STRUCTURAL EQUATION MODELING, *10*(1), 128–141

Revisiting Sample Size and Number of Parameter Estimates: Some Support for the N:q Hypothesis

Dennis L. Jackson
Tricon Global Restaurants, Inc.
Louisville, KY

A number of authors have proposed that determining an adequate sample size in structural equation modeling can be aided by considering the number of parameters to be estimated. While this advice seems plausible, little empirical support appears to exist. A previous study by Jackson (2001), failed to find support for this hypothesis, however, there were certain limitations to the study that likely led to the lack of findings. This study revisits the issue with a design modified to be more sensitive to practically significant effects of sample size to parameter estimate ratio. Consequently, some support was found for this hypothesis, notably among overall fit measures and tests. Results indicate that higher values of the observations per parameter ratio had a positive effect for some measures of fit. However, the overall effect was small relative to absolute sample size.

When planning research that will utilize structural equation modeling techniques, researchers are faced with the question of identifying an adequate sample size. Various approaches to arriving at a sample size have been suggested, such as a minimum sample size (e.g., 200), having a certain number of observations per measured variable, or through conducting power analyses (e.g., MacCallum, Browne, & Sugawara, 1996). Another suggested approach equates the necessary sample size to the number of parameters that must be estimated (e.g., Bentler & Chou, 1987; Bollen, 1989; Kline, 1998, Marsh, Balla, & McDonald, 1988; Mueller, 1997; Tanaka, 1987; Ullman, 1996), where higher values of the ratio of observations to parameters to be estimated (N:q) are preferred. In a previous study, Jackson (2001) failed to find support for this approach. However, the failure to find any practically significant effects for N:q likely had to do with the study design, namely not having enough variation in the independent variable of N:q (P. M. Bentler, personal communication, June 2001). My study represents an attempt to revisit this previous work using a modified design that incorporates more variance in this independent variable.

Requests for reprints should be sent to Dennis L. Jackson, Tricon Global Restaurants, Inc., 1441 Gardiner Lane, Louisville, KY 40213. E-mail: dennis.jackson@tricon-yum.com

The question of how large of a sample size is required for covariance structure modeling (CSM) is a deceptively difficult one to answer. A growing body of simulation work suggests that it is dependent upon several things, such as the number of indicators per latent variable (Gerbing & Anderson, 1985; Marsh, Hau, Balla, & Grayson, 1998; Velicer & Fava, 1998); the strength of association between the indicators and the latent variables (Bandalos, 1997; Gerbing & Anderson, 1985; Velicer & Fava, 1998), the degree of multivariate normality (West, Finch, & Curran, 1995) and estimation method (Fan, Thompson, & Wang, 1999; Fan & Wang, 1998; Tanaka, 1987).

The advice to consider sample size (*N*) in terms of number of parameters to be estimated (q) can be supported by a few different arguments. First, in CSM, the number of measured variables determines the number of variance and covariance elements that comprise the covariance matrix; however, it does not determine the number of model parameters that must be estimated. For instance, with 20 measured variables there are 210 unique elements in the variance–covariance matrix. However, any variety of models can be hypothesized to reproduce these unique elements, all the way from a highly saturated model (i.e., lots of parameters) to a very restrictive model with very few parameters. One can see how this fact would lead investigators to conclude that sample size should be based on the number of parameters being estimated, rather than, for instance, the number of measured variables.

A second argument, is based on achieving adequate power for hypothesis testing. MacCallum et al. (1996) tied sample size to the expected effect size and the degrees of freedom. More degrees of freedom and larger effect sizes mean that fewer observations are needed to achieve acceptable levels of power. Fewer degrees of freedom, implying more parameters being estimated, means larger sample and effect sizes need to be realized to achieve adequate power. It should be noted that MacCallum et al. (1996) acknowledged that there are other considerations for sample size besides hypothesis testing, such as parameter estimation precision.

Finally, there is the question of replication. When conducting research, a researcher hopes to arrive at a conclusion that will be replicated in future studies. Some researchers (e.g., Browne & Cudeck, 1993) have argued that less complex models will replicate better, especially with smaller samples. Furthermore, it has been argued that some fit indexes favor highly parameterized models when using large samples (Cudeck & Henly, 1991). Therefore, presumably the logic is that larger sample sizes may be required to replicate results from a model with a large number of parameters being estimated.

Recommendations for determining sample size, based on the number of parameters to be estimated, has made it into some recently published text books on SEM (Kelloway, 1998; Kline, 1998; Mueller, 1996). For instance, Kline (1998) suggests that, in the context of confirmatory factor analysis, N:q values in the range of 10:1 (10 observations per one estimated parameter) or even 20:1 seem appropriate. Whereas authors may offer different recommendations concerning the actual ratio needed, it is clear that the assertion that sample size should be considered in light

of the number of parameter estimates is somewhat ensconced in the knowledge-base of CSM. Though this seems like a reasonable assertion, and limitations of the previous work (Jackson, 2001) not withstanding, the assertion does not appear to have empirical support.

A question that remains is, if N:q is an important consideration in CSM, then in what ways will an inadequate N:q manifest itself? The advice of striving for a given N:q is presumably not dispensed for reasons of statistical power of an overall fit statistic, as sample size requirements relative to power can be estimated by other means. As mentioned earlier, it could be to increase the likelihood that the results will be replicated. Furthermore, it could be due to a desire to obtain more precise parameter estimates or more reliable statistical tests for individual parameter estimates. An assumption of this study is that, if a small sample size relative to the number of estimated parameters tends to result in solutions that are in some way inferior, then the effect should be detectable by examining overall measures of fit, overall statistical tests, parameter estimate variability, and parameter estimate precision.

The primary purpose of this study is to attempt to find support for the assertion that N:q is a meaningful way to think about sample size. Previous research suggests that an examination of N:q must be done in the context of other factors that influence the quality of CSM solutions, namely sample size, indicator reliability, and the number of measured variables per latent variable. My study is a simulation study and is necessarily vulnerable to limitations associated with such an approach. For instance, the variety of data idiosyncrasies and models that are presented in the literature cannot be addressed in simulation work. However, at this stage of the research in this area, and given the research question being posed, a simulation design seems to be an appropriate and defensible approach. Therefore, it was hypothesized that the ratio of the number of subjects to the number of estimated parameters would affect either the variance or bias in the parameter estimates, the values or variance of summary fit indexes, or some combination of these dependent variables. Consistent with previous findings (e.g., Anderson & Gerbing, 1984; Gerbing & Anderson, 1985; Velicer & Fava, 1998), it was hypothesized that sample size, indicator reliability, and the number of measured variables per factor would impact parameter estimate variability as well as certain fit indexes.

METHOD

Study Design

The design used for this research closely followed the one described by Jackson (2001) with two main differences. First, three levels of N:q were used in this study whereas the aforementioned study contained two; second, in the study this independent variable (N:q) had much more variability. The number of measured variables was held constant across all conditions (20 variables). In total, the design contained four independent

variables: sample size ($N = 50$, 100, 200, 400, and 800); the number of latent variables (hence the number of measured variables per latent variable); indicator reliability; and, the number of observations per estimated parameter (N:q).

There were five latent variable conditions: one latent variable having 20 indicators; two latent variables having 10 indicators each; three latent variables having seven indicators on the first two and six on the third; four latent variables having five indicators each; and five latent variables having four indicators each. All measured variables loaded on only one latent variable. Past work (Anderson & Gerbing, 1984) has shown significant effects for the number of indicators per latent variable on some overall fit indexes.

There were two conditions for indicator reliability, 0.60 and 0.80. For each condition, the population model had all loadings equal to 0.60 or 0.80 ($r^2 = 0.64$ and 0.36, respectively). Put another way, there were no mixed loading conditions. As noted previously, the reliability of the variables (strength of association between the variable and factor) has been found to influence CSM solutions (e.g., Anderson & Gerbing, 1984; Bandalos, 1997; Velicer & Fava, 1998).

In models with more than one factor, the data were generated so that the true value of the factor correlations was 0.50. This meant that, as each factor was added, the number of estimated parameters increased. With two factors, there was one additional parameter, the correlation between factor one and factor two. With three factors there were three additional parameters, with four there were six, and with five, there were 10 additional parameters. While the number of parameters to estimate was varied using this strategy, it was confounded with the number of variables per factor, since the number of variables was held constant.

The number of observations per parameter estimate was contrasted by varying another condition, namely adding constraints to the measured variable parameter estimates. Borrowing from psychometric theory, three conditions were created corresponding to three types of confirmatory factor models: congeneric; tau-equivalent; and parallel. In the congeneric condition, the model was identified by setting the variance of each latent variable to 1.0. No restrictions were placed upon factor loadings or error variances. In the tau-equivalent condition, the variance of the latent variables was set to 1.0 and, in addition, the lambda estimates (factor pattern loadings) were restricted to be equal for all 20 measured variables. Elements of the Phi matrix (factor correlations) were free to be estimated. Finally, the parallel condition had the same restrictions as the tau-equivalent condition and, in addition, restricted the error variances to be equal.

These three conditions allowed for much greater contrast in the N:q variable than in the previous study (Jackson, 2001). For instance, in the one factor congeneric condition, there were 40 parameters to be estimated (a factor loading for each of the 20 measured variables and corresponding error variances). In the tau-equivalent condition, there were 21 parameters to be estimated, the error variances for each of the 20 measured variables and the factor loading for these variables. In the

parallel condition, there were only two parameter estimates, the error variance and the factor loading for the measured variables. Staying with the one-factor example, this means that, for the smallest sample size ($N = 50$), the N:q ratio was 1.25 (1.25 observations per estimated parameter) for the congeneric condition, and 25 for the parallel condition. For the largest sample size ($N = 800$), the N:q ratio was 20 for the congeneric condition and 400 in the parallel condition.

By contrast, in the previous study (Jackson, 2001), N:q was manipulated by fixing the lambda element of one measured variable per latent variable to its true value. So, in the case of one latent variable, one additional parameter was fixed. In the five-factor condition, five additional parameters were fixed. This resulted in very little variation in the N:q variable. Staying with the one latent variable example, the N:q ratio was 1.25 for the smallest sample size in the low N:q condition and 1.28 in the high N:q condition. Even in conditions where more parameters were fixed (e.g., the five latent variable condition), there was very little variance between the high and low N:q conditions—1.00 vs. 1.11 at the smallest sample size and 16 vs. 17.78 at the largest sample size. Clearly, the present study provides a much better test of the N:q hypothesis.

Thus, this study examines the result of varying the number of parameters to be estimated by constraining factor pattern loadings and error variances to be equal. If the optimum sample size has something to do with the number of parameters being estimated, then these constraints should reveal something of this relationship. One estimation procedure, maximum likelihood, was used. The result was a Five (model) × Five (sample size) × Two (reliability) × Three (number of subjects per parameter, N:q) design.

Data Generation

The data used for this research were generated using SAS (1996) for the IBM mainframe. Data for the 20 observed variables were generated according to the models previously described. The data were generated using a random number generator incorporated in SAS (1996) that yields an asymptotic distribution with a mean of zero and standard deviation of one. Finally, 200 replications (covariance matrices) for each cell were generated. As the author did not have access to SAS's Interactive Matrix Language module, the more conventional method of sampling from a covariance matrix was not employed. Instead, data were generated using a modification of a program presented by Bernstein (1995, see Jackson, 2001, for an example). Care was taken to ensure the data generated from this program had a multivariate normal distribution.

The SAS procedure PROC CALIS (SAS, 1996) was used to fit the confirmatory factor analytic models to the data once they were generated. The statistics and parameter estimates generated by SAS were saved to a permanent file for subsequent analyses.

RESULTS

Convergence and Improper Solutions

Two of the sample covariance matrices—both involving low indicator reliability, small sample size, and fewer indicators per latent variable—resulted in SAS not converging on a solution. Results from these two replications were not considered in any further analyses. In addition, there were several instances of improper solutions (i.e., possible negative eigenvalues in the Phi matrix). These cases of improper solutions inevitably occurred with sample sizes of 50, and were most prevalent in models with more latent variables. Furthermore, there was a tendency for them to occur more often in the congeneric condition. These observations were not included in the analyses reported in this article. Furthermore, the Sums of Squares based on independently partitioned variance (Type III sums of squares in SAS) were interpreted.

Practical Significance

With the large number of observations considered in this study, many significant effects were anticipated. Because of this power, a criterion for practical significance was adopted just as in Anderson and Gerbing's study (1984), and only those effects that met the criterion, an effect size of at least 0.03 (using ω^2, e.g., Keppel, 1982), were interpreted and reported.

Measures of Fit

A number of fit indexes were analyzed as dependent variables in this study. The measures were chosen to represent the various classes of fit indexes; absolute fit indexes and Type 1, Type 2, and Type 3 incremental fit indexes. More detailed discussions of fit indexes can be found in Bollen (1989), Hu and Bentler (1995), Marsh, Balla, and Hau (1996), Muliak et al. (1989), and Tanaka (1993).

The four absolute fit indexes examined in this study were chi-square bias, root mean squared error of approximation (RMSEA; Steiger, 1990), the Goodness-of-Fit Index (GFI; Jöreskog & Sörbom, 1986), and the centrality index (CI; McDonald, 1989). With two exceptions, which will be described later in this article, the results were similar to previous findings (Jackson, 2001).

Sample size had a practically significant effect on chi-square bias ($\omega^2 = 0.237$), which was calculated by subtracting the expected value (*df*) from chi-square then dividing by the expected value. As the calculation implies, positive values of chi-square bias indicated that the obtained chi-square value was larger than expected and negative values indicated the obtained chi-square value was smaller than expected. Larger sample sizes were associated with smaller levels of

chi-square bias. The largest sample sizes were associated with slightly negative values of chi-square bias.

Additionally, sample size had a practically significant effect on RMSEA (ω^2 = 0.556), GFI (ω^2 = 0.977), and the CI (ω^2 = 0.458). In the case of RMSEA, smaller average values were associated with larger sample sizes and in the cases of CI and GFI, larger average values were associated with larger sample sizes. Using ω^2 as a guide, sample size had the only practically significant effect on GFI and CI. However, the effect of N:q on CI was close to the cutoff (ω^2 = 0.027), with higher values of N:q being associated with higher values of CI. The average CI for the congeneric condition was 0.93 and the average CI for the tau-equivalent and parallel conditions was 0.96 and 0.98, respectively. Many of the earlier mentioned effects were very similar to those found in the previous study (Jackson, 2001), thus the interested reader may consult that paper for a more detailed discussion of these effects. Table 1 contains mean and standard deviation values of CI and GFI by sample size.

The exceptions mentioned previously involved chi-square bias and RMSEA. In addition to sample size, N:q also had a practically significant effect on both of these fit measures (ω^2 = 0.075 and ω^2 = 0.034, respectively). Lower, even slightly negative values of chi-square bias were associated with higher N:q (more observations per estimated parameter). Similarly, lower average values of RMSEA were also associated with the higher N:q conditions. As can be seen in Table 2 the average chi-square bias value is lowest for the parallel model condition, which had the highest N:q ratio, followed by the tau-equivalent condition, which had the second highest N:q ratio. For the congeneric model, chi-square bias was positive for smaller sample sizes and approached zero for sample sizes of 400 and 800. In the tau-equivalent condition, the average chi-square underestimated its expected value (*df*) at sample sizes as small as 200 and in the parallel condition the average chi-square underestimated its expected value with sample sizes as small as 100. For larger sample sizes (N = 800), the chi-square value was, on average, negatively biased by as much as it was positively biased in the congeneric model at smaller

TABLE 1
Means and Standard Deviations for the Goodness-of-Fit Index (GFI) and Centrality Index (CI) for Varying Levels of Sample Size

	GFI		*CI*	
Sample Size	*M*	*SD*	*M*	*SD*
50	0.734	0.023	0.775	0.182
100	0.849	0.015	0.972	0.104
200	0.919	0.009	1.006	0.052
400	0.958	0.005	1.006	0.026
800	0.978	0.003	1.005	0.013
Total	0.888	0.089	0.953	0.132

TABLE 2
Chi-Square Bias and RMSEA for Varying Levels of Sample Size and N:q

	Chi-Square Bias				*RMSEA*			
Sample Size	*Congeneric*	*Tau-Equivalent*	*Parallel*	*Total*	*Congeneric*	*Tau-Equivalent*	*Parallel*	*Total*
M								
50	0.169	0.118	0.085	0.124	0.063	0.050	0.041	0.051
100	0.072	0.021	–0.015	0.026	0.027	0.018	0.013	0.019
200	0.032	–0.028	–0.060	–0.019	0.014	0.008	0.005	0.009
400	0.010	–0.049	–0.072	–0.037	0.008	0.004	0.003	0.005
800	–0.002	–0.063	–0.086	–0.051	0.005	0.003	0.002	0.003
Total	0.056	0.000	–0.030	0.009	0.023	0.017	0.013	0.018
SD								
50	0.092	0.096	0.096	0.101	0.023	0.025	0.026	0.026
100	0.102	0.106	0.108	0.111	0.019	0.018	0.016	0.018
200	0.107	0.110	0.113	0.116	0.013	0.010	0.009	0.011
400	0.111	0.112	0.114	0.118	0.009	0.007	0.006	0.008
800	0.109	0.117	0.116	0.119	0.006	0.005	0.004	0.005
Total	0.121	0.127	0.126	0.130	0.026	0.023	0.021	0.024

Note. RMSEA = root mean squared error of approximation.

sample sizes (e.g., $N = 100$). In addition, from examining the standard deviations in Table 2 it is apparent that the variability in RMSEA is smaller with larger sample sizes. Furthermore, it appears that it is slightly smaller with larger values of N:q relative to smaller values of N:q.

The three incremental fit indexes examined in this study were chosen to represent the three categories of these types of indexes; Type 1, Type 2, and Type 3 (see Hu & Bentler, 1995). The Normed fit index (NFI; Bentler & Bonett, 1980), the Nonnormed fit index (NNFI; Tucker & Lewis, 1973; Bentler & Bonett, 1980), and the comparative fit index (CFI; Bentler, 1990) were used to represent Type 1, Type 2, and Type 3 indexes, respectively. Results from analyses of the incremental fit indexes are similar to findings from the previous study (Jackson, 2001). For both the Type 1 and Type 3 fit indexes (NFI and CFI), there was a practically significant effect for sample size ($\omega^2 = 0.715$ and $\omega^2 = 0.431$, respectively), reliability of the indicators ($\omega^2 = 0.168$ and $\omega^2 = 0.049$, respectively), and the interaction between the two ($\omega^2 = 0.067$ and $\omega^2 = 0.089$, respectively). For the Type 2 index, NNFI, there was a practically significant effect for sample size and the interaction between sample size and the reliability of indicators ($\omega^2 = 0.363$ and 0.075, respectively). The main effect for indicator reliability did not quite reach practical significance ($\omega^2 = 0.025$). Means and standard deviations for the NFI, NNFI, and CFI, by sample size and indicator reliability, can be found in Table 3. In general, fit indexes were higher as sample size increased for each type of fit index and fit indexes were higher for conditions where the

TABLE 3
Fit Index Values for the NFI, NNFI, and CFI by Sample Size and Indicator Reliability

	Indicator Reliability								
	NFI			*NNFI*			*CFI*		
Sample Size	*0.60*	*0.80*	*Total*	*0.60*	*0.80*	*Total*	*0.60*	*0.80*	*Total*
M									
50	0.535	0.760	0.649	0.891	0.959	0.925	0.892	0.959	0.926
100	0.716	0.875	0.795	0.984	0.995	0.990	0.974	0.991	0.983
200	0.839	0.936	0.888	1.001	1.001	1.001	0.992	0.997	0.995
400	0.914	0.968	0.941	1.002	1.001	1.002	0.997	0.999	0.998
800	0.955	0.984	0.970	1.002	1.001	1.001	0.999	1.000	0.999
Total	0.793	0.905	0.849	0.976	0.991	0.984	0.971	0.989	0.980
SD									
50	0.069	0.040	0.126	0.091	0.037	0.077	0.076	0.031	0.067
100	0.050	0.021	0.088	0.048	0.017	0.037	0.032	0.011	0.025
200	0.030	0.011	0.054	0.023	0.008	0.017	0.013	0.004	0.010
400	0.017	0.006	0.030	0.012	0.004	0.009	0.006	0.002	0.004
800	0.009	0.003	0.016	0.006	0.002	0.004	0.003	0.001	0.002
Total	0.157	0.084	0.138	0.064	0.025	0.049	0.055	0.021	0.042

Note. NFI = Normed Fit Index; NNFI = Nonnormed Fit Index; CFI = comparative fit index.

indicator reliability was higher. In addition, the standard deviation of the fit indexes was smaller for larger sample sizes and higher indicator reliability conditions. Finally, the Type 1 fit index (NFI) performed more poorly than the Type 2 and Type 3 indexes at each level of sample size and in each indicator reliability condition. In short, it was more prone to underestimate its maximum value. This finding has been reported in previous work (e.g., Marsh et al., 1996).

Average fit values for each of the incremental fit indexes examined in this study were higher with larger sample sizes and greater indicator reliability. In addition, values of these fit indexes tended to increase more dramatically with increasing sample sizes under the lower reliability conditions. It should be noted that, though N:q did not have a practically significant effect on the incremental fit indexes, it did have a statistically significant effect on them and it accounted for a greater proportion of variance in this study than in the previous study for the Type 2 and Type 3 indexes ($\omega^2 = .029$ and 0.011, respectively). In both cases, there was a tendency for models with higher N:q values to have higher fit indexes.

This was most notable for the NNFI. For the congeneric condition, the mean NNFI was 0.974. For the tau-equivalent and parallel conditions, the mean NNFI value was 0.984 and 0.994, respectively. In addition, the interaction between N:q and sample size was statistically significant for NNFI, but had a practical significance value shy of the cutoff ($\omega^2 = .025$) as did the main effect for indicator reli-

ability (ω^2 = .029). As would be expected, conditions with higher indicator reliability resulted in higher average NNFI values (0.991 vs. 0.976). The interaction between N:q and sample size resulted from a slight tendency for NNFI in high N:q conditions to increase less with increasing sample size than for NNFI in lower N:q conditions. This appeared to be due to the fact that average NNFI values in high N:q conditions with small sample sizes were closer to their true value of 1.0 than average NNFI values in low N:q conditions and small sample sizes.

Parameter Estimates

Because of the way N:q was manipulated in the current design, certain aspects of the lambda estimates were not directly comparable across N:q conditions. Namely, the variance among lambda estimates could not be analyzed because in the tau-equivalent and parallel conditions, all lambda estimates were constrained to be equal. However, it was possible to compare the average bias in lambda estimates for the congeneric models with the bias in estimates for the tau-equivalent and parallel models. Furthermore, since no restrictions were placed on the correlations among the latent variables, both the bias and variance of phi estimates could be analyzed.

First, the average bias of lambda estimates was examined. Parameter bias was measured by subtracting the true parameter value from the observed parameter estimate (or average parameter estimate in the case of the congeneric model) and dividing by the true value of the parameter. The average across the 20 measured variables for the congeneric condition was compared to the lambda estimates from the tau-equivalent and parallel conditions. The results of analyzing the bias in lambda estimates was that there was no practically significant effect for any of the independent variables considered in this study or interactions among those independent variables. The effect for N:q came the closest to being practically significant ($\omega^2 = 0.012$). This magnitude of effect corresponds to small differences in the third decimal place of actual parameter estimates.

The fact that Phi estimates were not fixed in any way allowed for comparisons across the three N:q conditions. Since the true population values for all of these parameters was set to 0.50, any consistent variation or bias would be due to the independent variables, not differences in the true model parameters. There were no practically significant effects for phi parameter estimate bias. However, there were practically significant effects for the variation in these parameter estimates, as measured by the standard deviation of the estimates for each model fitted, using only those models with three or more factors. These practically significant effects were sample size ($\omega^2 = 0.304$) and indicator reliability ($\omega^2 = 0.033$). The variation among these parameter estimates was lower for the high indicator reliability condition, with the high indicator reliability condition yielding factor correlation standard deviations approximately 28% lower than low indicator reliability conditions. Additionally, smaller sample sizes were associated with greater variation in Phi parame-

ter estimates. The standard deviation of the factor correlations in the largest sample size ($N = 800$) was approximately four times smaller than in the smallest sample size ($N = 50$). Mean standard deviation for the Phi estimates can be found in Table 4.

DISCUSSION

The current research was designed to study the effect of varying sample size relative to the number of estimated parameters on a confirmatory factor analysis solution. The design was similar to the one used by Jackson (2001), with an important exception being that the N:q condition was varied more in this study. As an example, consider the model with one factor and 20 indicators. In the congeneric condition, there were 40 parameters to be estimated, one for each factor to variable path and one for the error variance associated with each manifest variable. At the other extreme, the parallel condition, there were only two parameters to be estimated, a single error variance and a single factor to variable path. This is because all factor loadings were constrained to be equal and all error variances were constrained to be equal. This meant that for the largest sample size ($N = 800$), in the congeneric condition there was a ratio of 20 observations per parameter estimate and for the parallel condition there were 20 times that many, or 400 observations per parameter estimate. By contrast, the variability of the sample size independent variable was less, with the largest condition being 16 times greater than the smallest condition ($N = 50$ being the smallest and $N = 800$ being the largest). Still, given this variability, N:q only had two practically significant effects, both on fit indexes, while absolute sample size had more and greater practically significant effects.

This study found some support for the proposition that sample size be considered in terms of N:q. However, in the current design, sample size had a much more profound effect although it varied less relative to N:q. The explanation for why N:q had a practically significant effect on chi-square bias and RMSEA appears to have

TABLE 4
Standard Deviations of Factor Correlation Estimates for Sample Size and Indicator Reliability

	Indicator Reliability		
Sample Size	*0.60*	*0.80*	*Total*
50	0.140	0.099	0.120
100	0.095	0.070	0.082
200	0.066	0.048	0.057
400	0.047	0.033	0.040
800	0.033	0.024	0.028
Total	0.076	0.055	0.065

to do with the use of degrees of freedom in their calculation. Degrees of freedom figures heavily in calculating chi-square bias and is taken into account in calculating RMSEA. An examination of the maximum likelihood fit statistic, on which chi-square is based, revealed that the congeneric model actually had the best fit, followed by the tau-equivalent and parallel models. This means that the added constraints of the tau-equivalent and parallel models served to deteriorate overall fit; however, the increase in degrees of freedom more than offset this deterioration for the two measures in question. This also explains why the effect of N:q on NNFI and CI came close to achieving practical significance, as degrees of freedom is also a component in their calculation. Because of the limitations of Monte Carlo simulations, it is very much premature to conclude that conventional wisdom be completely overthrown with respect to the N:q hypothesis. Its merit, however, certainly appears to be in need of more theoretical and empirical support. Furthermore, the general advice of basing sample size on some minimum value (e.g., 200 or more observations), ensuring indicators are carefully chosen and reliable, and ensuring there are an adequate number of indicators per latent variable seemingly provide more supportable guidelines for sample size than N:q (e.g., Anderson & Gerbing, 1984; Cohen, Cohen, & Velez, 1990; Gerbing & Anderson, 1985; Jackson, 2001).

More generally, however, it is important that the question of what is an appropriate sample size be further investigated. Structural equation modeling is based on asymptotic statistical theory. If certain assumptions are met and the sample size is large enough, statistical tests and parameter estimates can be trusted. Unfortunately, there is no easy number one can substitute for the phrase "large enough." Some rules of thumb appear to be in use; for example, 100 to 200 observations is a medium sample size (Kline, 1998). However, the N:q hypothesis appears to be a manifestation of an underlying assumption that sample size perhaps shouldn't be thought of in an absolute sense. Rather, features of a model, which the researcher is testing, should moderate this figure.

The problem, at this point, represents more than an esoteric argument over how best to conduct CSM. Researchers must justify their sample size when submitting their work to conferences or journals for presentation or publication. The practicalities and expense of obtaining a "large enough" sample come into play in the research design. Additionally, editors and reviewers must consider the research design when evaluating a manuscript for publication. This, naturally, includes making a judgement about the adequacy of the sample size. In short, firmer guidelines on sample size would be welcomed by researchers and reviewers alike.

In summary, this article describes research aimed at discovering support for the often-cited assertion that the appropriate sample size should be considered in light of the number of parameters being estimated. Some support was found for this assertion. Most notably, in conditions with higher N:q ratios, on average chi-square bias and RMSEA values were lower. Additionally, other fit indexes demonstrated a tendency to be impacted by N:q. Namely, NNFI and CI values were, on average,

larger with higher N:q ratios. The effect size for the latter two, however, was not as great as it was for chi-square bias and RMSEA. Given that this study offers some support for the N:q hypothesis, future research should be directed at better understanding these effects under varying conditions such as under different levels of approximation error or in testing models with structural parameters. Furthermore, research aimed at disentangling the relative impact of sample size, reliability of indicators, number of indicators per latent variable, N:q, and other determinants of model solutions over a variety of modeling conditions seems warranted.

ACKNOWLEDGMENT

I wish to thank Peter M. Bentler for his suggestions on the design of this study.

REFERENCES

Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, *49*, 155–173.

Bandalos, D. L. (1997). Assessing sources of error in structural equation models: The effects of sample size, reliability, and model misspecification. *Structural Equation Modeling*, *4*, 177–192.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238–246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88*, 588–606.

Bentler, P. M., & Chou, C. -P. (1987). Practical issues in structural modeling. *Sociological Methods & Research, 16*, 78–117.

Bernstein, I. H. (1995, April). *Simulation in teaching multivariate statistics*. Paper presented at the Workshop for the Society of Applied Multivariate Research, San Antonio, TX.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage.

Cohen, P., Cohen, J., & Velez, C. N. (1990). Problems in the measurement of latent variables in structural equations causal models. *Applied Psychological Measurement*, *14*, 183–196.

Cudeck, R., & Henly, S. J. (1991). Model selection in covariance structures analysis and the "problem" of sample size: A clarification. *Psychological Bulletin*, *109*, 512–519.

Gerbing, D. W., & Anderson, J. C. (1985). The effects of sampling error and model characteristics on parameter estimation for maximum likelihood confirmatory factor analysis. *Multivariate Behavioral Research*, *20*, 255–271.

Fan, X., & Wang, L. (1998). Effects of potential confounding factors on fit indices and parameter estimates for true and misspecified SEM models. *Educational & Psychological Measurement*, *58*, 701–736.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation methods, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling*, *6*, 56–83.

Hu, L. -T., & Bentler, P. M. (1995). Evaluating model fit. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 76–99). Thousand Oaks, CA: Sage.

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

Jöreskog, K. G., & Sörbom, D. (1986). *LISREL VI: Analysis of linear structural relationships by maximum likelihood and least squares methods*. Mooresville, IN: Scientific Software, Inc.

Kelloway, E. K. (1998). *Using LISREL for structural equation modeling: A researcher's guide.* Thousand Oaks, CA: Sage.

Keppel, G. (1982). *Design & analysis: A researcher's handbook* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford.

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, *1*, 130–149.

Marsh, H. W., Balla, J. R., & Hau, K. -T. (1996). An evaluation of incremental fit indices: A clarification of mathematical and empirical properties. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness of fit indexes in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin, 103*, 391–410.

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.

McDonald, R. P. (1989). An index of goodness-of-fit based on noncentrality. *Journal of Classification*, *6*, 97–103.

Mueller, R. O. (1996). *Basic principles of SEM: An introduction to LISREL and EQS*. New York: Springer.

Mueller, R. O. (1997). Structural equation modeling: Back to the basics. *Structural Equation Modeling, 4*, 353–369.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, D. (1989). Evaluation of goodness-of-fit indices for structural equation models. *Psychological Bulletin, 105*, 430–445.

SAS Institute, Inc. (1996). *SAS/STAT Software: Changes and enhancements through release 6.11*. Cary, NC: Author.

Steiger, J. H. (1990). Structural model evaluation and modification: An interval estimation approach. *Multivariate Behavioral Research*, *25*, 173–180.

Tanaka, J. S. (1987). "How big is big enough?": Sample size and goodness of fit in structural equation models with latent variables. *Child Development*, *58*, 134–146.

Tanaka, J. S. (1993). Multifaceted conceptions of fit in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 10–39). Newbury Park, CA: Sage.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika, 38*, 1–10.

Ullman, J. B. (1996). Structural equation modeling. In B. G. Tabachnick & L. S. Fidell (Eds.), *Using multivariate statistics* (pp. 709–811). New York: HarperCollins.

Velicer, W. F., & Fava, J. L. (1998). Effects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3*, 231–251.

West, S. G., Finch, J. F., & Curran, P. J. (1995). Structural equation models with nonnormal variables: Problems and remedies. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues and applications* (pp. 56–75). Thousand Oaks, CA: Sage.