Research Quarterly for Exercise and Sport

Vol. 82, No. 3, pp. 412–423

Sample Size and Power Estimates for a Confirmatory Factor Analytic Model in Exercise and Sport: A Monte Carlo Approach

Nicholas D. Myers, Soyeon Ahn, and Ying Jin

Monte Carlo methods can be used in data analytic situations (e.g., validity studies) to make decisions about sample size and to estimate power. The purpose of using Monte Carlo methods in a validity study is to improve the methodological approach within a study where the primary focus is on construct validity issues and not on advancing statistical theory. The purpose of this study is to demonstrate how Monte Carlo methods can be used to determine sample size and to estimate power for a confirmatory factor analytic model under model-data conditions commonly encountered in exercise and sport. Because the purpose is pursued by way of demonstration with the Coaching Efficacy Scale II–High School Teams, related sample size recommendations are provided: N≥ 200 for the theoretical model; N≥ 300 for the population model. Technical terms (e.g., coverage) are defined when necessary.

Key words: simulation, misspecification, ordinal, coaching efficacy

Rules of thumb for determining adequate sample size (N) are known to be of limited use in achieving an acceptable likelihood for desirable empirical outcomes (e.g., model convergence, statistical precision, statistical power) for a particular application of confirmatory factor analysis (CFA) with real data (Marsh, Hau, Balla, & Grayson, 1998). Common rules of thumb for determining adequate N for a particular application of CFA include, but are not limited to: $N \geq 200$, ratio of N to the number of variables in a model (p), $N/p \geq 10$; the ratio of N to the number of model parameters (q), $N/q \geq 5$; and an inverse relationship between construct reliability and adequate N. Even when model-data assumptions are made that are rarely observed in practice and simulated data are analyzed, the performance of these rules of thumb has limited the ability of methodologists to offer definitive guidelines for adequate N across the myriad of model-data conditions observed in practice (Gagné & Hancock, 2006; Jackson 2001, 2003). The core problem with these rules of thumb is that adequate N for CFA depends on many factors that typically vary across any two studies using real data and inexact theoretical models (e.g., distribution of variables, reliability of indictors, size of the model, degree of model misspecification). These factors can be directly modeled using Monte Carlo methods.

"Monte Carlo methods use random processes to estimate mathematical or physical quantities, to study distributions of random variables, to study and compare statistical procedures, and to study complex systems" (Gentle, 2005, pp. 1264–1265). Suppose that responses to a set of items are viewed as fallible reflective indicators of continuous latent variables within a CFA model (see Figure 1 for an application from exercise and sport). A population model for the indicators could be specified (see "Design Stage: Conceptualizing the Experiment" in the Method section for a worked example) and, using Monte Carlo methods, random samples of a particular size could be repeatedly

Submitted: October 19, 2009
Accepted: August 28, 2010

Nicholas D. Myers, Soyeon Ahn, and Ying Jin are with the Department of Educational and Psychological Studies at the University of Miami.

drawn from the population distribution (see "Generating Data Stage: Performing the Experiment" in the Method section for a worked example). Parameters of interest could be estimated in each random sample, and these could be combined to form an empirically generated sampling distribution for each parameter of interest, with the results summarized across replications (see Results section for a worked example). Monte Carlo methods, then, can be thought of as flexible experimental tools that can be used to artificially create (and hence study) theoretical (and hence unobserved) sampling distributions. The generality of this approach has allowed for two types of application: in studies of statistical methods and in data analysis (Gentle, 2003). This study fits within the second type of application. From this point forward the expression Monte Carlo methods is used when referring to the methodology in general, regardless of the type of application within which the methodology is applied.

Monte Carlo studies of statistical methods are used to advance statistical theory by testing the ability of particular quantitative methods to recover given population values under various conditions. As summarized by Bandalos (2006), the performance of structural equation modeling has often been investigated by manipulating various independent variables, including model type (e.g., CFA vs. latent variable path model), model size (e.g., number of observed variables), model complexity (e.g., number of parameters estimated), parameter values, sample size, level of nonnormality, and estimation method. Typical outcomes include characteristics of parameter estimates (e.g., bias and efficiency), relative standard error bias, and model-data fit indexes (e.g., η^2, root mean square error of approximation [RMSEA]). Because the primary goal of Monte Carlo studies of statistical methods is to advance statistical theory, and because only so many conditions can be manipulated in any one study, some conditions that may rarely be observed in practice are sometimes assumed (e.g., the population model is known, the data are continuous, all pattern coefficients are equal and large). Imposing such assumptions may result in findings of limited applicability to researchers who apply quantitative methods to real data (MacCallum, 2003). MacCallum argued that the models used in practice (i.e., theoretical model), at best, only approximate a more complex reality (i.e., population model) and that studies that integrate a reasonable degree of misspecification, via Monte Carlo methods in particular, may be more useful to researchers who apply quantitative methods to real data.

Monte Carlo methods can be used in data analysis (e.g., validity studies) to decide on *N* and to estimate power (π) for a particular application of quantitative methods (Muthén & Muthén, 2002). Specifically, two applied questions, "What *N* do I need to achieve a particular π level?" and "How much π will I have with a fixed *N*?" can be investigated. Few conditions may be manipulated because the primary goal of using Monte Carlo methods within a

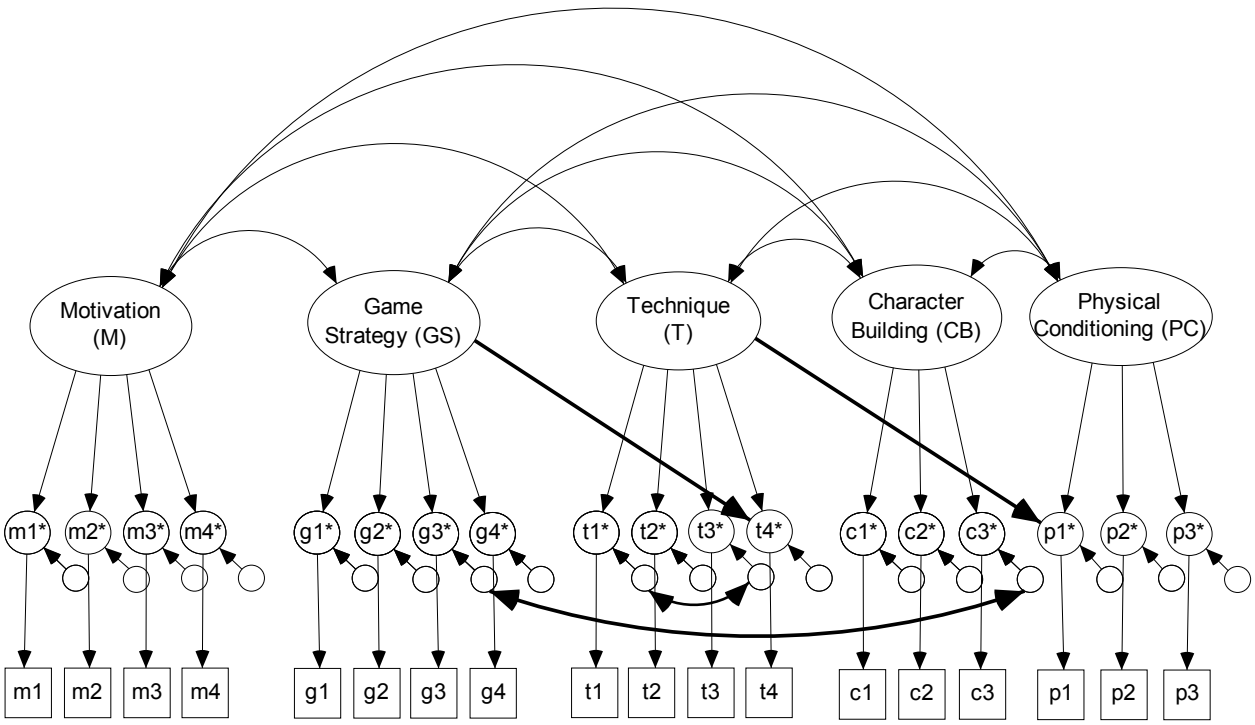

Figure 1. Population and theoretical measurement model for the Coaching Efficacy Scale II–High School Teams. The two bolded paths and the two bolded variances are fixed to = 0 in the theoretical model.

validity study is to improve the methodological approach within a particular study (e.g., providing a context-specific and empirically based rationale for how the necessary *N* for a desired level of π was determined) where the primary focus is on construct validity issues (e.g., evaluating, with a sufficient level of π, empirical evidence for a theoretically based a priori factor structure of a latent construct). The contribution of using Monte Carlo methods within a validity study, then, is toward improved application of a particular quantitative methodology within a substantive area and not to statistical theory.

A definition of π provided by Cohen (1988) is the probability of rejecting a truly false null hypothesis, H_0. A commonly used desired level of π is .80. Determining *N* for a CFA model based on a desired level of π is typically more useful than relying on commonly used rules of thumb, because the *N* needed for a particular application is influenced by many factors that can be directly modeled with Monte Carlo methods (Muthén & Muthén, 2002).

The conditions manipulated in the instructional paper by Muthén and Muthén (2002), in relation to a CFA model with continuous indicators, were the nonnormality of indicators and missing data. The CFA model itself consisted of two continuous latent variables (with equal variances) with five continuous indicators each (each of which had a .80 pattern coefficient). The focus was on answering the two applied questions in the previous paragraph in regard to a .25 correlation between the latent variables (i.e., the parameter of interest). Other model parameters, such as pattern coefficients, were not specified as parameters of interest.

A key assumption of Muthén and Muthén (2002) was that the population model and the theoretical model were identical. Thus the two applied questions were answered under the scenario where the theoretical model is exactly correct—a condition rarely, if ever, met in practice (MacCallum, 2003). Adopting this assumption, however, allowed for certain empirical criteria to be met before investigating the two applied questions regarding *N* and π. One criterion was that parameter estimate and standard error estimate bias did not exceed |10%|. A second criterion was that standard error estimate bias for the parameter of interest did not exceed |5%|. Bias is a systematic difference between a sample estimate and the relevant population value. A third criterion was that coverage remained between .91 and .98. Coverage is the proportion of replications for which the 95% confidence interval contains the parameter value. After meeting these three conditions, the two applied questions were explored. For reasons that will be detailed in the Methods section, this strategy will be altered in this study due to the inclusion of a level of model misspecification commonly observed in exercise and sport.

An informal review of volume 79 of *Research Quarterly for Exercise and Sport* (*RQES*) by the lead author suggested it is common for at least one CFA model to appear in a validity study. Some observed trends (exceptions exist) across these papers include single population models; many observed variables and multiple latent variables; fixing a number of pattern coefficients to satisfy theoretical identification requirements; ordinal data; evidence for approximate model-data fit (e.g., RMSEA ≤ .05) and against exact model-data fit (i.e., a statistically significant χ^2 value); variable pattern coefficients, factor variances, and covariances; a stronger focus on the pattern coefficient matrix, Λ, and the off-diagonal elements of the factor covariance matrix, ψ, than on the measurement error covariance matrix, Θ; and few multidimensional items and/or covariance between measurement errors. Rarely was an a priori plan for *N*, for a desired level of π, communicated. Rarely was an estimate of π, for a fixed *N*, communicated.

Muthén and Muthén (2002) did not provide an example of CFA models with ordinal data. Ordered categorical data (e.g., Likert-type scale) are common in exercise science and are nonnormal by definition due to the discrete nature of the metric (Muthén, 1984). Normal theory (NT) estimators (i.e., those most commonly used in structural equation modeling) assume that the data follow a conditional multivariate normal (MVN) distribution in the population. As reviewed by Finney and DiStefano (2006), violating the assumption of MVN with categorical data can produce untrustworthy results (e.g., inflated indices of model-data misfit, negatively biased parameter estimates, and negatively biased standard errors). The probability of observing untrustworthy results when categorical data are modeled with a NT estimator depends strongly on the degree of nonnormality and the number of ordered response options (e.g., DiStefano, 2002; Dolan, 1994; Muthén & Kaplan, 1985). In cases where the number of response options is less than five, Finney and DiStefano suggest using categorical variable methodology (CVM) with weighted least squares mean- and variance-adjusted estimation (WLSMV; Muthén, 1993). Modeling ordinal data under CVM, instead of as normal and continuous (even with a correction for nonnormality), can correct for attenuation in parameter estimates that may result due to the coarseness of the data.

The WLSMV estimator is similar to, but less computationally demanding than, the asymptotically distribution-free (ADF; Browne, 1984) estimator. The WLSMV estimator has generally outperformed (e.g., convergence to a proper solution) the ADF estimator under model-data conditions commonly observed in practice (e.g., Beauducel & Herzberg, 2006; Flora & Curran, 2004). The WLSMV estimator for categorical data also has generally outperformed (e.g., smaller bias in parameter estimates) parceling approaches for categorical items (Bandalos, 2008). While evidence for the WLSMV estimator is accumulating, this relatively new estimator has yet to be

studied extensively in Monte Carlo studies of statistical methods. Further, we are unaware of any published work using Monte Carlo methods in data analysis under WLSMV estimation.

The primary purpose of this study was to demonstrate how Monte Carlo methods can be used in a validity study to make decisions about *N* for a desired level of π, and, to estimate π for a fixed *N* under a CFA model with model-data conditions commonly encountered in exercise and sport. Two particular model-data conditions of special importance in this study are model misspecification and ordinal data modeled under CVM with WLSMV estimation. Because the purpose is pursued by way of demonstration with the Coaching Efficacy Scale II–High School Teams (CES II–HST), related sample size recommendations are provided.

Method

We adopted a nine-step procedure proposed by Paxton, Curran, Bollen, Kirby, and Chen (2001). The steps are conceptualized as occurring in three stages: design stage, generating-the-data stage, and interpreting results stage.

Design Stage: Conceptualizing the Experiment

Step 1: Research Questions. The two a priori research questions are:

1. What is the smallest *N* necessary to achieve at least .80 π for each parameter of interest?
2. Given a particular *N*, what is the π estimate for each parameter of interest?

The first question can be viewed as a design issue addressed before data collection. The second question can be viewed as a postdata-collection issue that may provide a useful context for subsequent results. Consistent with the lead author's review of *RQES*, parameters of interest are all the nonzero elements within Λ and the unique off-diagonal elements within ψ. Consistent with both MacCallum (2003) and the lead author's review of *RQES*, the two questions will be investigated when the theoretical model only approximates the population model.

The research questions are investigated in relation to a measurement model for the CES II–HST (Myers, Feltz, Chase, Reckase, & Hancock, 2008) to provide a demonstration. Development of the CES II–HST was based on previous research (e.g., Feltz, Chase, Moritz, & Sullivan, 1999) and relevant theory (e.g., Feltz & Chase, 1998; Feltz, Short, & Sullivan, 2008). The measurement model for the CES II–HST posits that five dimensions of coaching efficacy covary and influence responses to the items. Motivation is measured by four items and is defined as the confidence that coaches have in their ability to affect the psychological mood and psychological skills of their athletes. Game strategy is measured by four items and is defined as the confidence coaches have in their ability to lead during competition. Technique is measured by four items and is defined as the confidence coaches have in their ability to use her or his instructional and diagnostic skills during practices. Character building is measured by three items and is defined as the confidence coaches have in their ability to positively influence the character development of athletes through sport. Physical conditioning is measured with three items and is defined as the confidence coaches have in their ability to prepare athletes physically for participation in her or his sport.

Items within the CES II–HST are rated on an ordered three- or four-category scale, which is consistent with relevant psychometric research (Myers, Feltz, & Wolfe, 2008; Myers, Wolfe, & Feltz, 2005). A three-category structure is specified in this study. The measurement model for the CES II–HST hypothesizes that, for each item, a coach's true location on a continuous latent response variate (y^*) directly influences the category selected. Muthén and Muthén (2002) did not provide an example of CFA models with ordinal data.

The estimation method used is WLSMV (Muthén, 1993). Under WLSMV estimation, degrees of freedom typically are estimated, not fixed, for a particular model, and derivation of the confidence interval for an RMSEA point estimate is unknown (Muthén et al., 1997). It is noted that in Mplus Version 6 (Muthén & Muthén, 1998–2010), degrees of freedom under WLSMV estimation can now be computed in a more familiar way for a particular model (i.e., the difference between the number of parameters estimated in the unrestricted model and the number of parameters estimated in the restricted model). Details of this advance, along with a simulation study that shows a negligible difference in type I error rate between the two approaches outlined in this paragraph, are provided by Asparouhov and Muthén (2010).

Step 2a: Derive a Theoretical Model. The theoretical model is depicted in Figure 1 (after fixing to 0 the two bolded paths and the two bolded covariances). The theoretical model exhibits approximate fit to data from Myers et al. (2008): $\chi^2_R(78) = 115$, $p = .004$, RMSEA = .024, comparative fit index (CFI) = .992, and Tucker-Lewis Index (TLI) = .997. Providing evidence for at least close model fit should occur prior to determining π for parameter estimates within a model (Hancock, 2006).

Step 2b: Derive a Population Model. A population model is generated based on post hoc theorizing and modification indices from results of fitting the theoretical model to the data from Myers et al. (2008). Two pattern coefficients that were originally fixed to zero, $\lambda_{t4^*,GS}$ and $\lambda_{p1^*,T}$, were freely estimated. The content of t4 ("instruct all of the different positional groups of your athletes on appropri-

ate technique during practice") led us to postulate that responses to this item could indicate game strategy efficacy (because a coach must lead the entire team during competition) in addition to technique efficacy. The content of p1 ("prepare an appropriate plan for your athletes' off-season physical conditioning") led us to propose that responses to this item could indicate technique efficacy (because the item implies both instructional and diagnostic skills) in addition to physical conditioning efficacy. Two measurement error covariances that were originally fixed to zero, $\theta_{t1^*,t3^*}$ and $\theta_{g3^*,c3^*}$, were freely estimated. The content of t1 ("teach athletes the complex technical skills of your sport during practice") and t3 ("teach athletes appropriate basic technique during practice") led us to hypothesize that the similar wording of these items could produce a nonzero covariance between the residuals of these items. The content of g3 ("make effective personnel substitutions during competition") and c3 ("effectively promote good sportsmanship in your athletes") led us to propose a nonzero covariance between the residuals of these items because a personnel substitution during competition is occasionally done for the purpose of promoting good sportsmanship.

The four post hoc modifications are made to the theoretical model (see bolded arrows in Figure 1) to generate a population model that is consistent with the Myers et al. (2008) data: χ^2_R (76) = 91, p = .116, RMSEA = .016, CFI = .997, and TLI = .999. These modifications can be conceptualized as errors of omission in the theoretical model and likely represent common types of misspecifications in CFA models. While all modifications are statistically significant, the magnitude of each within the population model (the two standardized pattern coefficients are 0.14 and 0.10, and the two correlations are .26 and .25) is often classified as practically irrelevant (Thurstone, 1930). Omitting practically irrelevant parameters from a theoretical model is consistent with a long held belief that psychological models used in practice typically cannot be made to be exactly correct (MacCallum, 2003).

Step 3: Design Experiment. Conditions commonly encountered in measurement in exercise and sport, such as model misspecification, a range of parameter estimates, and ordinal data will be built into the code and will be reviewed in subsequent steps. An iterative approach is taken to investigate the first question. The first run will specify N = 799 consistent with Myers et al. (2008) to provide a baseline. Results will be examined as follows. With respect to the first question, the primary focus will be: Was the false H_0 that a particular parameter of interest, θ_i, was equal to 0, H_0: $\theta_i = 0$ rejected 80% of the time? Samples of different sizes will be drawn until the smallest N necessary to achieve $\pi \geq .80$ for each θ_i is determined. A relevant minimal sample size of $N \geq 200$ is adopted (Flora & Curran, 2004) for the first question. With respect to the second question, two sets of sample size, common N (300, 400, 500) and small N (50 and 100), will be selected, and π will be estimated for each θ_i. Small N conditions are included because the use of structural equation modeling with small N, although inadvisable in general, is observed in practice (Hau & Marsh, 2004). In each run, except for small N runs, 10,000 datasets will be generated, which is consistent with Muthén and Muthén (2002). In small N runs only 500 datasets will be generated because it is reasonable to hypothesize that these runs will be practically invalid (to be defined in Step 8) due to the combination of small N, model complexity, a small number of categories, and model misspecification (Boomsma, 1985).

Due to the complexity of estimation, the exact effects of misspecification on subsequent parameter estimates, and hence bias and coverage, are often unknowable a priori (Kaplan, 1988). For this reason, the strategy proposed by Muthén and Muthén (2002) is altered. Answers to the two applied questions are sought first while bias and coverage are monitored as secondary considerations. Parameter estimate bias is calculated consistent with Bandalos (2006, p. 401):

$$Bias\left(\hat{\theta}_i\right) = \sum_{j=1}^{n_r}\left(\frac{\left(\hat{\theta}_{ij} - \theta_i\right)}{\theta_i}\right) / n_r, \qquad (0)$$

where $\hat{\theta}_{ij}$ is the j^{th} sample estimate of the i^{th} population parameter θ_i, and n_r is the number of replications. Parameter estimate bias will be reported as a percentage, and values $\geq |10\%|$ will be noted. Standard error estimate bias is calculated consistent with Bandalos (2006, p. 401):

$$Bias\left(\widehat{SE}\left(\hat{\theta}_i\right)\right) = \sum_{j=1}^{n_r}\left(\frac{\widehat{SE}\left(\hat{\theta}_i\right)_j - SE\left(\hat{\theta}_i\right)}{SE\left(\hat{\theta}_i\right)}\right) / n_r, \qquad (1)$$

where $\widehat{SE}\left(\hat{\theta}_i\right)$ is the estimated standard error of $\hat{\theta}_i$ for the j^{th} replication, and $SE\left(\hat{\theta}_i\right)$ is an estimate of the population standard error of $\hat{\theta}_i$. Standard error estimate bias will be reported as a percentage and values $> |5\%|$ will be noted. Coverage also will be reported as a percentage and values outside of 91–98% will be noted.

Step 4: Choosing Values of Population Parameters. The population values used to generate the data will be taken from the results of imposing the population model on the Myers et al. (2008) data described in Step 2 (see Table 1). The variability across the practically relevant elements of Λ, .78 to 1.56, and ψ, .38 to .77, for the correlations and 0.33 to 0.69 for the variances, is consistent with the lead author's review of *RQES*. Standardized pattern coefficients are not provided in Table 1, but sufficient information for deriving these coefficients is provided (i.e., each unstandardized coefficient and the variance of each latent variable). For example, let $\lambda_{m2^*,M}$ represent the population coefficient from motivation efficacy to m2*. Standardized

$\lambda_{m2^*,M}$ is given by $\lambda_{m2^*,M}$*(SD_M/SD_{m2^*}). Therefore, the standardized value of $\lambda_{m2^*,M}$ = 1.02 * (.69/1) = 0.71. Practically relevant standardized pattern coefficients ranged from 0.57 to 0.89.

Step 5: Choosing Software. Analyses will be performed in Mplus 6.0 (Muthén & Muthén, 1998–2010). An annotated input file, where "!" and italicized text signify annotation, is available on request from the lead author and online at http://nicholas-myers.blogspot.com/. Further descriptions of key sections of the input file are provided.

Misspecification is created in two ways in this study. The first way is by specifying a population model under the Model Population section that differs from the theoretical model specified under the Model section (i.e., model error). The second way is by using sample data from Myers et al. (2008), where *N* was not extremely large, to generate estimates of the population values (i.e., sampling error). Including both model error and sampling error under a Monte Carlo approach is consistent with MacCallum and Tucker (1991). Average RMSEA across replications, $\overline{RMSEA}$, will be used to quantify the degree of misspecification (Bandalos, 2006).

Ordinal data are created by categorizing the data, which is drawn from a multivariate normal distribution. The population threshold values, which categorize the continuous data, are derived based on the Myers et al. (2008) data. These threshold values are *z* values determined by (a) the proportion of observations in or below a particular category and (b) the area under a standard normal curve. For example, there were 82 observations in the first category of g1. This proportion of observations relative to the total possible, 799, was .103. The threshold value between the first category and the second category of g1 (i.e., g1$1) is the *z* value that corresponds to cumulative area .103, which is -1.267. The population distribution for each item, then, was the same as observed in Myers et al.

Generating Data Stage: Performing the Experiment

Step 6: Executing the Simulations. The simulations are executed as described previously with additional data management issues noted in Step 8. Note that the only change to the code needed for each new *N* was the requested number of observations.

Table 1. Population/start values for the pattern coefficient matrix, Λ, and the covariance matrix, Ψ

Item	Population values for Λ					Start values for Λ				
	M	GS	T	CB	PC	M	GS	T	CB	PC
m1	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00
m2	1.02	0.00	0.00	0.00	0.00	1.02	0.00	0.00	0.00	0.00
m3	1.08	0.00	0.00	0.00	0.00	1.08	0.00	0.00	0.00	0.00
m4	1.08	0.00	0.00	0.00	0.00	1.08	0.00	0.00	0.00	0.00
g1	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
g2	0.00	1.03	0.00	0.00	0.00	0.00	1.03	0.00	0.00	0.00
g3	0.00	0.91	0.00	0.00	0.00	0.00	0.92	0.00	0.00	0.00
g4	0.00	1.05	0.00	0.00	0.00	0.00	1.05	0.00	0.00	0.00
t1	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
t2	0.00	0.00	0.94	0.00	0.00	0.00	0.00	0.90	0.00	0.00
t3	0.00	0.00	1.07	0.00	0.00	0.00	0.00	1.06	0.00	0.00
t4	0.00	0.18	0.78	0.00	0.00	0.00	0.00	0.92	0.00	0.00
c1	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	1.00	0.00
c2	0.00	0.00	0.00	1.05	0.00	0.00	0.00	0.00	1.05	0.00
c3	0.00	0.00	0.00	1.07	0.00	0.00	0.00	0.00	1.08	0.00
p1	0.00	0.00	0.12	0.00	1.00	0.00	0.00	0.00	0.00	1.00
p2	0.00	0.00	0.00	0.00	1.43	0.00	0.00	0.00	0.00	1.22
p3	0.00	0.00	0.00	0.00	1.56	0.00	0.00	0.00	0.00	1.33
	Population values for Ψ					Start values for Ψ				
	M	GS	T	CB	PC	M	GS	T	CB	PC
M	0.48	0.73	0.56	0.73	0.55	0.48	0.73	0.56	0.73	0.56
GS	0.40	0.62	0.77	0.51	0.54	0.40	0.62	0.78	0.52	0.55
T	0.32	0.51	0.69	0.42	0.53	0.33	0.52	0.73	0.42	0.55
CB	0.39	0.31	0.27	0.60	0.38	0.39	0.32	0.28	0.59	0.38
PC	0.22	0.25	0.25	0.17	0.33	0.26	0.29	0.31	0.20	0.44

Note. M = motivation; GS = game strategy; T = technique; CB = character building; PC = physical conditioning; for each Ψ, standardized values are provided above the main diagonal.

Step 7: File Storage. All of the files will be saved in the location from which the input file is run. A different folder was created for each run. An advantage to saving all of the generated datasets is that some datasets may need to be discarded, and a particular run repeated, for reasons that will be described below.

Step 8: Troubleshooting. The data collection plan is sometimes not exactly observed in the generation of the data, and the data analysis plan is sometimes not exactly observed in the analysis of the generated data (Bandalos, 2006). Two types of problems will be observed in the Results section. The first problem, absence of at least one observation in a particular category, is regarded as a data generation problem. Such a dataset is unusable, because a requirement of the analysis is that a constant number of thresholds are estimable within any particular item. The second type of problem, nonconvergent solutions and improper solutions (e.g., a negative variance), is regarded as a data analytic problem. Either of these solution types will result in the elimination of the relevant dataset. The number of unusable and eliminated datasets will be reported. A run will be considered practically invalid if the number of unusable and eliminated datasets is greater than 5% of the requested replications. This heuristic is adopted because having adequate power is not particularly useful if, with a given dataset, there is a nontrivial chance that the theoretical model either will not be estimable or will converge to an improper solution.

Runs with less than 5% of unusable and eliminated datasets will be repeated after problematic datasets are removed. Repeating a run after removing problematic datasets is sometimes referred to as an external Monte Carlo analysis. An annotated input file is available on request to the lead author and online at http://nicholas-myers.blogspot.com/. The context for this run is given in the Results section.

Interpreting Results Stage: Findings from the Experiment

Step 9: Summarizing Results. For each run (e.g., 10,000 replications for the $N = 300$ run), Mplus integrates the relevant information across replications into a single output file. All of the information for the Results discussed in this step is provided under the Model Results section of the output file provided by Mplus. Key parameter-level results from the baseline run will be depicted in Table 2. The first column will specify parameters of interest (e.g., $\lambda_{\mathrm{ml,M}}$ denotes the path from latent motivation to latent response variate m1*; $\psi_{\mathrm{M,GS}}$ denotes the covariance between latent motivation and latent game strategy; etc.). The second column, $\hat{\pi}$, provides the percentage of replications in which $H_0{:}\theta_i = 0$ is rejected. The third column, θ_i, provides the population value for the i^{th} parameter. The fourth column, $\overline{\hat{\theta}_i}$, provides the average estimate for θ across replications. The fifth column, $Bias\left(\hat{\theta}_i\right)\%$, provides the relevant parameter bias value as a percentage. The sixth column, $SE\left(\hat{\theta}_i\right)$, provides an estimate of the relevant population standard error (i.e., *SD* of $\hat{\theta}_i$ across replications). The seventh column, $\overline{SE\left(\hat{\theta}_i\right)}$, provides the average *SE* estimate for $\hat{\theta}_i$ across replications. The eighth column, $Bias\left(\overline{SE\left(\hat{\theta}_i\right)}\right)\%$, provides the relevant standard error bias value as a percentage. The ninth column, $Cover_{95\%\,CI}$, provides the percentage of replications in which the 95% confidence interval (CI) includes the population value. Model-level information will be summarized in the text.

Results

Question 1

For the baseline run ($N = 799$), all 10,000 generated datasets are usable. Each dataset converges to a proper solution when fit to the theoretical model and $\overline{RMSEA} = .018$. Key parameter-level results from the baseline run are depicted in Table 2. Power is 100% for each $H_0{:}\theta_i = 0$. Parameter estimate bias exceeds |10%| for 3 of 13 pattern coefficients, *Bias* ($\hat{\lambda}_{t4^*,T} = 19.1\%$), *Bias* ($\hat{\lambda}_{t\beta2^*,PC} = -14.7\%$), *Bias* ($\hat{\lambda}_{t\beta2^*,PC} = -15.1\%$), and for 4 of 10 latent variable covariances, *Bias* ($\hat{\psi}_{M,PC} = 18.1\%$), *Bias* ($\hat{\psi}_{GS,PC} = 19.4\%$), *Bias* ($\hat{\psi}_{T,PC} = 23.1\%$), *Bias* ($\hat{\psi}_{CB,PC} = 18.2\%$). Standard error estimate bias never exceeds |5%|. Coverage is less than 91% for 4 of 13 pattern coefficients, $Cover_{95\%\,CI}$ ($\hat{\lambda}_{t2^*,T} = 86.0\%$), $Cover_{95\%\,CI}$($\hat{\lambda}_{t4^*,T} = 1.6\%$), $Cover_{95\%\,CI}$($\hat{\lambda}_{t2^*,PC} = 25.9\%$), $Cover_{95\%\,CI}$($\hat{\lambda}_{t\beta3^*,PC} = 23.2\%$), and for 4 of 10 latent variable covariances, $Cover_{95\%\,CI}$($\hat{\psi}_{M,PC} = 66.0\%$), $Cover_{95\%\,CI}$ ($\hat{\psi}_{GS,PC} = 57.8\%$), $Cover_{95\%\,CI}$($\hat{\psi}_{T,PC} = 45.8\%$), $Cover_{95\%\,CI}$($\hat{\psi}_{CB,PC} = 80.5\%$). A partial output file with annotation is available on request to the lead author and online at http://nicholas-myers.blogspot.com/.

For the final run ($N = 200$), 9,979 of the 10,000 generated datasets are usable. A small number of usable datasets, 163 (or 1.6%), converge to an improper solution when fit to the theoretical model and are eliminated. The final run is repeated with 9,816 usable datasets (or 98.2%) in an external Monte Carlo analysis. Each dataset converges to a proper solution when fit to the theoretical model and $\overline{RMSEA} = .019$. Key parameter-level results are similar to the baseline run (see Table 2). Power is ≥ 98.4% for each $H_0{:}\theta_i = 0$. Parameter estimate bias exceeds |10%| for the same three pattern coefficients, *Bias* ($\hat{\lambda}_{t4^*,T} = 19.5\%$), *Bias* ($\hat{\lambda}_{t\beta2^*,PC} = -14.1\%$), *Bias* ($\hat{\lambda}_{t\beta3^*,PC} = -14.8\%$), and the same four latent variable covariances, *Bias* ($\hat{\psi}_{M,PC} = 20.0\%$), *Bias* ($\hat{\psi}_{GS,PC} = 21.4\%$), *Bias* ($\hat{\psi}_{T,PC} = 25.1\%$), *Bias* ($\hat{\psi}_{CB,PC} = 20.2\%$), as in the baseline run. Standard error estimate bias exceeds |5%| for only *Bias* (*SE* ($\hat{\lambda}_{t4^*,T} = -6.6\%$)). Coverage is less than 91% for three of the same pattern coefficients, $Cover_{95\%\,CI}$($\hat{\lambda}_{t4^*,T} = 41.4\%$), $Cover_{95\%\,CI}$ ($\hat{\lambda}_{t\beta2^*,PC} = 63.2\%$), $Cover_{95\%\,CI}$($\hat{\lambda}_{t\beta3^*,PC} = 60.3\%$), and for the same four latent variable covari-

ances, $Cover_{95\%CI}(\hat{\psi}_{M,PC} = 87.1\%)$, $Cover_{95\%CI}(\hat{\psi}_{GS,PC} = 83.4\%)$, $Cover_{95\%CI}(\hat{\psi}_{T,PC} = 78.9\%)$, $Cover_{95\%CI}(\hat{\psi}_{CB,PC} = 90.8\%)$, as in the baseline run.

The finding for Question 1 is that a relatively small sample ($N = 200$) provides ample π to reject each $H_0: \theta_i = 0$. Problematic bias values and coverage values, however, are observed at both the small sample and baseline sample. Thus, while there is ample π, a few absolute $Bias\left(\hat{\theta}_i\right)$ values are relatively large, and the 95% CI around a few $\hat{\theta}_i$ too frequently exclude θ_i.

Question 2

Results from the common N (300, 400, and 500) runs closely follow results from the baseline run and the final run. The number of usable datasets range from to 9,998 ($N = 300$) to 10,000 ($N = 400$ and 500). The number of usable datasets that converge to an improper solution when fit to the theoretical model, range from 19 ($N = 300$) to 1 ($N = 500$) and are eliminated. For each run $\overline{RMSEA} = .017$. Power is ≥ 99.9% for each $H_0: \theta_i = 0$. In each of the three runs, parameter estimate bias exceeds |10%| for the same three pattern coefficients at approximately the same percentages, and for the same four latent variable covariances at approximately the same percentages, as in the baseline run (see Table 2) and the final run. Standard error estimate bias exceeds |5%| for only Bias ($SE(\hat{\lambda}_{t4^*,T} = -5.2\%)$) and only when $N = 300$. Coverage is less than 91% for the same three or four pattern coefficients at approximately the same percentages, and for the same four latent variable covariances at approximately the same percentages as in the baseline run and the final run. The small N runs ($N = 50$ and 100) were considered practically invalid because 361 (or 72.2%) and 89 (or 17.8%) of the generated datasets are problematic.

The answer to Question 2 with common N is similar to the answer to Question 1. Commonly used sample sizes provide ample π to reject each $H_0: \theta_i = 0$. Similarly problematic bias values and coverage values, however, continue to be observed at each sample size.

A Post Hoc Question

Given the repeated observation of a few problematic bias values and a few low coverage values, a post hoc question is: To what degree might the relatively minor model

Table 2. Key parameter-level results from baseline run ($N = 799$)

Parameter	$\hat{\pi}$ (%)	θ_i	$\overline{\hat{\theta}_i}$	*Bias* $(\hat{\theta}_i)$ (%)	*SE* $(\hat{\theta}_i)$	$\overline{SE\,(\hat{\theta}_i)}$	*Bias* (*SE* $(\hat{\theta}_i)$) (%)	$Cover_{95\%CI}$ (%)
$\lambda_{m1^*,M}$	—	1.00	1.00	—	0.000	0.000	—	—
$\lambda_{m2^*,M}$	100.0	1.02	1.02	0.3	0.059	0.059	0.0	95.2
$\lambda_{m3^*,M}$	100.0	1.08	1.08	0.2	0.060	0.059	-1.0	95.2
$\lambda_{m4^*,M}$	100.0	1.08	1.09	0.2	0.060	0.059	-0.8	94.8
$\lambda_{g1^*,GS}$	—	1.00	1.00	—	0.000	0.000	—	—
$\lambda_{g2^*,GS}$	100.0	1.03	1.03	0.1	0.041	0.040	-1.5	94.9
$\lambda_{g3^*,GS}$	100.0	0.91	0.92	0.9	0.042	0.042	0.0	94.9
$\lambda_{g4^*,GS}$	100.0	1.05	1.04	0.0	0.041	0.041	-1.0	94.9
$\lambda_{t1^*,T}$	—	1.00	1.00	—	0.000	0.000	—	—
$\lambda_{t2^*,T}$	100.0	0.94	0.91	-3.2	0.036	0.036	-1.1	86.0
$\lambda_{t3^*,T}$	100.0	1.07	1.06	-0.2	0.034	0.036	4.7	95.7
$\lambda_{t4^*,T}$	100.0	0.78	0.93	19.1	0.037	0.036	-3.5	1.6
$\lambda_{c1^*,CB}$	—	1.00	1.00	—	0.000	0.000	—	—
$\lambda_{c2^*,CB}$	100.0	1.05	1.06	0.2	0.061	0.061	-0.2	95.2
$\lambda_{c3^*,CB}$	100.0	1.07	1.08	1.4	0.063	0.063	-0.5	95.1
$\lambda_{p1^*,PC}$	—	1.00	1.00	—	0.000	0.000	—	—
$\lambda_{p2^*,PC}$	100.0	1.43	1.22	-14.7	0.079	0.078	-1.6	25.9
$\lambda_{p3^*,PC}$	100.0	1.56	1.33	-15.1	0.086	0.084	-1.8	23.2
$\Psi_{M,GS}$	100.0	0.40	0.40	0.1	0.028	0.027	-2.2	94.4
$\Psi_{M,T}$	100.0	0.32	0.33	3.0	0.028	0.027	-0.4	93.6
$\Psi_{M,CB}$	100.0	0.39	0.39	-0.1	0.029	0.029	0.0	95.0
$\Psi_{M,PC}$	100.0	0.22	0.26	18.1	0.026	0.025	-1.2	66.0
$\Psi_{GS,T}$	100.0	0.51	0.52	2.7	0.027	0.027	-0.4	91.4
$\Psi_{GS,CB}$	100.0	0.31	0.32	2.3	0.030	0.029	-1.0	94.1
$\Psi_{GS,PC}$	100.0	0.25	0.29	19.4	0.028	0.027	-2.9	57.8
$\Psi_{T,CB}$	100.0	0.27	0.28	2.4	0.031	0.031	-1.0	94.2
$\Psi_{T,PC}$	100.0	0.25	0.31	23.1	0.029	0.028	-3.8	45.8
$\Psi_{CB,PC}$	100.0	0.17	0.20	18.2	0.027	0.027	-1.1	80.5

Note. *SE* = standard error.

misspecifications be responsible for these problems? This question is interesting because evidence for some degree of misspecification is generally consistent with applied measurement (MacCallum, 2003). All of the runs previously described are repeated with the exception that the theoretical model is slightly altered to mirror the population model exactly (i.e., the population model is assumed known). Simply, the four practically irrelevant parameters described previously (and fixed to = 0 in the theoretical model) are now freely estimated in the theoretical model. From this point forward, for simplicity, the set of practically *irrelevant* parameter estimates is denoted $\hat{\theta}_j$, whereas the set of parameter estimates of *interest* is denoted $\hat{\theta}_i$.

The final run (N= 200) is now considered practically invalid because 702 (or 7.2%) of the generated datasets are problematic. Results from the baseline (N= 799) and common N (300, 400, and 500) runs *without* model misspecification are generally very similar to each other and differ from the parallel runs *with* model misspecification in similar ways. Therefore, results from only the baseline run (N= 799) are reported in detail. This run is selected because the results with model misspecification were previously reported both in the text and in Table 2. Power for each H_0:$\theta_i = 0$ across runs, ≥ 99.8%, is similar to parallel previous runs with model misspecification.

For the baseline run (N= 799), each dataset converges to a proper solution when fit to the theoretical model without misspecification and $\overline{RMSEA} = .005$. Key parameter-level results from the baseline run are depicted in Table 3. Parameter estimate bias never exceeds |10%| within $\hat{\theta}_i$ (the maximum value is .7%) and ranges from -2.5 to -1.1 within $\hat{\theta}_j$. Standard error estimate bias never exceeds |5%| within $\hat{\theta}_i$ (the maximum absolute value is 3.1%) and ranges from -3.3% to -0.3% within $\hat{\theta}_j$. Coverage is never less than 91% within either $\hat{\theta}_i$ or $\hat{\theta}_j$. Power is 100% for each H_0:$\theta_i = 0$ and ranges from 44.5% to 75.4% for each H_0:$\theta_i = 0$. Power for each H_0:$\theta_i = 0$ varies across runs (e.g., ranges from 20.3% to 36.2% when N= 300).

Table 3. Key parameter-level results from the baseline run (N = 799) without model misspecification

Parameter	$\hat{\pi}$ (%)	θ_i	$\overline{\hat{\theta}_i}$	*Bias* ($\hat{\theta}_i$) (%)	*SE* ($\hat{\theta}_i$)	$\overline{\widehat{SE}\,(\hat{\theta}_i)}$	*Bias* ($\widehat{SE}\,(\hat{\theta}_i)$) (%)	$Cover_{95\%CI}$ (%)
$\lambda_{m1^*,M}$	—	1.00	1.00	—	0.000	0.000	—	—
$\lambda_{m2^*,M}$	100.0	1.02	1.02	0.3	0.059	0.059	0.0	95.2
$\lambda_{m3^*,M}$	100.0	1.08	1.08	0.2	0.060	0.059	-0.8	95.2
$\lambda_{m4^*,M}$	100.0	1.08	1.09	0.2	0.060	0.059	-0.8	94.8
$\lambda_{g1^*,GS}$	—	1.00	1.00	—	0.000	0.000	—	—
$\lambda_{g2^*,GS}$	100.0	1.03	1.03	0.1	0.041	0.040	-1.7	94.8
$\lambda_{g3^*,GS}$	100.0	0.91	0.91	0.1	0.042	0.042	0.2	95.2
$\lambda_{g4^*,GS}$	100.0	1.05	1.04	0.0	0.041	0.041	-1.2	94.9
$\lambda_{t4^*,GS}$	45.3	0.18	0.18	-1.9	0.105	0.102	-3.3	95.1
$\lambda_{t1^*,T}$	—	1.00	1.00	—	0.000	0.000	—	—
$\lambda_{t2^*,T}$	100.0	0.94	0.94	0.1	0.045	0.045	-0.2	95.3
$\lambda_{t3^*,T}$	100.0	1.07	1.07	0.2	0.037	0.037	-0.3	95.1
$\lambda_{t4^*,T}$	100.0	0.78	0.79	0.7	0.106	0.102	-3.1	94.8
$\lambda_{p1^*,T}$	47.1	0.12	0.11	-1.1	0.062	0.061	-1.0	94.8
$\lambda_{c1^*,CB}$	—	1.00	1.00	—	0.000	0.000	—	—
$\lambda_{c2^*,CB}$	100.0	1.05	1.06	0.2	0.061	0.061	-0.7	95.2
$\lambda_{c3^*,CB}$	100.0	1.07	1.07	0.2	0.062	0.062	-0.2	95.0
$\lambda_{p1^*,PC}$	—	1.00	1.00	—	0.000	0.000	—	—
$\lambda_{p2^*,PC}$	100.0	1.43	1.43	0.4	0.129	0.127	-1.5	94.8
$\lambda_{p3^*,PC}$	100.0	1.56	1.57	0.5	0.148	0.145	-2.4	94.3
$\psi_{M,GS}$	100.0	0.40	0.40	0.3	0.028	0.027	-2.2	94.4
$\psi_{M,T}$	100.0	0.32	0.32	0.2	0.028	0.028	-1.1	94.6
$\psi_{M,CB}$	100.0	0.39	0.39	0.3	0.029	0.029	0.3	95.0
$\psi_{M,PC}$	100.0	0.22	0.22	0.5	0.027	0.027	-0.7	94.7
$\psi_{GS,T}$	100.0	0.51	0.51	0.2	0.028	0.028	-1.4	94.5
$\psi_{GS,CB}$	100.0	0.31	0.31	0.5	0.030	0.029	-1.7	94.7
$\psi_{GS,PC}$	100.0	0.25	0.25	0.5	0.030	0.029	-2.3	94.5
$\psi_{T,CB}$	100.0	0.27	0.27	0.5	0.031	0.031	-1.3	94.6
$\psi_{T,PC}$	100.0	0.25	0.25	0.4	0.031	0.031	-2.2	94.4
$\psi_{CB,PC}$	100.0	0.17	0.17	0.7	0.026	0.026	-1.1	94.6
t1*,t3*	44.5	0.07	0.07	-1.5	0.037	0.037	-0.5	94.9
c3*,g3*	75.4	0.10	0.10	-2.5	0.036	0.036	0.3	95.1

Note. *SE* = standard error.

The findings related to Question 3 suggest that the relatively small degree of model misspecification appears to be responsible for the problematic bias values and for the low coverage values previously observed. This finding is consistent with Monte Carlo studies of statistical methods that manipulated model misspecification (e.g., Kaplan, 1988). Thus, for correctly specified models across all practically valid runs (i.e., $N \geq 300$), there is ample power for each H_0: $\theta_i = 0$. $Bias\left(\hat{\theta}_i\right)$ values are within generally accepted levels, and the 95% CI around each $\hat{\theta}_i$ almost always includes θ_i. Thus, it appears that while introducing model misspecification via Monte Carlo methods in a data analytic situation has the advantage of more closely reflecting practice, it has the disadvantage of introducing additional difficulties (MacCallum, 2003).

Discussion

Monte Carlo methods have long been used to advance statistical theory. There have been several recent calls to use Monte Carlo methods as a tool to improve applications of quantitative methods in substantive research (e.g., MacCallum, 2003; Muthén & Muthén, 2002). The primary purpose of this study is to demonstrate how Monte Carlo methods can be used to decide on sample size and to estimate power for a CFA model under model-data conditions commonly encountered in measurement in exercise and sport. Because the purpose is pursued by way of demonstration with the CES II–HST, related sample size recommendations are provided: $N \geq 200$ for the theoretical model, $N \geq 300$ for the population model.

The two questions, "What sample size do I need to achieve a particular level of power?" and "How much power will I have with a fixed sample size?" commonly arise in validity studies. Annotated Mplus code for investigating these questions in relation to the measurement model for the CES II–HST is available upon request to the lead author and online at http://nicholas-myers.blogspot.com/. Combining the two annotated input files with Figure 1 provides an example of how to translate models into code. The code can be altered in relatively minor ways for other CFA models common to exercise and sport. The authors of this study focus on these questions in relation to the pattern coefficients and the covariances between the latent variables (to maintain a reasonable focus) but the code does not need to be altered if the focus expands to include the variances of the latent variables. The findings for the first two questions in relation to the CES II–HST are: as little as 200 (for the theoretical model) or 300 (for the population model), and, at least 99.9% for sample sizes of 300, 400, and 500.

An important assumption embedded in the code is that the theoretical model only approximates the population model. The first two questions, then, are investigated under the common scenario that there is evidence against exact model-data fit. An apparent consequence of the misspecification in this case is that, over repeated sampling, a few parameter estimates are biased and the confidence interval around a parameter estimate too frequently excludes the population value in a few instances. While this result is consistent with statistical theory (Kaplan, 1988), it also serves as a reminder that a level of misfit that may be regarded as trivial in practice may have troubling effects on parameters of conceptual interest. Thus, it is likely that other validity studies where there is evidence against exact fit experience similar problems. Incorporating Monte Carlo methods in validity studies may provide information on where these effects may be occurring and encourage sustained efforts toward generating closer approximations of population models. Such efforts must be balanced against model generation based only on modification indices (MacCallum, 1986). As MacCallum demonstrated, a post hoc specification search based only on empirical information (e.g., modification indexes) frequently results in the acceptance of a post hoc theoretical model that may be consistent with a particular dataset but, more importantly, may be inconsistent with the true model (i.e., population model).

Primary limits of this study include the use of a single (and not extremely large) dataset to generate a population model and population values, the exclusion of other conditions commonly found in practice, a sample size recommendation that happens to converge on a common rule of thumb, and the focus on only a particular CFA model. The use of the Myers et al. data (2008) and post hoc theorizing to generate a population model and population values can be viewed as a limitation of this study in the sense that, because the observed dataset was not extremely large and was treated as the population, sampling error was likely nontrivial. The population model and the population values proposed, therefore, are unlikely to be exactly correct and should be viewed as only reasonable hypotheses. We view this limitation as tolerable (and necessary as the population model and population values are unlikely to ever be truly known) given the paucity of research with the CES II–HST and the intended broader contribution of the study. Other conditions commonly encountered in practice, but not modeled in this study, include missing data and continuous data. Muthén and Muthén (2002) provide examples of how to impose both conditions. The sample size recommendations for the CES II–HST forwarded in this study (i.e., $N \geq 200$ for the theoretical model, $N \geq 300$ for the population model) happen to converge at, or near to, a common rule of thumb (i.e., $N \geq 200$). This convergence should not be viewed as either general support for this rule of thumb or as evidence against the usefulness of the

Monte Carlo approach described in this study. Implementing the Monte Carlo approach advocated in this study will always provide more information (e.g., bias) than will unexplored adherence to the $N \geq 200$ rule of thumb. As such, we advocate that in research situations similar to those described in this study, researchers in exercise and sport should strongly consider implementing the Monte Carlo approach described in this study to make decisions about N and/or to estimate π (as opposed to relying on related rules of thumb). Last, this study focuses on only a particular CFA model, which limits the breadth of the contribution of the study.

References

Asparouhov, T., & Muthén, B. O. (2010). *Simple second order chi-square correction.* Retrieved from Mplus website: http://www.statmodel.com/download/WLSMV_new_chi21.pdf

Bandalos, D. L. (2006). The use of Monte Carlo studies in structural equation modeling research. In R. C. Serlin (Series Ed.), G. R. Hancock, & R. O. Mueller (Vol. Eds.), *Structural equation modeling: A second course* (pp. 385–462). Greenwich, CT: Information Age.

Bandalos, D. L. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Structural Equation Modeling, 15,* 211–240.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13,* 186–203.

Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in LISREL maximum likelihood estimation. *Psychometrika, 50,* 229–242.

Browne, M. W. (1984). Asymptotic distribution-free methods in the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62–83.

Cohen, J. (1988). *Statistical power for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

DiStefano, C. (2002). The impact of categorization with confirmatory factor analysis. *Structural Equation Modeling, 9,* 327–346.

Dolan, C. V. (1994). Factor analysis of variables with 2, 3, 5, and 7, response categories: A comparison of categorical variable estimators using simulated data. *British Journal of Mathematical and Statistical Psychology, 47,* 309–326.

Feltz, D. L., & Chase, M. A. (1998). The measurement of self-efficacy and confidence in sport. In J. L. Duda (Ed.), *Advancements in sport and exercise psychology measurement* (pp. 65–80). Morgantown, WV: Fitness Information Technology.

Feltz, D. L., Chase, M. A., Moritz, S. E., & Sullivan, P. J. (1999). A conceptual model of coaching efficacy: Preliminary investigation and instrument development. *Journal of Educational Psychology, 91,* 765–776.

Feltz, D. L., Short, S. E., & Sullivan, P. J. (2008). *Self-efficacy in sport.* Champaign, IL: Human Kinetics.

Finney, S. J., & DiStefano, C. (2006). Nonnormal and categorical data in structural equation modeling. In R. C. Serlin (Series Ed.), G. R. Hancock, & R. O. Mueller (Vol. Eds.), *Structural equation modeling: A second course* (pp. 269–313). Greenwich, CT: Information Age.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9,* 466–491.

Gagné, P., & Hancock, G. R. (2006). Measurement model quality, sample size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research, 41,* 65–83.

Gentle, J. E. (2003). *Random number generation and Monte Carlo methods* (2nd ed.). New York: Springer.

Gentle, J. E. (2005). Monte Carlo simulation. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral science* (Vol. 3, pp. 1264–1271). Hoboken, NJ: Wiley.

Hancock, G. R. (2006). Power analysis in covariance structure modeling. In R. C. Serlin (Series Ed.), G. R. Hancock, & R. O. Mueller (Vol. Eds.), *Structural equation modeling: A second course* (pp. 69–115). Greenwich, CT: Information Age.

Hau, K-T., & Marsh, H. W. (2004). The use of item parcels in structural equation modeling: Non-normal data and small sample sizes. *British Journal of Mathematical and Statistical Psychology, 57,* 327–351.

Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equation Modeling, 8,* 205–223.

Jackson, D. L. (2003). Revisiting sample size and number of parameter estimates: Some support for the N:q hypothesis. *Structural Equation Modeling, 10,* 128–141.

Kaplan, D. (1988). The impact of specification error on the estimation, testing, and improvement of structural equation models. *Multivariate Behavioral Research, 23,* 467–482.

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin, 100,* 107–120.

MacCallum, R. (2003). Working with imperfect models. *Multivariate Behavioral Research, 38,* 113–139.

MacCallum, R., & Tucker, L. R. (1991). Representing sources of error in the common factor model: Implications for theory and practice. *Psychological Bulletin, 100,* 502–511.

Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220.

Muthén, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika, 49,* 115–132.

Muthén, B. O. (1993). Goodness of fit with categorical and other non-normal variables. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 205–243). Newbury Park, CA: Sage.

Muthén, B. O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171–189.

Muthén, L. K., & Muthén, B. O. (1998–2010). *Mplus user's guide* (6th ed.). Los Angeles, CA: Authors.

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.

Myers, N. D., Feltz, D. L., Chase, M. A., Reckase, M. D., & Hancock, G. R. (2008). The Coaching Efficacy Scale II–High School Teams. *Educational and Psychological Measurement, 68,* 1059–1076.

Myers, N. D., Feltz, D. L., & Wolfe, E. W. (2008). A confirmatory study of rating scale category effectiveness for the Coaching Efficacy Scale. *Research Quarterly for Exercise and Sport, 79,* 300–311.

Myers, N. D., Wolfe, E. W., & Feltz, D. L. (2005). An evaluation of the psychometric properties of the Coaching Efficacy Scale for American coaches. *Measurement in Physical Education and Exercise Science, 9,* 135–160.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte Carlo experiments: Design and implementation. *Structural Equation Modeling, 8,* 287–312.

Thurstone, L. L. (1930). The learning function. *Journal of General Psychology, 3,* 469–470.

Authors' Note

Please address correspondence concerning this article to Nicholas D. Myers, Department of Educational and Psychological Studies, University of Miami, Merrick Building 311E, Coral Gables, FL, 33124-2040.

E-mail: nmyers@miami.edu