Journal of Modern Applied Statistical Methods

Volume 9 | Issue 2 Article 5

11-1-2010

Recommended Sample Size for Conducting Exploratory Factor Analysis on Dichotomous Data

Robert H. Pearson
University of Northern Colorado, robert.pearson@unco.edu

Daniel J. Mundform
New Mexico State University, mundfrom@nmsu.edu

Journal of Modern Applied Statistical Methods
November 2010, Vol. 9, No. 2, 359-368

1538 – 9472/10/$95.00

REGULAR ARTICLES

Recommended Sample Size for Conducting Exploratory Factor Analysis on Dichotomous Data

Robert H. Pearson
University of Northern Colorado,
Greeley, CO USA

Daniel J. Mundfrom
New Mexico State University,
Las Cruces, NM USA

Minimum sample sizes are recommended for conducting exploratory factor analysis on dichotomous data. A Monte Carlo simulation was conducted, varying the level of communalities, number of factors, variable-to-factor ratio and dichotomization threshold. Sample sizes were identified based on congruence between rotated population and sample factor loadings.

Key words: Exploratory Factor Analysis, dichotomous data, sample size.

Introduction

Selecting a sample size is one of the most important decisions to be made when planning an empirical study. Often the choice is based on the minimum necessary sample size to obtain reliable results from the statistical procedures to be conducted. For many procedures (e.g., *t*-test, *F*-test) an exact minimum can be found which will allow relationships in the population (if they exist) to be detected with high probability. The issue of sample size for exploratory factor analysis (EFA) is not as straightforward, however, because an exact minimum cannot easily be found analytically and because the procedure's use involves a greater degree of subjectivity.

Although factor analysis has been used in a vast array of scientific fields, it is most frequently used as a tool to investigate the structure of scores obtained via psychometric measures. Such research seeks to identify and possibly measure a small number of unobservable traits that are hypothesized to explain a large portion of the covariation among observed variables. The statistical problem for EFA is the estimation of communalities and - perhaps more importantly - factor loadings. If the results of a factor analysis are to be useful beyond a particular study, then the estimated loadings must be reasonable approximations of true population loadings. Thus, reliable guidelines for selecting a sample size that is likely to produce a factor solution which closely matches a population factor structure would be a boon to researchers planning factor analytic studies.

Until recently, most of the published sample size recommendations were simplified rules based on experts' experience. Several of the most frequently cited guidelines are absolute numbers. Gorsuch (1983) and Kline (1994) suggested sampling at least 100 subjects. Comrey and Lee (1992) provided the following scale of sample size adequacy: 50 – very poor, 100 – poor, 200 – fair, 300 – good, 500 – very good, and 1,000 or more – excellent. Authors have also proposed minimum ratios of sample size to the number of variables (*n*:*p*). Cattell (1978) suggested three to six subjects per variable, Gorsuch (1983) suggested this ratio be at least five and both Everitt (1975) and

Robert H. Pearson is an Assistant Professor of Applied Statistics in the Department of Applied Statistics & Research Methods. His research interests include multivariate statistics, statistical computing and factor analysis. Email: robert.pearson@unco.edu. Daniel J. Mundfrom is an Associate Professor of Applied Statistics in the Department of Economics. His research interests include multivariate statistics, statistical methods and applications of statistics. Email: mundfrom@nmsu.edu.

Nunnally (1978) recommended sampling at least ten times as many subjects as variables.

MacCallum, Widaman, Zhang, and Hong (1999) demonstrated mathematically and empirically that sample size requirements are contingent upon two aspects of the factor structure. Specifically, they showed that both mathematical overdetermination (the extent to which the common factors are sufficiently represented by an adequate number of variables) and the size of communalities have a considerable effect on the agreement between sample and population factor loadings. In a Monte Carlo study they showed that communality had an estimated effect size ($\hat{\omega}^2$) nearly three times greater than sample size and overdetermination had an effect nearly as large as sample size. Mundfrom, Shaw and Ke (2005) subsequently provided sample size recommendations for 180 population conditions on the basis of a Monte Carlo study that varied the number of factors, the ratio of variables to factors (an important aspect of overdetermination) and communalities.

In practice, data are often measured on ordinal or nominal scales, particularly in the social sciences (Hip & Bollen, 2006; Lee & Song, 2003; Schoenberg & Arminger, 1989). Exploratory factor analysis is often applied to ordinal or dichotomous data to examine their relationship with underlying factors (Baños & Franklin, 2002; Mundfrom, Bradley, & Whiteside, 1994; Tomás-Sábado & Gómez-Benito, 2005). Many authors have suggested other approaches for this situation (Bartholomew & Knott, 1999; Bock & Aitkin, 1981; Muthén, 1978), however, a traditional factor analysis can be useful as long as a meaningful and interpretable set of factors can be identified, regardless of the measurement level of the input data. Johnson and Wichern (2002) refer to this as the WOW criterion: "If, while scrutinizing the factor analysis, the investigator can shout 'Wow, I understand these factors,' the application is deemed successful" (p. 524).

Darlington (1997) described this use of factor analysis as heuristic rather than absolute. It is understood that any factor solution is only one among many that are possible. If the retained factor structure can be cross-validated or together with other evidence supports a broader theory, then the analysis is successful. Mulaik (1989) discussed how this approach fits with theory development throughout science:

> Theoretical physics, for example, is continuously occupied with differing speculations designed to synthesize the same sets of diverse experimental data. All of these differing theoretical speculations may yield models that fit equally well the data already at hand, but in time some or all of these speculative models may be eliminated from further consideration by their inconsistency with new data obtained to test certain predictions derived from them. (p. 54)

For a factor solution to be replicable across studies it must represent a structure that truly exists in the population.

The primary purpose of this study was to provide sample size recommendations for researchers who are planning factor analytic studies that will involve dichotomous variables. It was also of interest to compare the results of this study to requirements for continuous data (Mundfrom, et al., 2005). From a methodological standpoint, the extent to which these results differ from those found by Mundfrom, et al. (2005) lends insight into the effect that scale of measurement has on this statistical procedure. Because the case of dichotomous data is the most extreme departure from continuity, these recommendations represent an upper bound for minimum necessary sample size. Therefore, these recommendations were also intended to serve as conservative guidelines for EFA of ordinal data.

Methodology

Monte Carlo simulation was used for this study. Population data were generated using the SAS System v9.1.3 (SAS Institute Inc., 2007). One-hundred matrices of dichotomous data, each conceptually representing a unique population of 100,000 observations on p variables, were generated for each condition determined by four manipulated variables: the number of common factors (m), the variable-to-factor ratio (p:m), the

variable communalities and the dichotomization threshold. Populations were randomly generated using the following two-stage process.

In the first stage, the procedure described by Tucker, Koopman, and Linn (1969) was used to randomly generate population correlation matrices with specified factor structures. A total of 180 factor structures were investigated by crossing the number of factors ($1 \leq m \leq 6$), the variable-to-factor ratio ($3 \leq p{:}m \leq 12$), and the variable communalities. Three levels of variable communalities were examined: high, in which communalities were randomly assigned values of 0.6, 0.7 or 0.8; wide, in which they could have values from 0.2 to 0.8 in increments of 0.1; and low, in which they could have values of 0.2, 0.3, or 0.4 (Tucker, Koopman & Linn, 1969). Ten correlation matrices were generated for each factor structure.

In the second stage, ten matrices of binary data were generated from each population correlation matrix (R). Each data matrix consisted of 100,000 rows of values on p dichotomous variables. First, a matrix X was created by taking the product of the Cholesky root of R and a matrix of multivariate-normal deviates. Elements of each column of X were then dichotomized according to three conditions. In the first condition, all variables were dichotomized to have a 50/50 split. This condition results in the smallest amount of information loss due to dichotomization (Cohen, 1983) and can be considered the best case. In the second condition, all variables were dichotomized to have an 80/20 split. This condition was used in simulation studies by Parry and McArdle (1991) and Weng and Cheng (2005), and is similar to the 84/16 split used by Bernstein and Teng (1989) which they likened to item distributions found in symptom description scales such as in the MMPI or a difficult ability test. In the remaining condition, half of the variables were dichotomized using an 80/20 split and half using a 50/50 split.

Because differences in item means limit the maximum possible value of the product-moment correlation it was important to investigate the resulting effect on factor loading estimates. As a result, one-hundred population data matrices (hereafter referred to as populations) were generated for each combination of communality level, number of factors, variable-to-factor ratio and dichotomization threshold.

Each population was factor analyzed using maximum likelihood estimation and varimax rotation. One-hundred simple random samples of a specific size were then selected from each population. If a sample correlation matrix was non-positive-definite, another was generated and used instead. Each sample was factor analyzed and the rotated factor loadings were compared to those in the population using a coefficient of congruence.

Sample sizes were chosen by first starting with a sample size that was too small based on the recommendations of Mundfrom, et al. (2005). Sample sizes were then increased systematically according to the following algorithm:

- while $n < 30$, it was increased by 1;
- while $30 \leq n < 100$, it was increased by 5;
- while $100 \leq n < 300$, it was increased by 10;
- while $300 \leq n < 500$, it was increased by 20;
- while $500 \leq n < 1,000$, it was increased by 50;
- while $n \geq 1,000$, it was increased by 200.

This system of increments is nearly identical to that used by Mundfrom, et al. (2005). The procedure was stopped when the sample and population correlation matrices met criteria based on a coefficient of congruence. These criteria are defined below. The procedure was also stopped if a sample size greater than 5,000 was necessary.

In summary, a $3 \times 6 \times 10 \times 3$ factorial design was implemented, corresponding to the experimental variables communality level, number of factors, variable-to-factor ratio, and dichotomization threshold, resulting in a total of 540 population conditions. One-hundred populations were randomly generated for each population condition and 100 samples were taken from each population for every sample size considered. Thus, a total of 10,000 samples

were taken for each population condition and sample size combination.

Coefficient of Congruence

A coefficient of congruence was calculated to assess the degree of correspondence between the sample and population solutions (MacCallum, et al., 1999; Tucker, et al., 1969). The coefficient for the k^{th} factor was calculated using the formula:

$$\phi_k = \frac{\sum_{j=1}^{p} \lambda_{jk(s)} \lambda_{jk(t)}}{\sqrt{\left(\sum_{j=1}^{p} \lambda_{jk(s)}^2\right)\left(\sum_{j=1}^{p} \lambda_{jk(t)}^2\right)}},$$

where $\lambda_{jk(t)}$ is the true population factor loading for variable j on factor k, and $\lambda_{jk(s)}$ is the corresponding sample loading. To assess the degree of congruence for a given solution, the mean value of ϕ_k across the m factors was computed and denoted K. For any solution with m factors there were $m!$ possible arrangements of the factors and therefore $m!$ possible values of K. The maximum value of K was used for each solution, thus representing the sample solution that was most similar to the targeted population solution.

For each population, 100 samples were taken and factor analyzed, resulting in 100 values of K. The fifth percentile of these coefficients, denoted K_{95}, was used to represent the lower bound of a 95% confidence interval for a particular population. Subsequently, 100 values of K_{95} were obtained for each population condition, corresponding to the 100 generated populations.

MacCallum, et al. (1999) provided the following guidelines for interpreting values of the coefficient of congruence: 0.98 to 1.00 = excellent, 0.92 to 0.98 = good, 0.82 to 0.92 = borderline, 0.68 to 0.82 = poor, and below 0.68 = terrible. Because the purpose of this study was to determine minimum recommended sample sizes, only those that provided good and excellent levels of agreement were retained. For a given population condition and sample size, the proportions of K_{95}s that were greater than 0.92 and 0.98 were respectively denoted P_{92} and P_{98}.

For a particular condition, a sample size was determined to meet the good criterion if either of the following occurred (Mundfrom, et al., 2005):

- The P_{92} from three successive sample sizes was at least 0.95.
- The P_{92} from two successive sample sizes was at least 0.95, the P_{92} from the next sample size was less than 0.95 and the P_{92} from the next two successive sample sizes was at least 0.95.

The same system was used to select a sample size to meet the excellent criterion. Thus, for every population condition, two sample sizes were chosen as recommendable according to the two criteria.

Results

Minimum necessary sample sizes were identified using a Monte Carlo simulation that manipulated four population characteristics. Factor structures were determined by crossing three levels of communality (high, wide and low), six numbers of factors (1 to 6), and ten variable-to-factor ratios (3 to 12). The three variable distributions considered were 50/50, 80/20 and a third distribution, hereafter referred to as mix, for which half the variables had a 50/50 split and half had an 80/20 split. The minimum necessary sample sizes for each of the 540 population conditions and two agreement criteria are presented in Tables 1, 2, and 3 for the high, wide and low levels of communality respectively.

A few cautions should be observed when interpreting these results. First, the methodology employed did not consider sample sizes beyond 5,000, so this was an artificial ceiling in this study. Secondly, frequent computational errors occurred for conditions when the p:m ratio was three: all results for these conditions should be interpreted cautiously. In addition, the three conditions involving one-factor models with p:m = 3 could not be run by SAS PROC FACTOR with maximum likelihood estimation. Thirdly, the observed results for the mix condition were unstable for models with four to six factors. This instability may be an

Table 1: Minimum Sample Size for Two Agreement Criteria - High Level of Communality

p:m	*Excellent (0.98) Criterion*						*Good (0.92) Criterion*					
	F1	*F2*	*F3*	*F4*	*F5*	*F6*	*F1*	*F2*	*F3*	*F4*	*F5*	*F6*
50/50 Variable Distribution												
3	.	1,200	3,000	5,000	5,000	5,000	.	400	1,400	3,800	5,000	5,000
4	120	270	750	1,600	5,000	5,000	40	90	380	800	3,600	5,000
5	80	280	460	1,800	5,000	5,000	35	85	180	550	2,600	5,000
6	75	250	500	650	1,800	2,600	28	85	200	250	650	700
7	70	250	340	750	1,000	1,200	26	85	120	360	340	400
8	60	270	260	500	1,800	1,000	23	100	90	170	340	460
9	55	320	200	400	1,200	1,400	22	95	65	150	300	700
10	65	260	200	290	480	1,400	25	75	70	110	140	420
11	55	200	220	440	380	800	22	85	75	150	130	250
12	50	160	250	400	550	900	20	60	100	150	170	280
80/20 Variable Distribution												
3	.	2,000	5,000	5,000	5,000	5,000	.	420	5,000	5,000	5,000	5,000
4	230	750	1,600	5,000	5,000	5,000	75	320	900	3,200	3,800	5,000
5	170	900	1,200	2,400	4,400	5,000	65	340	400	900	1,400	4,600
6	150	360	800	2,400	3,800	5,000	55	120	250	500	1,400	2,000
7	130	340	1,200	1,600	3,200	2,200	55	120	420	950	1,200	1,600
8	120	270	650	1,600	2,000	2,000	50	110	230	300	650	900
9	120	240	700	800	1,600	1,800	50	75	190	420	500	650
10	100	320	400	600	950	1,400	45	100	180	200	360	380
11	100	240	440	800	1,400	1,000	45	75	150	290	460	380
12	95	400	700	1,200	850	1,400	45	120	180	320	250	460
Half 50/50 and Half 80/20												
3	.	5,000	2,200	5,000	5,000	5,000	.	4,200	800	5,000	5,000	5,000
4	180	2,000	5,000	5,000	5,000	5,000	55	600	4,000	5,000	5,000	5,000
5	130	480	1,400	2,400	5,000	5,000	40	300	550	1,400	1,400	5,000
6	120	480	1,000	3,200	4,200	5,000	45	190	380	2,200	1,800	3,400
7	95	480	950	1,400	1,600	3,200	40	160	320	460	600	850
8	95	260	500	2,600	1,800	3,000	40	85	180	1,200	600	1,200
9	85	200	340	600	1,200	3,200	35	65	140	240	360	650
10	85	180	340	480	1,800	3,800	35	60	120	160	550	1,200
11	75	140	320	380	1,800	3,600	27	50	100	140	900	750
12	80	190	240	440	650	1,800	30	55	80	150	220	550

Note: F1 denotes one-factor models, F2 two-factor models, etc.

Table 2: Minimum Sample Size for Two Agreement Criteria - Wide Level of Communality

p:m	*Excellent (0.98) Criterion*						*Good (0.92) Criterion*					
	F1	*F2*	*F3*	*F4*	*F5*	*F6*	*F1*	*F2*	*F3*	*F4*	*F5*	*F6*
50/50 Variable Distribution												
3	.	4,000	5,000	5,000	5,000	5,000	.	1,800	5,000	5,000	5,000	5,000
4	700	1,400	5,000	5,000	5,000	5,000	200	480	2,400	5,000	5,000	5,000
5	320	1,400	5,000	5,000	5,000	5,000	95	480	1,400	5,000	5,000	4,600
6	250	950	1,600	2,800	4,000	3,600	75	380	550	1,000	2,200	1,400
7	280	360	1,000	1,600	5,000	5,000	90	180	360	550	1,600	1,600
8	150	460	600	1,400	3,600	3,800	50	190	210	380	1,800	1,400
9	210	650	600	1,800	1,200	2,200	65	170	230	460	420	850
10	150	420	600	1,600	1,400	1,600	55	150	220	550	420	550
11	140	320	700	1,200	1,600	1,600	45	110	210	320	460	550
12	170	440	500	700	950	1,600	55	140	170	180	320	550
80/20 Variable Distribution												
3	.	5,000	5,000	5,000	5,000	5,000	.	5,000	5,000	5,000	5,000	5,000
4	650	5,000	5,000	5,000	5,000	5,000	180	2,000	5,000	5,000	5,000	5,000
5	500	2,000	3,800	5,000	5,000	5,000	160	850	1,400	5,000	5,000	5,000
6	440	1,200	2,000	5,000	5,000	5,000	140	460	500	3,000	5,000	5,000
7	340	1,800	1,800	2,800	4,400	5,000	110	550	600	800	1,600	3,200
8	340	950	1,200	3,000	2,800	4,400	110	270	420	700	1,400	1,600
9	320	550	1,000	1,400	2,600	5,000	100	230	300	550	750	2,000
10	240	550	1,000	1,600	2,200	3,600	85	200	360	550	750	1,400
11	220	400	850	1,200	1,600	2,200	75	130	270	360	480	650
12	210	420	650	950	1,600	1,800	70	140	180	320	460	600
Half 50/50 and Half 80/20												
3	.	4,200	5,000	5,000	5,000	5,000	.	2,200	4,200	5,000	5,000	5,000
4	600	1,800	5,000	5,000	5,000	5,000	200	1,200	5,000	5,000	5,000	5,000
5	290	900	5,000	5,000	5,000	5,000	90	460	3,800	5,000	5,000	5,000
6	300	750	3,600	5,000	5,000	5,000	85	300	1,400	1,200	1,800	5,000
7	210	700	900	5,000	5,000	5,000	70	200	420	2,000	2,800	2,200
8	210	850	1,600	5,000	2,800	5,000	70	300	360	1,200	1,200	2,400
9	210	1,200	650	2,600	2,600	3,000	70	380	220	900	1,200	1,600
10	180	750	800	1,200	1,400	3,000	55	260	250	460	550	850
11	190	500	750	1,600	2,000	5,000	65	180	280	420	600	1,600
12	280	700	1,000	1,200	3,600	3,600	85	240	240	340	1,200	1,400

Note: F1 denotes one-factor models, F2 two-factor models, etc.

Table 3: Minimum Sample Size for Two Agreement Criteria - Low Level of Communality

p:m	*Excellent (0.98) Criterion*						*Good (0.92) Criterion*					
	F1	*F2*	*F3*	*F4*	*F5*	*F6*	*F1*	*F2*	*F3*	*F4*	*F5*	*F6*
50/50 Variable Distribution												
3	.	5,000	5,000	5,000	5,000	5,000	.	5,000	5,000	5,000	5,000	5,000
4	950	3,000	5,000	5,000	5,000	5,000	280	1,200	2,000	5,000	5,000	5,000
5	900	5,000	3,800	5,000	5,000	5,000	270	1,800	1,600	5,000	5,000	5,000
6	650	2,600	3,600	5,000	5,000	5,000	200	1,200	1,400	3,600	5,000	5,000
7	460	2,400	1,600	3,000	5,000	5,000	140	750	600	1,200	5,000	2,800
8	400	950	2,200	5,000	5,000	5,000	120	340	700	1,800	5,000	5,000
9	380	1,400	2,600	2,800	5,000	3,400	120	480	900	1,000	1,600	1,400
10	380	600	1,800	2,200	3,200	4,200	110	180	750	1,000	1,200	1,600
11	340	850	1,400	1,800	5,000	3,200	95	260	400	500	5,000	1,200
12	290	1,000	1,600	2,000	5,000	5,000	85	320	700	700	2,400	2,600
80/20 Variable Distribution												
3	.	5,000	5,000	5,000	5,000	5,000	.	5,000	5,000	5,000	5,000	5,000
4	1,800	5,000	5,000	5,000	5,000	5,000	550	2,600	3,800	5,000	5,000	5,000
5	2,000	5,000	5,000	5,000	5,000	5,000	550	2,600	5,000	5,000	5,000	5,000
6	1,200	2,200	5,000	5,000	5,000	5,000	320	750	2,600	5,000	5,000	5,000
7	800	2,600	2,800	5,000	5,000	5,000	230	650	1,200	2,000	2,800	5,000
8	700	1,800	5,000	5,000	5,000	5,000	200	480	5,000	5,000	3,000	5,000
9	700	1,600	3,400	4,400	4,600	5,000	200	600	1,000	1,800	2,000	4,600
10	600	1,800	3,400	2,400	5,000	5,000	180	650	1,200	800	2,400	2,600
11	550	1,400	2,800	2,800	4,400	5,000	160	420	650	950	1,600	3,200
12	550	1,000	1,200	2,400	4,400	4,400	160	360	1,000	850	1,600	1,600
Half 50/50 and Half 80/20												
3	.	5,000	5,000	5,000	5,000	5,000	.	5,000	5,000	5,000	5,000	5,000
4	2,000	5,000	5,000	5,000	5,000	5,000	600	5,000	3,000	5,000	5,000	5,000
5	950	5,000	5,000	5,000	5,000	5,000	260	2,600	5,000	5,000	5,000	5,000
6	700	1,800	5,000	5,000	5,000	5,000	220	700	5,000	5,000	5,000	5,000
7	550	1,800	5,000	5,000	5,000	5,000	170	500	5,000	3,000	3,800	2,800
8	550	1,600	2,600	5,000	5,000	5,000	170	600	1,200	1,600	2,600	2,800
9	420	1,400	2,400	5,000	5,000	5,000	130	460	950	2,000	2,800	3,800
10	460	1,200	5,000	2,400	5,000	5,000	140	400	1,800	850	4,400	2,000
11	400	1,000	2,800	2,600	4,000	5,000	120	260	1,000	950	1,600	2,800
12	360	2,800	1,800	4,000	2,800	5,000	110	700	650	2,600	950	1,800

Note: F1 denotes one-factor models, F2 two-factor models, etc.

artifact of the methodology used to generate the data.

Overall, the sample sizes needed to analyze dichotomous data are higher than those needed for continuous data as presented by Mundfrom, et al. (2005). For many models with high communalities, three or fewer factors, and high *p*:*m* ratios, sample sizes below 100 are likely to achieve good agreement. Conversely, sample sizes in the thousands are necessary to meet that criterion for most cases when all variables have low communalities or the factors are weakly determined.

Some relationships are apparent from Tables 1 and 2. For a given distribution, level of communality and number of factors, the necessary sample size tends to decrease sharply as the *p*:*m* ratio increases until some elbow after which changes in sample size are very small. This elbow tends to occur at *p*:*m* ratios between seven and ten. For a fixed *p*:*m* ratio, the minimum sample size tends to increase as the number of factors increases. These relations mimic those reported by Mundfrom, et al. (2005) for continuous data, but with more extreme patterns.

Among the three dichotomization conditions, the 50/50 distribution generally requires the lowest sample size. No generalizations are evident as to which of the 80/20 and mix conditions require a lower sample size. The disparity between continuous and binary conditions is smallest for the most well-defined factor structures, especially those with high *p*:*m* ratios. Differences among the binary distribution conditions tend to be small relative to their differences from the continuous data requirements.

Conclusion

One purpose of this study was to provide sample size recommendations to be used by researchers planning studies involving factor analysis of dichotomous data; these are provided in Tables 1, 2 and 3. Although the requirements for analyzing binary data are uniformly higher than those for continuous data across varied aspects of factor model design, they are still reasonable for well-defined factor models. A sample size of 100, which Gorsuch (1983) called the absolute minimum and Comrey and Lee (1992) labeled as poor, is enough to achieve a good level of agreement for models having one or two factors, as well as for three-factor models with at least 24 variables when communalities are high and variables have a symmetric distribution. When the *p*:*m* ratio is high, a sample size of 300 results in good agreement for many models in the wide communality condition and all three examined variable distribution conditions. This sample size is also enough to achieve excellent loading agreement for small models (one or two factors) when variables have high communalities.

The necessary sample size to achieve good agreement between sample and population loadings is grossly inflated for poorly-defined factor models. When communalities are all in the low range, sample sizes in the thousands are necessary for most of the examined conditions. The same is true for most models having four or more factors and *p*:*m* ratios of five or lower.

Another goal of this study was to investigate how dichotomization affects the necessary sample size for EFA. Cohen (1983) showed that when two continuous variables with a joint correlation of *r* are dichotomized at their means, the correlation between the resulting variables is attenuated to a value of .637*r*. One effect of the reduced correlations is that the communalities estimates are concordantly reduced. As described by Schiel and Shaw (1992), 36% of the information is lost when a perfectly reliable continuous variable is dichotomized at the mean. Hence, the communalities are deterministically reduced and additional error is present in the correlation estimates themselves.

MacCallum, et al. (1999) illustrated the role that sampling error has in the formula for the sample factor model. In the presence of sampling error the unique factors will neither have zero correlations with each other nor with the common factors. The terms that are affected by this error are weighted by the size of the unique factor loadings, which are inversely related to communalities.

In summary, dichotomization results in increased sampling error in correlation estimates and attenuated correlation coefficients, which in turn results in decreased communalities. The latter outcome produces larger unique variances which places more weight on the lack of fit

terms in the sample factor model. Thus, there is more sampling error and more weight placed on its detrimental effects.

Dichotomization has the greatest deleterious impact on necessary sample size when communalities are low, the ratio of variables to factors is low or the number of factors is high. The direct and interaction effects of communality follow directly from the previous argument. The other two characteristics affect the overdetermination of common factors. Although the variable-to-factor ratio is not the sole basis of overdetermination, it is an important aspect of it. Many authors have suggested the importance of having a high *p*:*m* ratio (Comrey & Lee, 1992; Tucker, Koopman, and Linn, 1969).

Mundfrom, et al. (2005) demonstrated that the *p*:*m* ratio both has a strong direct relationship with sample size for a fixed *m* as well as a moderating effect on the relationships between sample size, communality, and the number of factors. Moreover, the results of the present study show that the ratio also moderates the effects of dichotomization and variable distribution. At high *p*:*m* ratios, the sample size requirements between the 50/50, 80/20, and mix distributions are fairly similar and in some cases (high communalities, one or two factors) are not that discrepant from those for continuous data. On the contrary, when the ratio is low and the common factors have a low degree of overdetermination, then other changes to the factor model have dramatic consequences on the necessary sample size.

Unless extremely large samples are tenable, some general strategies are recommended when binary data will be factor analyzed. Using variables with high communalities substantially reduces sample size requirements. However, this aspect of the study may be the most difficult to control in practice, especially in survey development. A more manageable design aspect is the *p*:*m* ratio. Having at least eight variables per factor is advised, and a ratio of ten or more should be preferred. This practical step may ameliorate unexpected problems of skewed variables and occasional low communalities.

Results of this study provide direct guidelines to applied researchers who are selecting a sample size for research that will involve exploratory factor analysis of dichotomous data. It is also intended for these results to serve as conservative guidelines for research involving ordinal data. Although the use of dichotomous measures does necessitate larger samples, if many high-quality indicators are used to measure a small number of factors, then applied researchers can be confident that a small to moderate sample size will be adequate to produce a reliable factor solution.

References

Baños, J. H., & Franklin, L. M. (2002). Factor structure of the mini-mental state examination in adult psychiatric inpatients. *Psychological Assessment*, *14*(*4*), 397-400.

Bartholomew, D. J., & Knott, M. (1999). *Latent Variable Models and Factor Analysis*. London: Arnold.

Bock, R. & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*, 443-459.

Browne, M. (1968). A comparison of factor analytic techniques. *Psychometrika*, *33*, 267-334.

Cattell, R. (1978). *The Scientific Use Of Factor Analysis*. New York: Plenum.

Cohen, J. (1983). The cost of dichotomization. *Applied Psychological Measurement*, *7*(*3*), 249-253.

Comrey, A., & Lee, H. (1992). *A first course in factor analysis*. Hillsdale, NJ: Erlbaum.

Darlington, R. (1997). *Factor Analysis*. Retrieved June 2, 2008, from http:// www.psych.cornell.edu/darlington/factor.htm.

Everitt, B. (1975). Multivariate analysis: The need for data, and other problems. *British Journal of Psychiatry*, *126*, 237-240.

Gorsuch, R. L. (1983). *Factor Analysis* (*2nd Ed*.). Hillsdale, NJ: Erlbaum.

Gorsuch, R. L. (1997). Exploratory factor analysis: its role in item analysis. *Journal of Personality Assessment*, *68*(*3*), 532-560.

Hip, J., & Bollen, K. (2006). Model Fit in Structural Equation Models with Censored, Ordinal, and Dichotomous Variables: Testing Vanishing Tetrads. *Sociological Methodology*, *33*, 267-305.

Johnson, R., & Wichern, D. (2002). *Applied Multivariate Statistical Analysis* (*5th Ed.*). Upper Saddle River, NJ: Prentice Hall.

Kline, P. (1994). *An Easy Guide To Factor Analysis*. New York: Routledge.

Lee, S., & Song, X. (2003). Bayesian analysis of structural equation models with dichotomous variables. *Statistics in Medicine*, *22*(*19*), 3073-3088.

MacCallum, R., Widaman, K., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(*1*), 84-99.

Mulaik, S. (1989). Blurring the distinctions between component analysis and common factor analysis. *Multivariate Behavioral Research*, *25*(*1*), 53-59.

Mundfrom, D., Bradley, R., & Whiteside, L. (1993). A factor analytic study of the infant/toddler and early childhood versions of the HOME Inventory. *Educational and Psychological Measurement*, *53*, 479-489.

Mundfrom, D., Shaw, D., & Ke, T. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing*, *5*(*2*), 159-168.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. *Psychometrika*, *43*, 551-560.

Nunnally, J. (1978). *Psychometric Theory* (*2nd Ed.)*. New York : McGraw-Hill.

Parry, C., & McArdle, J. (1991). An applied comparison of methods for least-squares factor analysis of dichotomous variables. *Applied Psychological Measurement, 15*(*1*), 35-46.

Schiel, J., & Shaw, D. (1992). Information retention as a function of the number of intervals and the reliability of continuous variables. *Applied Measurement in Education, 5*(*3*), 213-223.

Schoenberg, R., & Arminger, G. (1989). Latent variable models of dichotomous data. *Multivariate Behavioral Research*, *18*(*1*), 164-182.

Tomás-Sábado, J., & Gómez-Benito, J. (2005). Construction and Validation of the Death Anxiety Inventory (DAI). *European Journal of Psychological Assessment*, *21*(*2*), 108-114.

Tucker, R., Koopman, R., & Linn, R. (1969). Evaluation of factor analytic research procedures by means of simulated correlation matrices. *Psychometrika*, *34*(*4*), 421-459.

Weng, L., & Cheng, C. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement, 65*(*5*), 697-716.