, D. (2017b). Small sample methods for multilevel modeling: A colloquial elucidation of REML and the Kenward-Roger correction. *Multivariate Behavioral Research,* Advance online publication, doi:10.1080/00273171.2017.1344538

McNeish, D. (2016a). On using Bayesian methods to address small sample problems. *Structural Equation Modeling*, *23*, 750-773. doi:10.1080/10705511.2016.1186549

McNeish, D. (2016b). Using data-dependent priors to mitigate small sample bias in latent growth models: A discussion and illustration using M*plus*. *Journal of Educational and Behavioral Statistics*, *41*, 27-56. doi:10.3102/1076998615621299

McNeish, D.(2014). Modeling sparsely clustered data: Design-based, model-based, and single-level methods. *Psychological Methods*, *19*, 552-563. doi:10.1037/met0000024

McNeish, D., & Harring, J. R. (2017). Correcting model fit criteria for small sample latent growth models with incomplete data. *Educational and Psychological Measurement,* advance online publication, doi:10.1177/0013164416661824.

McNeish, D., Stapleton, L.M., & Silverman, R.D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 22*, 114-140. doi:10.1037/met0000078

McNeish, D., & Stapleton, L. M. (2016a). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review*, *28*, 295-314. doi:10.1007/s10648-014-9287-x

McNeish, D. & Stapleton, L.M. (2016b). Modeling clustered data with very few clusters.

Multivariate Behavioral Research, 51, 495-518. doi:10.1080/00273171.2016.1167008

Miles, J., & Shevlin, M. (2007). A time and a place for incremental fit indices. *Personality and Individual Differences*, *42*, 869-874. doi:10.1016/j.paid.2006.09.022

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological Methods*, *17*, 313-335. doi:10.1037/a0026802

Muthén, L. K., & Muthén, B. O. (2012). *Mplus user's guide* (7th ed.). Los Angeles: Muthén & Muthén.

Muthén, L. K., & Muthén, B. O. (2002). How to use a Monte Carlo study to decide on sample size and determine power. *Structural Equation Modeling*, *9*, 599-620. doi:10.1207/S15328007SEM0904_8

Nevitt, J., & Hancock, G. R. (2004). Evaluating small sample approaches for model test statistics in structural equation modeling. *Multivariate Behavioral Research*, *39*, 439-478. doi:10.1207/S15327906MBR3903_3

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Ren, S., Lai, H., Tong, W., Aminzadeh, M., Hou, X., & Lai, S. (2010). Nonparametric bootstrapping for hierarchical data. *Journal of Applied Statistics*, *37*, 1487-1498. doi:10.1080/02664760903046102

Roberts, B. W., & Del Vecchio, W. F. (2000). The rank-order consistency of personality traits from childhood to old age: a quantitative review of longitudinal studies. *Psychological Bulletin*, *126*, 3-25. doi:10.1037/0033-2909.126.1.3

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: a meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*, 1-25. doi:10.1037/0033-2909.132.1.1

Russell, D. W. (2002). In search of underlying dimensions: The use (and abuse) of factor analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin*, *28*, 1629-1646. doi:10.1177/014616702237645

Saris, W. E., Satorra, A., & Van der Veld, W. M. (2009). Testing structural equation models or detection of misspecifications?. *Structural Equation Modeling*, *16*, 561-582. doi:10.1080/10705510903203433

Satorra, A., & Bentler, P. M. (2001). A scaled difference chi-square test statistic for moment structure analysis. *Psychometrika*, *66*, 507-514. doi:10.1007/BF02296192

Savalei, V. (2012). The relationship between root mean square error of approximation and model misspecification in confirmatory factor analysis models. *Educational and Psychological Measurement*, *72*, 910-932. doi:10.1177/0013164412452564

Savalei, V. (2010). Small sample statistics for incomplete nonnormal data: Extensions of complete data formulae and a Monte Carlo comparison. *Structural Equation Modeling*, *17*, 241-264. doi:10.1080/10705511003659375

Scherbaum, C. A., & Ferreter, J. M. (2009). Estimating statistical power and required sample sizes for organizational research using multilevel modeling. *Organizational Research Methods*, *12*, 347-367. doi:10.1177/1094428107308906

Schuurman, N. K., Grasman, R. P. P. P., & Hamaker, E. L. (2016). A comparison of inverse wishart prior specifications for covariance matrices in multilevel autoregressive models. *Multivariate Behavioral Research*, *51*, 185-206. doi:10.1080/00273171.2015.1065398

Sherman, M., & Cessie, S. L. (1997). A comparison between bootstrap methods and generalized estimating equations for correlated outcomes in generalized linear models. *Communications in Statistics-Simulation and Computation*, *26*, 901-925.

doi:10.1080/03610919708813417

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

Snijders, T. A., & Bosker, R. J. (1993). Standard errors and sample sizes for two-level research. *Journal of Educational Statistics*, *18*, 237-259. doi:10.2307/1165134

Spiegelhalter, D., Myles, J. P., Jones, D. R., & Abrams, K. R. (1999). An introduction to Bayesian methods in health technology assessment. *British Medical Journal, 319*, 508-512.

Stapleton, L. M., Pituch, K. A., & Dion, E. (2015). Standardized effect size measures for mediation analysis in cluster-randomized trials. *The Journal of Experimental Education*, *83,* 547-582. doi:10.1080/00220973.2014.919569

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science*, *57*, 748-761. doi:10.1111/ajps.12001

Steiger, J. H. (2007). Understanding the limitations of global fit assessment in structural equation modeling. *Personality and Individual Differences*, *42*, 893-898. doi:10.1016/j.paid.2006.09.017

Steiger, J. H., & Lind, J. C. (1980, May). Statistically based tests for the number of common factors. Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA

Swain, A. J. (1975). Analysis of parametric structures for variance matrices (Unpublished doctoral dissertation). Department of Statistics, University of Adelaide, Australia.

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with" well fitting" models. *Journal of Abnormal Psychology*, *112,* 578-598. doi:10.1037/0021-843X.112.4.578

van de Schoot, R., Winter, S., Ryan, O., Zondervan-Zwijnenburg, M. & Depaoli, S. (2017). A systematic review of Bayesian papers in psychology: The last 25 years. *Psychological Methods,* Advance Online Publication, doi:10.1037/met0000100.

van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: the case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology*, *6*. doi:10.3402/ejpt.v6.25216

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development, 85*, 842-860. doi:10.1111/cdev.12169

van de Schoot, R., Lugtig, P., & Hox, J. (2012). A checklist for testing measurement invariance. *European Journal of Developmental Psychology*, *9*, 486-492. doi:10.1080/17405629.2012.686740

van der Leeden, R., Meijer, E., & Busing, F. M. T. A. (2008). Resampling multilevel models. In J. de Leeuw & E. Meijer (Eds.), Handbook of multilevel analysis (pp. 401–433). New York: Springer.

Vickers, A. J. (2003). How many repeated measures in repeated measures designs? Statistical issues for comparative trials. *BMC Medical Research Methodology*, *3*, 22. doi:10.1186/1471-2288-3-22

Wang, J., Carpenter, J. R., & Kepler, M. A. (2006). Using SAS to conduct nonparametric residual bootstrap multilevel modeling with a small number of groups. *Computer Methods and Programs in Biomedicine, 82*, 130-143. doi:10.1016/j.cmpb.2006.02.006

What Works Clearinghouse. (2013). *Procedures and standards handbook* (Version 3).. US Department of Education. Washington, DC: Institute of Education Sciences.

Wolf, E. J., Harrington, K. M., Clark, S. L., & Miller, M. W. (2013). Sample size requirements for structural equation models an evaluation of power, bias, and solution propriety. *Educational and Psychological Measurement, 73*, 913-934. doi:10.1177/0013164413495237

Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. T*he Annals of Statistics, 14*, 1261-1295. doi:10.1214/aos/1176350142

Yuan, K. H. (2005). Fit indices versus test statistics. *Multivariate Behavioral Research*, *40*, 115-148. doi:10.1207/s15327906mbr4001_5

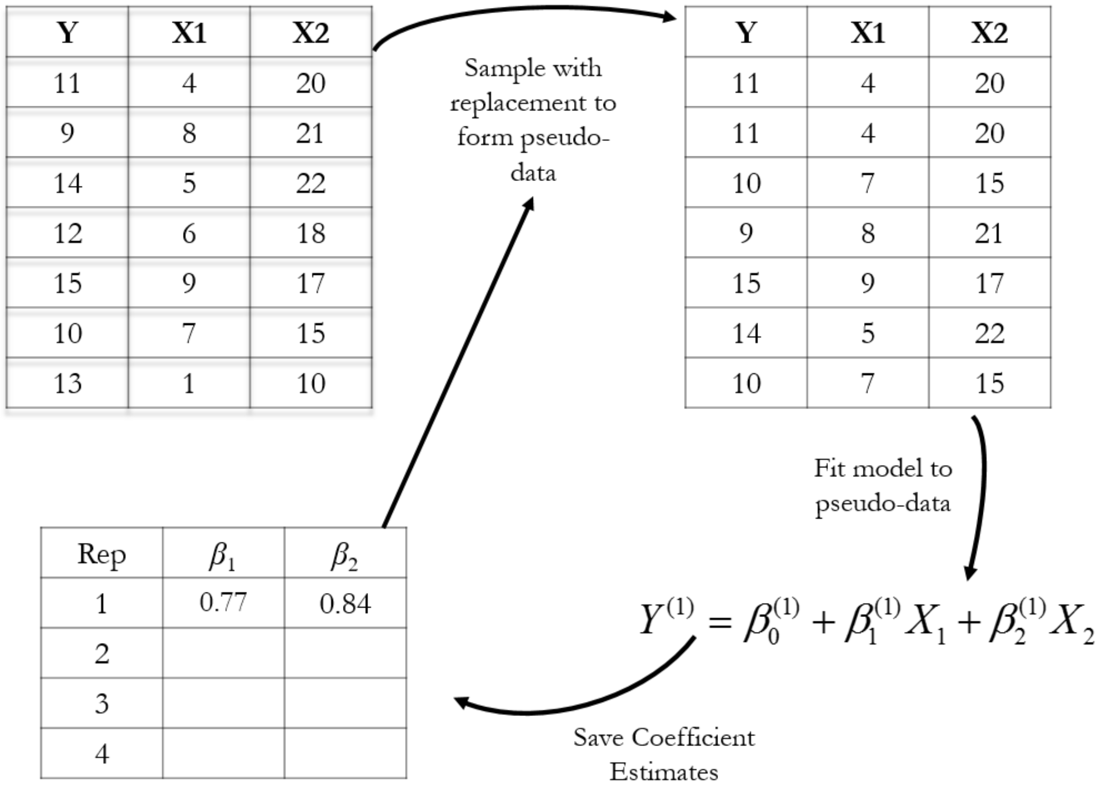
Yuan, K. H., & Bentler, P. M. (1997). Mean and covariance structure analysis: Theoretical and practical improvements. *Journal of the American statistical association*, *92*, 767-774. doi:10.1080/01621459.1997.10474029

Yuan, K. H., Tian, Y., & Yanagihara, H. (2015). Empirical correction to the likelihood ratio statistic for structural equation modeling with many variables. *Psychometrika*, *80,* 379-405. doi:10.1007/s11336-013-9386-5

Table 1

Summary of supporting arguments, counterarguments, and alternative suggested methods

	Bootstrapping Multilevel Data	Using T_{ML} to Assess Model Fit	Switching to Bayesian Estimation
Common Supporting Arguments	• Bootstrapping does not require asymptotic assumptions about the shape of the sampling distribution • Fisher Information is not required to compute sampling variability	• Overpowered argument does not apply with small samples • Test statistic has constant interpretation • Assessment of good or poor fit is not as reliant on model type, size, or factor loadings, etc.	• Sampling variability is based on the posterior distribution, not on asymptotic methods • Prior distributions can be used to augment the limited amount of information contained in the data
Arguments Against Conventional Recommendations	• Standard bootstrapping procedures use resampling • With small multilevel datasets, pseudo-datasets may not be representative of the original data • Underestimates sampling variability	• The test statistic is chi-square distributed asymptotically, but not with small samples • *p*-values are not trustworthy • Well-fitting models are much more likely to be rejected and deemed poorly fitting with small samples	• Bayesian methods are theoretically equipped to better handle small samples, but the necessary precautions are not taken in common implementations • Diffuse priors with small samples often led to more biased estimates that are less likely to detect effects
Suggestions for Best Practice	• Likelihood corrections like REML or Kenward-Roger • Alternative bootstrapping algorithms that do not resample such as wild bootstrapping	• Heuristic small sample corrections such as those by Swain, Yuan, or Bartlett • Small sample test statistics, most of which were developed by Yuan and Bentler	• Prior information is needed to in order to capitalize of Bayes' advantageous small sample properties • Otherwise, small sample frequentist methods can often yield more desirable statistical properties



Y	X1	X2
11	4	20
9	8	21
14	5	22
12	6	18
15	9	17
10	7	15
13	1	10

Sample with replacement to form pseudo-data

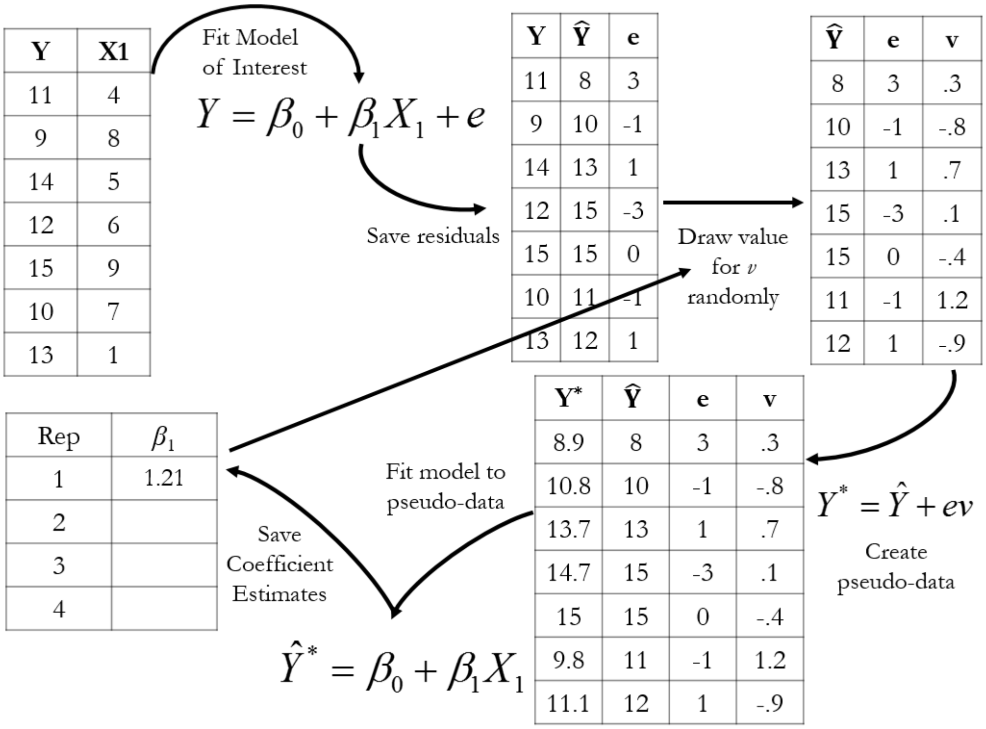
Y	X1	X2
11	4	20
11	4	20
10	7	15
9	8	21
15	9	17
14	5	22
10	7	15

$$Y^{(1)} = \beta_0^{(1)} + \beta_1^{(1)} X_1 + \beta_2^{(1)} X_2$$

Save Coefficient Estimates

Rep	β_1	β_2
1	0.77	0.84
2		
3		
4		

Figure 1. Algorithm for standard bootstrap that resamples the data. For simplicity, the figure shows single level data. For multilevel data, instead of selecting individuals, entire clusters would be sampled.



Y	X1
11	4
9	8
14	5
12	6
15	9
10	7
13	1

Y	$\hat{Y}$	e
11	8	3
9	10	-1
14	13	1
12	15	-3
15	15	0
10	11	-1
13	12	1

$\hat{Y}$	e	v
8	3	.3
10	-1	-.8
13	1	.7
15	-3	.1
15	0	-.4
11	-1	1.2
12	1	-.9

Y*	$\hat{Y}$	e	v
8.9	8	3	.3
10.8	10	-1	-.8
13.7	13	1	.7
14.7	15	-3	.1
15	15	0	-.4
9.8	11	-1	1.2
11.1	12	1	-.9

Rep	β_1
1	1.21
2	
3	
4	

Figure 2. Algorithm for wild bootstrap that creates pseudo-data by perturbing the residuals rather than resampling data. For simplicity, the figure shows single level data.

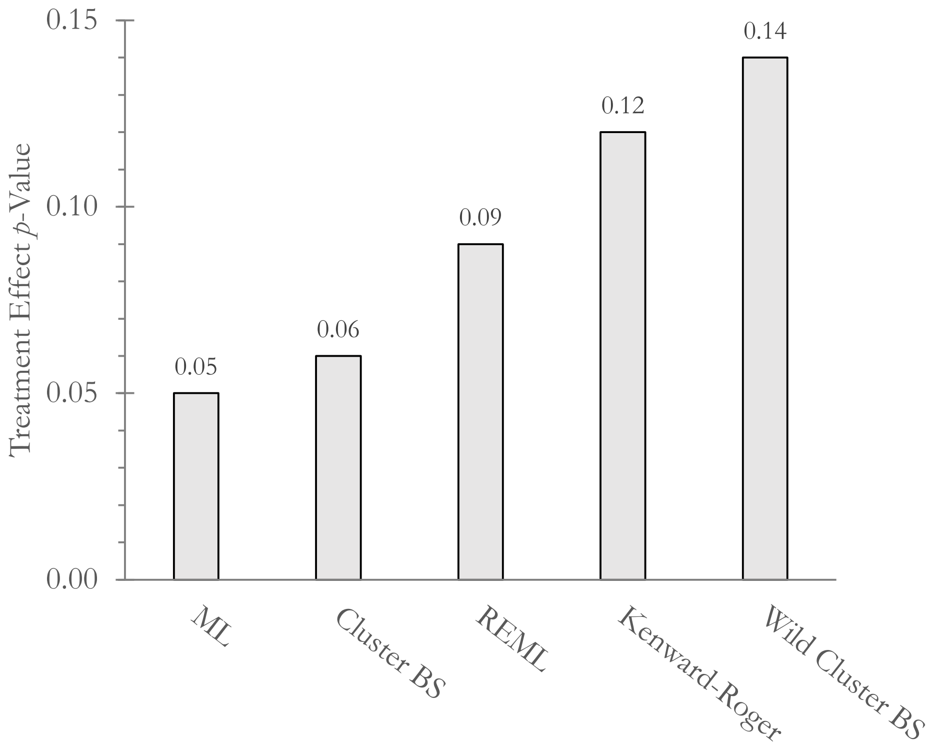


Figure 3. Comparison of *p*-values for the treatment effect of the Head Start data using 5 different methods. Methods on the horizontal-axis are arranged by magnitude of the *p*-value. The estimated treatment effect was constant across methods at 2.76. ML = Maximum Likelihood, BS = Bootstrap, REML = Restricted Maximum Likelihood, Kenward-Roger = Restricted Maximum Likelihood with a Kenward-Roger Correction.

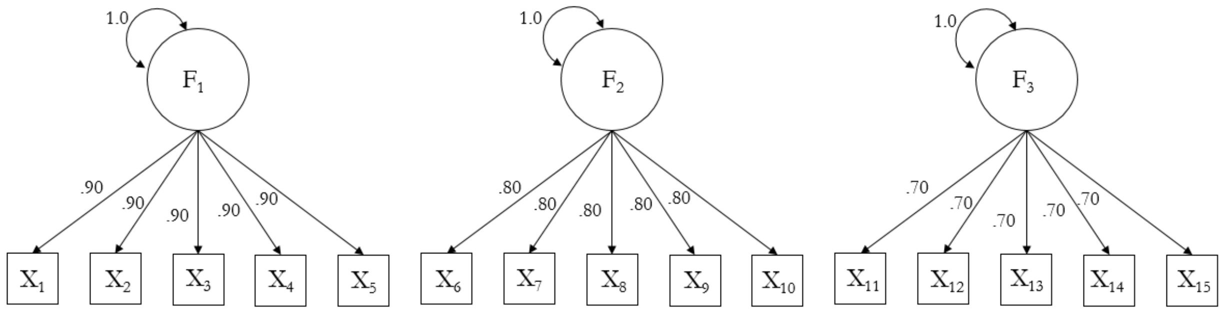


Figure 4. Data generation model for CFA simulation. Factor loadings are in a standardized metric. Residual variances are not shown but are equal to 1 – (Factor Loading)^2. Factor variances are set to 1 and factors are uncorrelated. There is no mean structure, no cross-loadings, and no residual covariances.

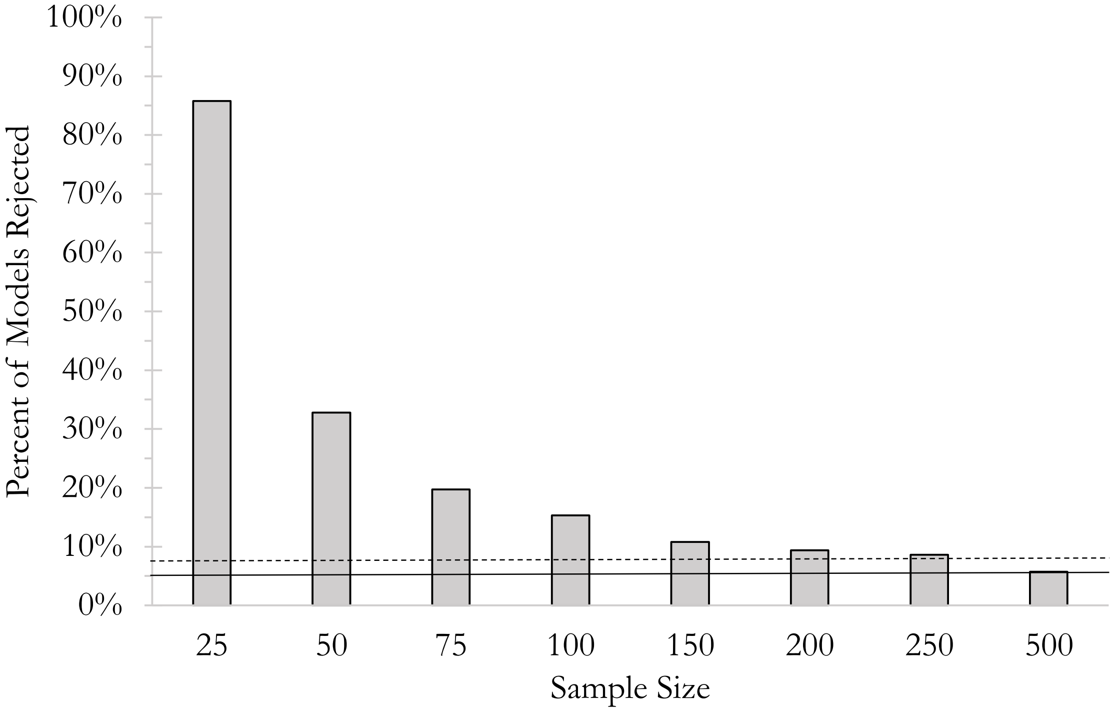


Figure 5. Percentage of truly perfect fitting CFA models that are rejected, by sample size. The solid horizontal line is the nominal 5% rate. The dashed black horizontal line represents 7.5% which is the maximum tolerable rate after considering random fluctuations (Bradley, 1978). The respective percentages are 85.8, 32.8, 19.7, 15.3, 10.8, 9.4, 8.6, and 5.7.

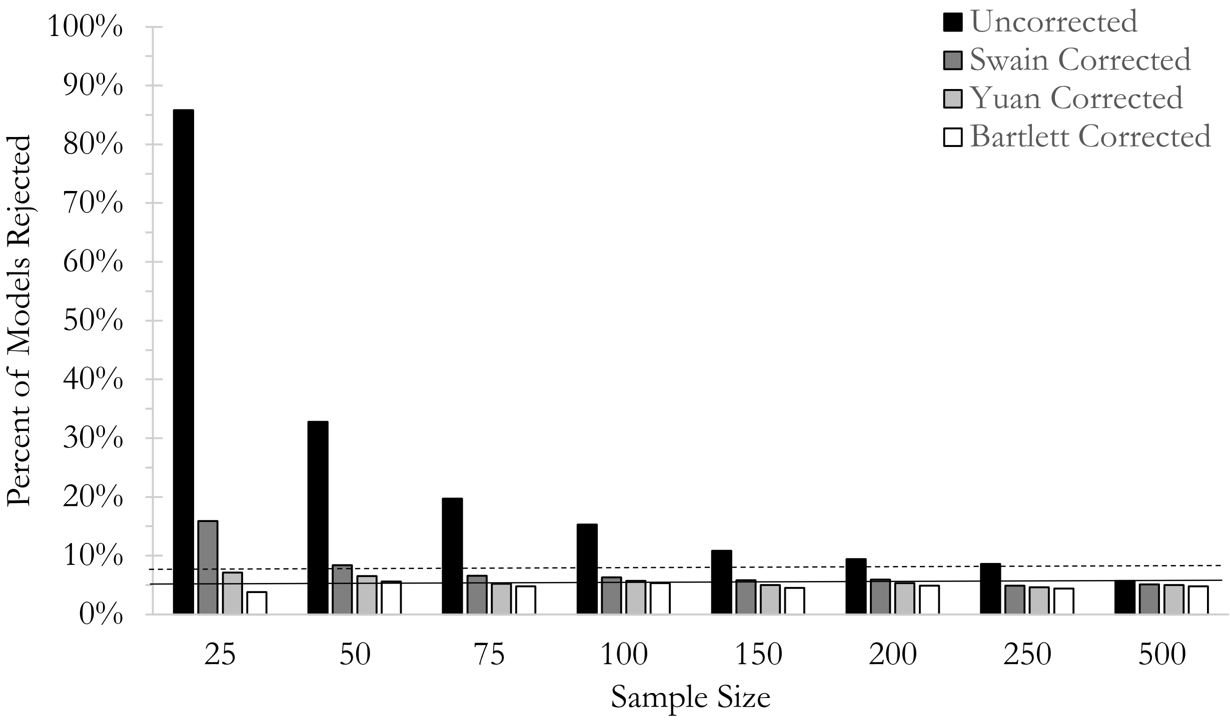


Figure 6. Percentage of perfect CFA models that are rejected, by sample size using uncorrected tests and three different small sample corrections (Swain, Yuan, and Bartlett). The solid horizontal line is the nominal 5% rate. The dashed black horizontal line represents 7.5% which is the maximum tolerable rate after considering random fluctuation (Bradley, 1978).

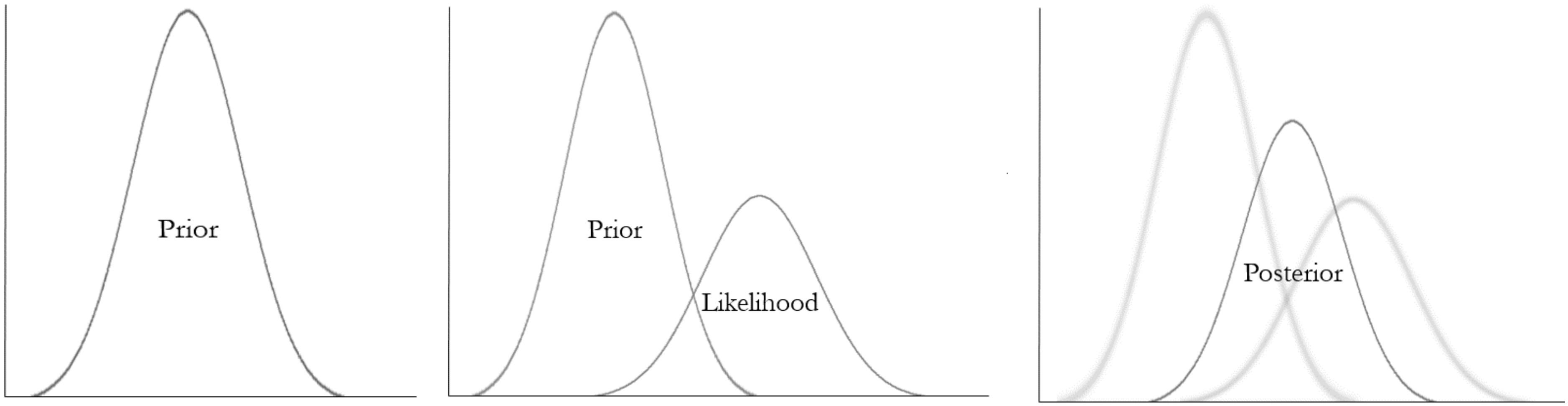


Figure 7. Conceptual diagram of how the Bayesian posterior distribution is formed. The prior is specified first before inspecting the data, information from the data is contained in the likelihood, then the likelihood and the posterior are combined to yield the posterior. The summary statistics of the posterior distribution are conceptually related to quantities estimated in the frequentist framework.

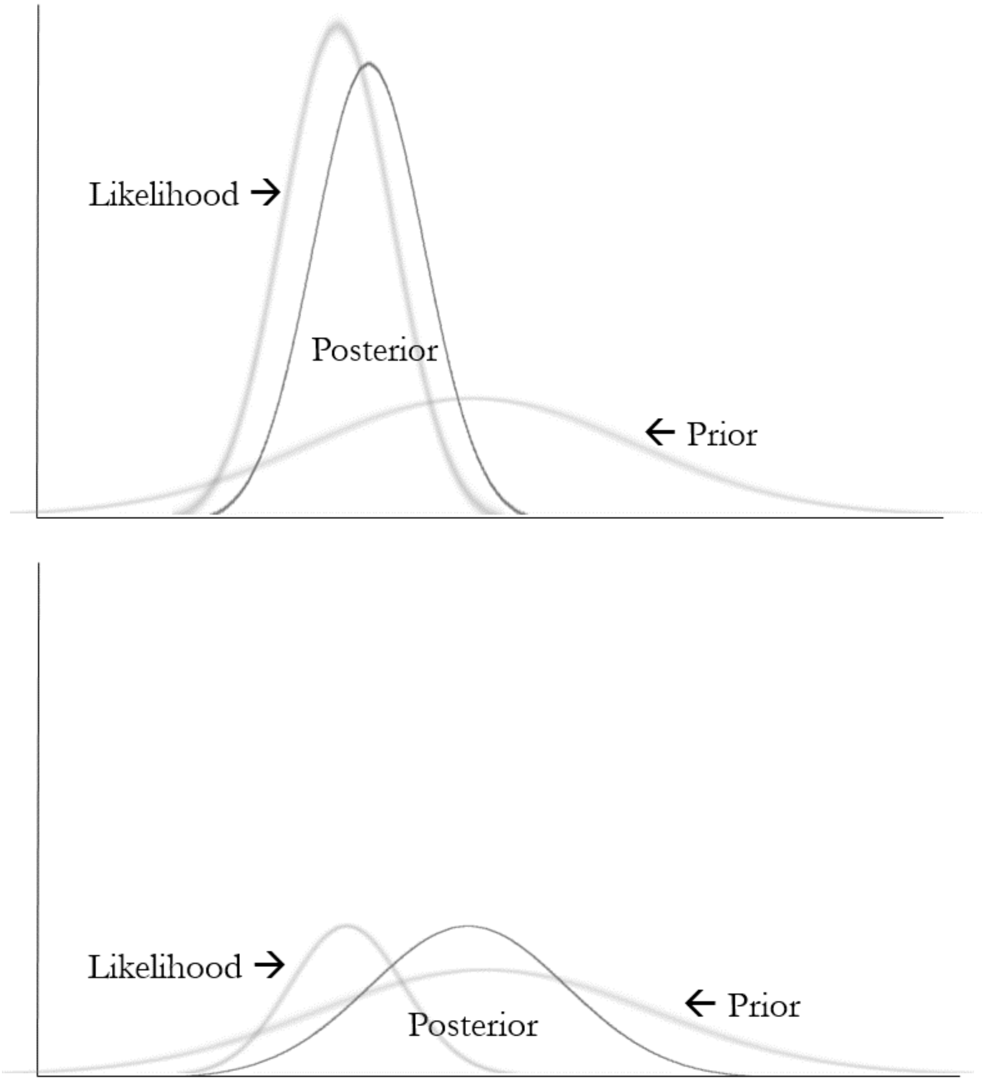


Figure 8. Comparison of the influence of diffuse priors for large samples (top panel) and small samples (bottom panel). With large samples, the likelihood dominates the prior so the posterior is strongly influenced by the likelihood. In small samples, the likelihood does not contain enough information to overtake the prior, so the prior plays a large role in forming the posterior. The uncertainty of the diffuse prior results in a posterior with more variability than contained in the likelihood and leads to a posterior that is less faithful to the data.

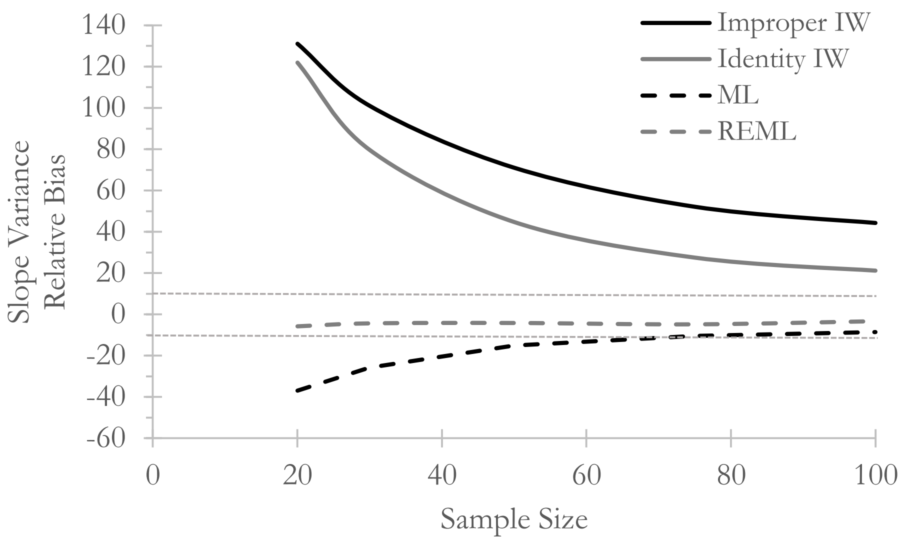


Figure 9a. Plot of relative biases of the slope variance for two Bayesian diffuse priors and two likelihood methods for sample sizes between 20 and 100. IW= Inverse Wishart, ML = Maximum Likelihood, REML = Restricted Maximum Likelihood. Small dashed grey lines indicate -10% and +10% relative bias; estimates between the two small dashed lines are considered acceptable.

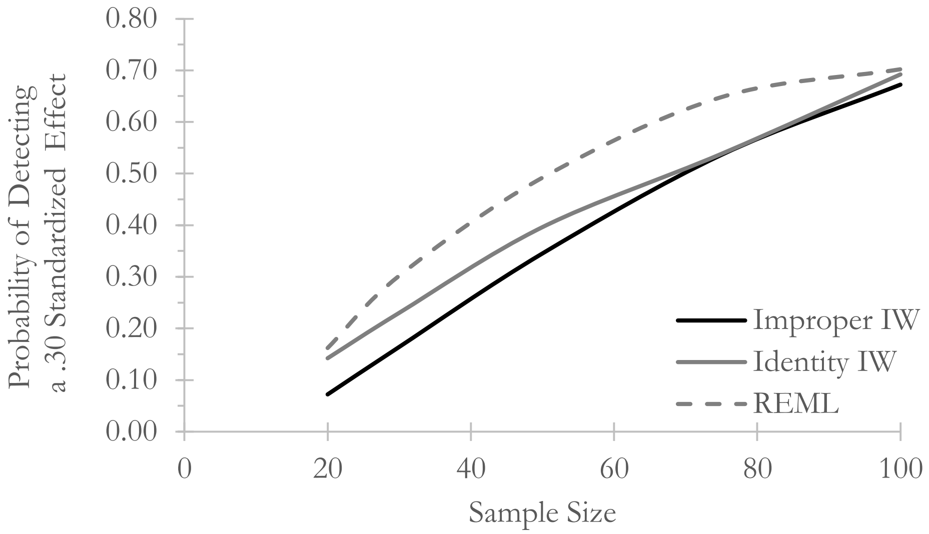


Figure 9b. Probability to detect a time-invariant predictor with a standardized coefficient of 0.30 for sample sizes between 20 and 100. IW= Inverse Wishart, REML = Restricted Maximum Likelihood. Maximum Likelihood is not shown because the inflated Type-I error rates distort interpretation of power.