Multivariate Behavioral Research, 44:147–181, 2009

ISSN: 0027-3171 print/1532-7906 online
DOI: 10.1080/00273170902794206

Psychology Press
Taylor & Francis Group

Exploratory Factor Analysis With Small Sample Sizes

J. C. F. de Winter,* D. Dodou,* and P. A. Wieringa
Department of BioMechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, The Netherlands

Exploratory factor analysis (EFA) is generally regarded as a technique for large sample sizes (N), with $N = 50$ as a reasonable absolute minimum. This study offers a comprehensive overview of the conditions in which EFA can yield good quality results for N below 50. Simulations were carried out to estimate the minimum required N for different levels of loadings (λ), number of factors (f), and number of variables (p) and to examine the extent to which a small N solution can sustain the presence of small distortions such as interfactor correlations, model error, secondary loadings, unequal loadings, and unequal p/f. Factor recovery was assessed in terms of pattern congruence coefficients, factor score correlations, Heywood cases, and the gap size between eigenvalues. A subsampling study was also conducted on a psychological dataset of individuals who filled in a Big Five Inventory via the Internet. Results showed that when data are well conditioned (i.e., high λ, low f, high p), EFA can yield reliable results for N well below 50, even in the presence of small distortions. Such conditions may be uncommon but should certainly not be ruled out in behavioral research data.

Exploratory factor analysis (EFA) is one of the most widely used statistical methods in psychological research (Fabrigar, Wegener, MacCallum, & Strahan, 1999), prompted by the need to go beyond the individual items of tests and questionnaires to reveal the latent structure that underlies them. Factor analyses

*These authors contributed equally to this work.

Correspondence concerning this article should be addressed to Joost de Winter, Department of BioMechanical Engineering, Faculty of Mechanical, Maritime and Materials Engineering, Delft University of Technology, Mekelweg 2, 2628 CD Delft, The Netherlands. E-mail: j.c.f.dewinter@tudelft.nl

are generally performed with large sample sizes. A study of the literature easily shows that applying EFA to small sample sizes is treated with caution. Researchers are discouraged from using EFA when their sample size (*N*) is too small to conform to the norms presented in the state of the art in factor analysis. Many early recommendations focused on the importance of absolute sample size. Guilford (1954) recommended a minimum sample size of 200 for consistent factor recovery. Comrey (1973) suggested a range of minimum sample sizes, from 50 (very poor) to 1,000 (excellent) and advised researchers to obtain sample sizes larger than 500. Gorsuch (1974) characterized sample sizes above 200 as large and below 50 as small. Cattell (1978) proposed that 500 would be a good sample size to aim at, commenting that in the context of most problems, however, 250 or 200 could be acceptable. Other researchers focused on the number of cases per variable (*N*/*p*) and recommendations range from 3:1–6:1 (Cattell, 1978) to 20:1 (Hair, Anderson, Tatham, & Grablowsky, 1979), with the latter advising researchers to obtain the highest cases-per-variable ratio possible in order to minimize the chance of overfitting the data.

Later studies showed that those propositions were inconsistent (Arrindell & Van der Ende, 1985) and recommendations on absolute *N* and the *N/p* ratio have gradually been abandoned as misconceived (Jackson, 2001; MacCallum, Widaman, Zhang, & Hong, 1999). Meanwhile, a number of studies have pointed out that not only sample size but also high communalities (Acito & Anderson, 1980; Pennell, 1968) as well as a large number of variables per factor (*p/f*; Browne, 1968; Tucker, Koopman, & Linn, 1969) contribute positively to factor recovery. Recently, a steeply increasing number of simulation studies has investigated the determinants of reliable factor recovery and shown that minimum sample size is a function of several parameters. There are no absolute thresholds: minimum sample size varies depending on the level of communalities, loadings, number of variables per factor, and the number of factors (Gagné & Hancock, 2006; MacCallum, Widaman, Preacher, & Hong, 2001; MacCallum et al., 1999; Marsh, Hau, Balla, & Grayson, 1998; Velicer & Fava, 1998). A considerable part of the literature on sample size recommendations has been reviewed by Velicer and Fava and MacCallum et al. (1999).

MacCallum et al. (1999) developed a theoretical framework for the effects of sample size on factor recovery and provided a basis for the contention that there are no absolute thresholds for a minimum sample size. This framework is based on earlier theoretical analyses presented by MacCallum and Tucker (1991), subsequently extended by MacCallum et al. (2001). The framework indicates that factor recovery improves as (a) sample size increases, (b) communalities increase, and (c) *p/f* increases; the effect of *p/f* decreases as communalities increase, and it may also interact with the sample size. Although the simulations in MacCallum et al. (1999) and MacCallum et al. (2001) applied a minimum *N* of 60, their theoretical framework should be applicable to smaller sample sizes

as well. However, it remains undefined how small a sample size can be and still yield acceptable solutions.

Only a very limited number of studies on the role of sample size in factor analysis have investigated real or simulated samples sized smaller than 50, probably because this is considered a reasonable absolute minimum threshold (Velicer & Fava, 1998). A few earlier studies recognized that sample sizes of 30 (Geweke & Singleton, 1980, having tested sample sizes as small as 10) or 25 (Bearden, Sharma, & Teel, 1982) can be adequate but, as Anderson and Gerbing (1984) noted, the latter study was limited and its findings should not be generalized. In a Monte Carlo study on confirmatory factor analysis (CFA) with sample sizes ranging from 25 to 400, Boomsma (1982) characterized factor analyzing sample sizes smaller than 100 as "dangerous" and recommended using sample sizes larger than 200 for safe conclusions. A subsampling study of Costello and Osborne (2005) indicated that for a sample size as small as 26, only 10% of the samples recovered the correct factor structure, whereas 30% of the analyses failed to converge and 15% had Heywood cases. A study by Marsh and Hau (1999) specifically devoted to small sample sizes in CFA used a minimum of 50 and warned that reducing the sample size from 100 to 50 can dramatically increase the number of improper solutions. Sapnas and Zeller (2002) determined adequate sample sizes for principal component analysis and suggested that a sample size between 50 and 100 was adequate to evaluate psychometric properties of measures of social constructs. This study, however, has been criticized for methodological errors and for failing to explain under which conditions a small sample EFA may be feasible (Knapp & Sawilowsky, 2005). In a more recent work, Zeller (2006) concluded that a sample size between 10 and 50 was sufficient for two dimensions and 20 variables. A simulation study by Preacher and MacCallum (2002) on applying EFA in behavior genetics clearly showed that for communalities between .8 and .9 and two factors EFA can yield reliable solutions even for sample sizes as small as 10. Another recent Monte Carlo study by Mundfrom, Shaw, and Ke (2005) also showed that if communalities are high and the number of factors is small, factor analysis can be reliable for sample sizes well below 50. Finally, Gagné and Hancock (2006) found that a sample size of 25 yielded no incidences of nonconvergent or Heywood cases when loadings were as high as .8 and $p/f = 12$. The majority of these studies, however, did not investigate factor recovery when deviating from a simple structure, a situation most likely to be encountered in real data. An exception was the study by Preacher and MacCallum, but it included model error as the sole distortion.

This article aims to offer a comprehensive overview of the conditions in which EFA can yield good quality results for small sample sizes. A number of simulations were carried out to examine how the level of loadings and communalities, the number of factors, and the number of variables influence

factor recovery and whether a small sample solution can sustain the presence of distortions such as interfactor correlation, model error, secondary loadings, unequal loadings, and unequal *p/f*. Next, we conducted a subsampling study on a psychological dataset of individuals who filled in the 44-item Big Five Inventory. The dataset was part of the Gosling-Potter Internet Personality Project studying volunteers assessed via the Internet (Gosling, Vazire, Srivastava, & John, 2004; Srivastava, John, Gosling, & Potter, 2003).

SIMULATION STUDIES

The majority of previous Monte Carlo studies that examined the role of sample size in factor analysis estimated factor recovery for a predefined range of sample sizes. In contrast, this study estimated the minimum sample size that would yield a sample solution in good congruence with a population pattern (assuming a simple population pattern with common factors and equal loadings) for a combination of determinants (i.e., factor loadings, number of factors, and number of variables). The study subsequently introduced a number of small distortions to a population pattern to investigate factor recovery in a realistic context.

MINIMUM SAMPLE SIZE AS A FUNCTION OF DETERMINANTS

Method

The minimum sample size was estimated for population conditions with varying factor loadings ($\lambda = .2, .4, .6, .8, .9$), number of factors ($f = 1, 2, 3, 4, 8$), and number of variables ($p = 6, 12, 24, 48, 96$), except for $p < 2f$. The numerical ranges of the factors and variables were chosen to be representative for general factor analytical practice in psychological research (Henson & Roberts, 2006).

For each of the conditions under investigation, population solutions were defined to exhibit a simple pattern with equal loadings, as equal a number of loading variables per factor as possible, no secondary loadings, orthogonal factors, and no model error.[1] See Table 1 for an example.

The minimum sample size for each condition was estimated by means of a proportional controller. A Tucker's congruence coefficient (K) of .95 was considered the minimum threshold for "good agreement" (Lorenzo-Seva & Ten

[1]This article defines simple structure as a special case of Thurstonian simple structure, also called independent cluster structure or ideal simple structure.

TABLE 1
Example of Population Pattern
($\lambda = .8$, $f = 3$, $p = 24$)

.8	0	0
.8	0	0
.8	0	0
.8	0	0
.8	0	0
.8	0	0
.8	0	0
.8	0	0
0	.8	0
0	.8	0
0	.8	0
0	.8	0
0	.8	0
0	.8	0
0	.8	0
0	.8	0
0	0	.8
0	0	.8
0	0	.8
0	0	.8
0	0	.8
0	0	.8
0	0	.8
0	0	.8

Berge, 2006).[2] The controller tuned N so that K converged to the .95 threshold. More precisely, the following procedure was repeated 5,000 times:

1. Based on the population solution, a sample observation matrix (Nxp) was generated, using a method described by Hong (1999).
2. The Pearson correlation matrix of the sample observation matrix was submitted to principal axis factoring (maximum number of iterations: 9,999; iterative procedure continues until the maximum absolute difference of communalities was smaller than 10^{-3}) and oblique direct quartimin rotation (i.e., oblimin with gamma = 0; Bernaards & Jennrich, 2005) by extracting f factors. To prevent optimism bias by screening solutions, unscreened data were used, that is, solutions that yielded Heywood cases

[2]Lorenzo-Seva and Ten Berge (2006) suggest the .95 threshold for good agreement on the basis of judgments of factor similarity by factor analytic experts. Note that others have used a .92 threshold for good and .98 for excellent agreement (MacCallum et al., 2001).

(one or more variables with communalities equal to or higher than 1) were included in further analysis.

3. To recover the order and sign of the loadings, the *K*s for the factor combinations ($f \times f$) between the sample solution and the population solution were calculated. Next, the reordering procedure of the sample solution started with the highest absolute K of the $f \times f$ calculated *K*s and proceeded toward the lowest K until the sign and order of all factors were recovered.
4. K was calculated between each reordered sample solution and the population pattern.
5. A new N was calculated as $N(i+1) = N(i) - N(i) \cdot (K - .95)$, rounding away from $N(i)$. In other words, if $K > .95$, N was reduced, whereas, if $K < .95$, N was increased. Initial N, that is $N(1)$, was set at 1,000. A minimum N of 5 was set for controller stabilization. If N exceeded 10,000, the controlling phase was terminated and no estimated N was provided.

After the 5,000th repetition, the mean N of the last 4,500 repetitions was calculated, hereafter referred to as $N_{estimate}$. The first 500 iterations were omitted so that $N_{estimate}$ was based on the *N*s after the controller had stabilized.

The quality of $N_{estimate}$ was assessed during a verification phase. That is, for 5,000 new repetitions, median K, mean K, 5th percentile of K, the mean factor score correlation coefficient (*FSC*), and the proportion of sample solutions exhibiting one or more Heywood cases were calculated. The factor score correlation coefficient was inspired by the comparability coefficient described by Everett (1983). Bartlett factor scores based on the sample solution and Bartlett factor scores based on the population pattern were calculated. *FSC* was then defined as the correlation between the sample factor scores and the population factor scores, averaged over the f factors. Heywood variables were omitted when calculating the factor scores of the sample solution. Finally, Cohen's d effect size (*ES*) was calculated between the fth and $(f+1)$th eigenvalues of the unreduced correlation matrix as a descriptive measure of the size of their "gap".[3] An $ES = 4$ was considered a gap of adequate size; assuming two independent normally distributed eigenvalues with equal standard deviations and applying a threshold

[3]The *ES* index in this article was calculated from the eigenvalues of the unreduced correlation matrices (UCM, with 1s in the diagonal). It has been argued that it is more conceptually sensible to use the eigenvalues of the reduced correlation matrix (RCM, with communality estimates in the diagonal) when the goal is to identify the number of common factors (Fabrigar et al., 1999; Preacher & MacCallum, 2003). We have repeated the subsampling study with *ES* based on the RCM (with communality estimates based on squared multiple correlations). Results showed that the difference in *ES* based on the UCM and the *ES* based on the RCM was always smaller than 10% and that overall average *ES* was higher for the UCM as compared with the RCM.

in the middle (i.e., $ES = 2$ from both means) implies that the correct number of factors can be identified in 95.5% of the solutions.

Results

The results in Table 2 show that factor recovery can be reliable with sample sizes well below 50. In agreement with the theoretical framework of MacCallum et al. (1999), lower sample sizes were needed when the level of loadings (λ; therefore the communalities) was high, the number of factors (f) small, and the number of variables (p) high. For loadings higher than .8 and one factor, even sample sizes smaller than 10 were sufficient for factor recovery. The level of loadings was a very strong determinant. For example, when loadings were as high as .9, and even with a high number of factors ($f = 4$) and a limited number of variables ($p = 12$), a sample size of 12 sufficed.

A larger number of variables improved factor recovery, particularly when loadings were low. No practical objection for performing EFA was found in conditions where the number of variables exceeded the sample size. In fact, increasing the number of variables reduced the minimum N, also when $p > N$.

Table 2 shows that for constant mean K, ES was lowest when high λ was combined with low p and low N, indicating that researchers should be cautious when deciding on the number of factors, particularly under such circumstances. In most conditions, however, ES was greater than 4. The highest ES was found in patterns with low λ, high p, and high N. Increasing p was beneficial for ES, even when N was decreased.

Table 2 also reveals the different tendencies of the factor recovery indices. Although the mean/median K was kept constant at .95, the 5th percentile of K systematically increased with an increase of p, signifying a favorable distributional change of K. FSC consistently and strongly improved with an increase of p, profiting from the additional information provided by extra variables. The proportion of sample solutions exhibiting Heywood cases reduced with higher p, whereas it increased for higher λ. This phenomenon can be attributed to the fact that increased λ elevates the risk of communalities higher than 1, due to sampling error. Note that the presence of Heywood cases was not detrimental to the recovery of the population pattern per se, as high K and high FSC could still be obtained in the unscreened solutions.

A more detailed analysis was conducted to gain insight into the interactions between the determinants. Mean K, mean FSC, and ES were calculated for a wide range of f (between 1 and 8) and p (logarithmically spaced between 10 and 200), except for $p < 2f$, for six combinations of sample sizes (small: $N = 25$, medium: $N = 100$, and high: $N = 1{,}000$) and levels of loadings (low: $\lambda = .4$ and high: $\lambda = .9$). The results are shown in Figure 1. Increasing N was always beneficial. Also apparent is that f had a relatively strong influence,

TABLE 2
Estimated *N* for Satisfactory Factor Recovery for Different Factor Loadings (λ), Numbers of Factors (*f*), and Numbers of Variables (*p*). For Each Condition, the Median, Mean, and 5th Percentile (*P5*) of Tucker's Congruence Coefficient (*K*), the Mean Factor Score Correlation Coefficient (*FSC*), the Proportion of Sample Solutions Exhibiting One or More Heywood Cases, and Cohen's d Effect Size (*ES*) Between the *f*th and (*f* + 1)th Eigenvalues Are Shown

λ	*f*	*p*	$N_{estimate}$	*Median K*	*Mean K*	*P5 K*	*Mean FSC*	*Heywood Cases*	*ES*
.2	1	6	1,524	.961	.950	.879	.952	.000	6.23
		12	752	.955	.950	.902	.960	.000	6.36
		24	470	.953	.951	.919	.970	.000	7.53
		48	339	.952	.950	.927	.980	.000	8.98
		96	274	.951	.950	.933	.987	.000	10.43
	2	6	5,849	.958	.950	.883	.949	.000	6.78
		12	2,571	.954	.950	.916	.955	.000	6.91
		24	1,438	.952	.950	.927	.962	.000	8.38
		48	918	.950	.950	.934	.971	.000	10.61
		96	676	.950	.950	.939	.981	.000	13.47
	3	12	5,363	.953	.950	.914	.952	.000	6.88
		24	2,829	.951	.950	.931	.959	.000	8.33
		48	1,732	.950	.950	.937	.967	.000	11.16
		96	1,197	.950	.950	.942	.976	.000	14.96
	4	24	4,602	.950	.950	.932	.956	.000	7.94
		48	2,750	.950	.950	.939	.964	.000	11.24
		96	1,827	.950	.950	.942	.973	.000	15.57
	8	48	8,695	.950	.950	.942	.957	.000	10.18
		96	5,390	.950	.950	.945	.964	.000	15.75
.4	1	6	102	.963	.950	.871	.954	.004	5.31
		12	64	.955	.948	.890	.968	.000	5.16
		24	52	.953	.949	.905	.982	.000	5.66
		48	46	.952	.949	.915	.990	.000	6.05
		96	44	.952	.950	.919	.995	.000	6.46
	2	6	370	.960	.950	.874	.946	.015	6.61
		12	186	.954	.950	.911	.963	.000	6.16
		24	134	.951	.949	.924	.976	.000	7.02
		48	112	.951	.950	.931	.986	.000	8.21
		96	101	.950	.950	.936	.992	.000	9.26
	3	6	1,159	.953	.949	.888	.938	.001	8.84
		12	353	.954	.950	.916	.958	.001	6.27
		24	234	.951	.950	.929	.972	.000	7.48
		48	186	.951	.950	.936	.983	.000	9.28
		96	163	.950	.950	.939	.990	.000	10.80
	4	12	589	.954	.950	.913	.953	.006	6.40
		24	349	.951	.950	.932	.969	.000	7.28
		48	270	.950	.950	.938	.980	.000	9.63
		96	230	.950	.950	.941	.988	.000	11.94

(*continued*)

TABLE 2
(*Continued*)

λ	*f*	*p*	$N_{estimate}$	*Median K*	*Mean K*	*P5 K*	*Mean FSC*	*Heywood Cases*	*ES*
.4	8	24	977	.951	.950	.934	.958	.001	6.53
		48	678	.950	.950	.942	.972	.000	9.34
		96	541	.950	.950	.944	.982	.000	13.62
.6	1	6	18	.965	.940	.813	.946	.046	4.42
		12	15	.961	.943	.844	.969	.004	4.39
		24	13	.955	.940	.840	.982	.001	4.46
		48	12	.952	.938	.847	.991	.000	4.61
		96	12	.951	.940	.860	.995	.000	4.69
	2	6	59	.960	.950	.877	.945	.120	5.35
		12	39	.955	.948	.896	.968	.001	5.15
		24	34	.952	.948	.913	.983	.000	5.66
		48	31	.951	.948	.920	.991	.000	6.19
		96	30	.951	.948	.923	.995	.000	6.43
	3	6	208	.950	.949	.903	.913	.404	8.42
		12	67	.954	.949	.910	.964	.009	5.26
		24	55	.951	.949	.924	.981	.000	5.94
		48	50	.949	.948	.929	.989	.000	6.77
		96	49	.951	.950	.934	.994	.000	7.72
	4	12	99	.952	.949	.911	.956	.051	5.20
		24	78	.951	.950	.929	.978	.000	5.97
		48	71	.950	.949	.935	.988	.000	7.31
		96	68	.950	.950	.938	.993	.000	8.56
	8	24	179	.951	.950	.933	.967	.004	5.36
		48	156	.950	.950	.941	.983	.000	7.59
		96	146	.950	.950	.943	.990	.000	10.15
.8	1	6	6	.984	.935	.686	.955	.331	4.95
		12	6	.983	.948	.786	.975	.176	4.93
		24	6	.982	.954	.822	.987	.099	5.06
		48	6	.982	.957	.840	.994	.048	5.24
		96	6	.981	.959	.850	.997	.025	5.24
	2	6	12	.960	.941	.815	.948	.542	3.75
		12	11	.956	.940	.843	.972	.134	3.85
		24	10	.949	.937	.856	.983	.044	3.80
		48	10	.949	.940	.876	.990	.008	4.07
		96	10	.949	.941	.882	.993	.002	4.32
	3	6	21	.958	.949	.886	.901	.845	3.73
		12	17	.952	.942	.873	.969	.181	3.64
		24	17	.952	.947	.907	.985	.011	4.08
		48	17	.952	.949	.916	.992	.000	4.53
		96	17	.952	.949	.922	.995	.000	4.84
	4	12	24	.954	.948	.900	.967	.325	3.61
		24	23	.952	.948	.919	.984	.010	4.11

(*continued*)

TABLE 2
(*Continued*)

λ	*f*	*p*	$N_{estimate}$	*Median K*	*Mean K*	*P5 K*	*Mean FSC*	*Heywood Cases*	*ES*
.8	4	48	23	.951	.949	.927	.991	.000	4.72
		96	23	.950	.949	.929	.994	.000	5.15
	8	24	45	.950	.949	.927	.977	.105	3.71
		48	47	.951	.950	.938	.989	.000	5.10
		96	47	.950	.950	.939	.993	.000	6.31
.9	1	6	5	.996	.961	.837	.978	.442	7.19
		12	5	.996	.978	.895	.989	.338	7.07
		24	5	.996	.978	.901	.994	.247	7.00
		48	5	.996	.981	.914	.997	.183	7.11
		96	5	.995	.978	.905	.998	.135	7.22
	2	6	7	.974	.951	.816	.968	.771	3.37
		12	6	.958	.931	.748	.976	.704	3.07
		24	6	.955	.934	.806	.984	.567	3.17
		48	6	.954	.935	.814	.988	.457	3.19
		96	6	.954	.937	.835	.991	.344	3.24
	3	6	8	.954	.929	.769	.896	.938	2.56
		12	9	.955	.939	.838	.975	.660	2.83
		24	9	.952	.940	.867	.985	.350	2.91
		48	9	.950	.941	.878	.989	.149	3.05
		96	9	.949	.941	.886	.991	.064	3.17
	4	12	12	.956	.944	.861	.974	.772	2.72
		24	12	.951	.943	.887	.985	.277	2.92
		48	12	.949	.945	.904	.990	.061	3.14
		96	12	.948	.944	.908	.992	.016	3.32
	8	24	23	.953	.947	.919	.982	.532	2.84
		48	24	.950	.949	.930	.990	.017	3.50
		96	25	.951	.951	.936	.993	.000	4.11

Note. All solutions were based on 5,000 repetitions.

with mean *K* and mean *FSC* reducing when *f* increased. The effect of *p*, on the one hand, depended on λ: for low λ, increasing *p* resulted in higher mean *K* and *ES*, whereas, for high λ, increasing *p* had a much smaller positive effect. For *FSC*, on the other hand, a higher *p* was beneficial for both high and low λ. These findings agree with the theoretical framework presented by MacCallum et al. (1999), demonstrating that a high *p/f* is advantageous to factor recovery and that this effect diminishes with increasing λ. However, the present results also showed that *p/f* is not a comprehensive measure, as *p* and *f* have clearly distinct effects on factor recovery.

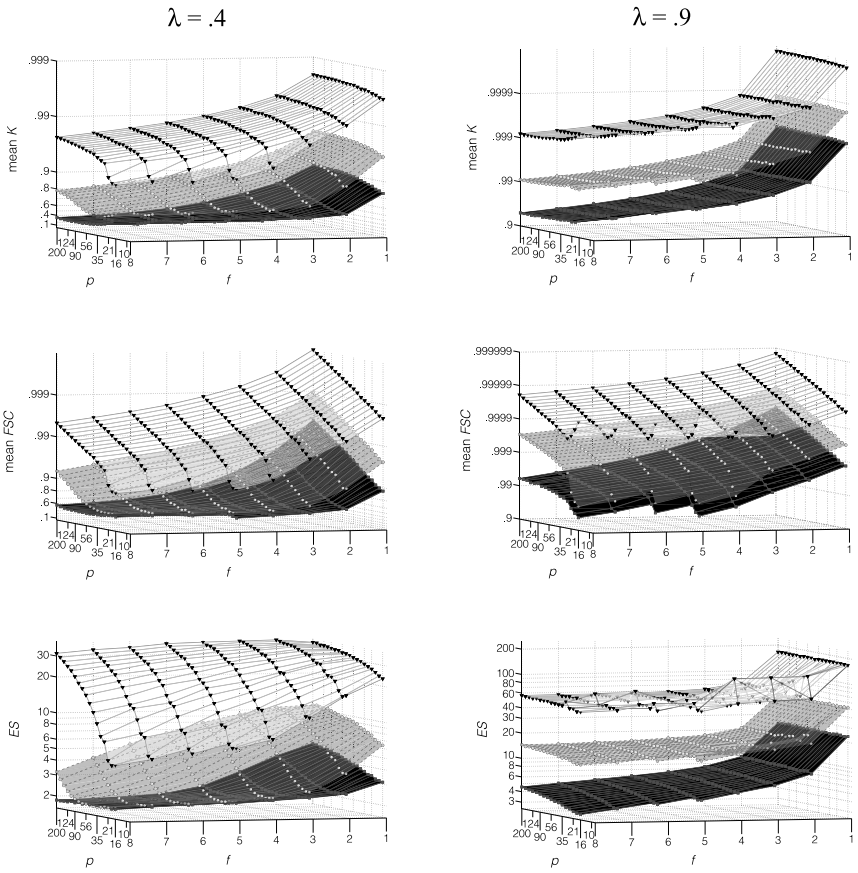


FIGURE 1 Main effects and interactions among the determinants of factor analysis. The plots show the factor recovery indices (mean Tucker's congruence coefficient (*K*), mean factor score correlation coefficient (*FSC*), and the Cohen's *d* effect size (*ES*) between the *f* th and ($f + 1$)th eigenvalues) as functions of a wide range of *f* and *p*, and two levels of loadings ($\lambda = .4$ and $\lambda = .9$). Three levels of sample size are shown in each plot, that is, $N = 25$ (black), 100 (grey), and 1,000 (white).

THE ROLE OF DISTORTIONS

Method

In reality, models rarely exhibit a perfectly simple structure. Moreover, models are imperfect, leaving a part of reality undepicted. For this reason, we systematically evaluated the role of various distortions (divided into 13 groups of four conditions each) in a baseline population solution with a small *N*

but large λ, low f, and large p ($N = 17, \lambda = .8, f = 3, p = 24$). The corresponding pattern solution is shown in Table 1. Iterative principal factor analysis was performed with oblimin rotation for all the investigated groups of distortions. Sufficient repetitions were performed for each condition so that the 95% confidence interval of the mean K was narrower than .001. The design of the simulation is summarized in Table 3. As an example, Table 4 shows the first population pattern of each investigated group.

Group 1: Interfactor correlation (ifc = .1, .3, .5, .7) for one pair of factors. This group examined the effect of various levels of *ifc* between two factors.

Group 2: Interfactor correlation (ifc = .1, .3, .5, .7) between all factors. Same as Group 1, but here all three combinations of factors were correlated, providing a more severe test case.

Group 3: Model error, altering the amount of variance. To investigate whether model error plays a role in factor recovery for small N, random model error was introduced for every repetition by means of 200 minor factors explaining (a) .05, (b) .1, (c) .15, and (d) .2 of the variance. The data generation parameter determining the distribution of successive minor factors was set at $\varepsilon = .1$.

Group 4: Model error, altering the distribution of the minor factors. In this group, the distribution of minor factors explaining .2 of the variance was altered by varying ε from .05 to .15, .25, and .35. Larger values of ε causes the contribution of the minor factors to be more skewed in favor of the earlier factors in the sequence (MacCallum & Tucker, 1991).

Group 5: Low loadings (.6) added. The simulation results in Table 2 on the factor analytic determinants showed that the addition of variables improved factor recovery. Adding low loading variables, however, inevitably decreases average communalities among all variables. Considering that both communalities and the number of variables are important determinants of factor recovery, the question is whether the addition of variables improves factor recovery, even when the added variables reduce average communalities. For this reason, the behavior of population patterns with a number (6, 12, 24, 96) of extra variables with low loadings (.6) was studied.[4]

[4]A loading of .6 was considered low for the sample size ($N = 17$) under investigation. This was based on the findings of the first part of the simulations (Table 2): for $\lambda = .6$, $f = 3$, $p = 24$, the required minimum N for good agreement ($K = .95$) was 55.

TABLE 3
Design of the Simulation Study Investigating the Role of Distortions

Group 1: Interfactor Correlation for One Pair of Factors	*Group 8: Unequal Loadings Between Factors*
Small (.1)	Small deviations (.85/.8/.75)
Medium (.3)	Medium deviations (.9/.8/.69)
Large (.5)	Large deviations (.95/.8/.61)
Very large (.7)	Very large deviations (.99/.8/.55)
Group 2: Interfactor Correlation Between All Factors	*Group 9: Unequal Loadings Within Factors*
Small (.1)	Small deviations (.85/.75)
Medium (.3)	Medium deviations (.9/.69)
Large (.5)	Large deviations (.95/.61)
Very large (.7)	Very large deviations (.99/.55)
Group 3: Model Error; Altering the Amount of Variance[a]	*Group 10: Secondary Loadings*
Small (.05)	2 variables; loadings .2/-.2
Medium (.10)	2 variables; loadings .4/-.4
Large (.15)	4 variables; loadings .2/-.2
Very large (.20)	4 variables; loadings .4/-.4
Group 4: Model Error; Altering the Distribution of Minor Factors[a]	*Group 11: Random Distortions of All Loadings*[a]
ε = small (.05)	Small (range .05)
ε = medium (.15)	Medium (range .10)
ε = large (.25)	Large (range .15)
ε = very large (.35)	Very large (range .20)
Group 5: Low Loadings (.6) Added	*Group 12: Unequal p/f (One Weak Factor)*[b]
Adding 6 variables	p/f = 8, 8, 6
Adding 12 variables	p/f = 8, 8, 4
Adding 24 variables	p/f = 8, 8, 3
Adding 96 variables	p/f = 8, 8, 2
Group 6: Low Loadings (.6) Replacing High Loadings	*Group 13: Unequal p/f (Two Weak Factors)*[b]
Replacing 3 variables	p/f = 8, 6, 6
Replacing 6 variables	p/f = 8, 4, 4
Replacing 12 variables	p/f = 8, 3, 3
Replacing 18 variables	p/f = 8, 2, 2
Group 7: Altering the Number of Variables	
p = 12	
p = 15	
p = 18	
p = 48	

[a] A different population pattern was produced for each repetition for all conditions of groups 3, 4, and 11. [b] The numbers refer to the variables per factor with a .8 loading.

TABLE 4
Population Patterns Used in the Simulations

Group 5			*Group 6*			*Group 7*		
.8	0	0	.8	0	0	.8	0	0
.8	0	0	.8	0	0	.8	0	0
.8	0	0	.8	0	0	.8	0	0
.8	0	0	.8	0	0	.8	0	0
.8	0	0	.8	0	0	0	.8	0
.8	0	0	.8	0	0	0	.8	0
.8	0	0	.8	0	0	0	.8	0
.8	0	0	.6	0	0	0	.8	0
0	.8	0	0	.8	0	0	0	.8
0	.8	0	0	.8	0	0	0	.8
0	.8	0	0	.8	0	0	0	.8
0	.8	0	0	.8	0	0	0	.8
0	.8	0	0	.8	0			
0	.8	0	0	.8	0			
0	.8	0	0	.8	0			
0	.8	0	0	.6	0			
0	0	.8	0	0	.8			
0	0	.8	0	0	.8			
0	0	.8	0	0	.8			
0	0	.8	0	0	.8			
0	0	.8	0	0	.8			
0	0	.8	0	0	.8			
0	0	.8	0	0	.8			
0	0	.8	0	0	.6			
.6	0	0						
.6	0	0						
0	.6	0						
0	.6	0						
0	0	.6						
0	0	.6						

Group 8			*Group 9*			*Group 10*		
.85	0	0	.85	0	0	.8	.2	0
.85	0	0	.747	0	0	.8	-.2	0
.85	0	0	.85	0	0	.8	0	0
.85	0	0	.747	0	0	.8	0	0
.85	0	0	.85	0	0	.8	0	0
.85	0	0	.747	0	0	.8	0	0
.85	0	0	.85	0	0	.8	0	0
.85	0	0	.747	0	0	.8	0	0

(*continued*)

TABLE 4
(*Continued*)

Group 8			*Group 9*			*Group 10*		
0	.8	0	0	.85	0	0	.8	0
0	.8	0	0	.747	0	0	.8	0
0	.8	0	0	.85	0	0	.8	0
0	.8	0	0	.747	0	0	.8	0
0	.8	0	0	.85	0	0	.8	0
0	.8	0	0	.747	0	0	.8	0
0	.8	0	0	.85	0	0	.8	0
0	.8	0	0	.747	0	0	.8	0
0	0	.747	0	0	.85	0	0	.8
0	0	.747	0	0	.747	0	0	.8
0	0	.747	0	0	.85	0	0	.8
0	0	.747	0	0	.747	0	0	.8
0	0	.747	0	0	.85	0	0	.8
0	0	.747	0	0	.747	0	0	.8
0	0	.747	0	0	.85	0	0	.8
0	0	.747	0	0	.747	0	0	.8
Group 11			*Group 12*			*Group 13*		
.791	-.019	-.011	.8	0	0	.8	0	0
.807	.002	.004	.8	0	0	.8	0	0
.805	.017	-.012	.8	0	0	.8	0	0
.778	-.013	.014	.8	0	0	.8	0	0
.819	-.024	-.018	.8	0	0	.8	0	0
.811	-.018	.014	.8	0	0	.8	0	0
.776	-.001	.001	.8	0	0	.8	0	0
.812	.019	.004	.8	0	0	.8	0	0
-.010	.818	.001	0	.8	0	0	.8	0
.013	.825	-.018	0	.8	0	0	.8	0
-.020	.813	-.009	0	.8	0	0	.8	0
-.014	.819	.018	0	.8	0	0	.8	0
-.001	.793	-.003	0	.8	0	0	.8	0
.019	.811	-.022	0	.8	0	0	.8	0
.004	.789	.019	0	.8	0	0	0	.8
-.004	.815	-.012	0	.8	0	0	0	.8
.016	-.015	.789	0	0	.8	0	0	.8
.010	-.004	.786	0	0	.8	0	0	.8
-.003	.006	.805	0	0	.8	0	0	.8
-.002	.003	.803	0	0	.8	0	0	.8
.019	-.013	.818	0	0	.8			
.007	.008	.786	0	0	.8			
-.018	-.018	.790						
-.014	.001	.804						

Note. The first condition described in the text is shown for each group.

Group 6: Low loadings (.6) replacing high loadings. A number (3, 6, 12, 18) of high loading variables were replaced with low loading (.6) variables. This was expected to cause a stronger distortion than Group 5 because the number of high loading variables was also reduced.

Group 7: Altering the number of variables. The number of variables was altered from 24 to 12, 15, 18, and 48 in order to investigate whether a discontinuity appears in factor recovery when factor analyzing samples in which the number of variables exceeds the sample size.

Group 8: Unequal loadings between factors. The level of the loadings among the three factors was varied in such a way that the average communalities of all variables remained equal to the baseline condition (i.e., .64). The following four combinations were investigated: (a) .85/.8/.75 ($= \sqrt{.8^2 - (.85^2 - .8^2)}$), (b) .9/.8/.69, (c) .95/.8/.61, and (d) .99/.8/.55.

Group 9: Unequal loadings within factors. The loadings within each of the three factors were alternated in such a way that the average communalities were equal to those in the baseline condition. The following four combinations of alternate nonzero loadings were investigated: (a) .85/.75 ($= \sqrt{.8^2 - (.85^2 - .8^2)}$), (b) .9/.69, (c) .95/.61, and (d) .99/.55.

Group 10: Secondary loadings. The effect of adding two or four secondary loadings of low (.2) as well as high (.4) level was examined. Alternating signs of the secondary loadings were used to prevent rotation toward a different solution.

Group 11: Random distortions of all loadings. In reality, population patterns are not homogeneous. Therefore, random distortions of all loadings were introduced. More precisely, four levels of uniform random loadings (ranges .05, .1, .15, and .2) were added to the baseline.

Group 12: Unequal p/f (one weak factor). Equal *p/f* rarely occurs in reality. Therefore the third factor was weakened by decreasing the number of variables that loaded on this factor.

Group 13: Unequal p/f (two weak factors). Factors 2 and 3 were weakened by decreasing the number of variables that loaded on these factors. This group tested the impact of weakening two out of three factors on factor recovery.

Results

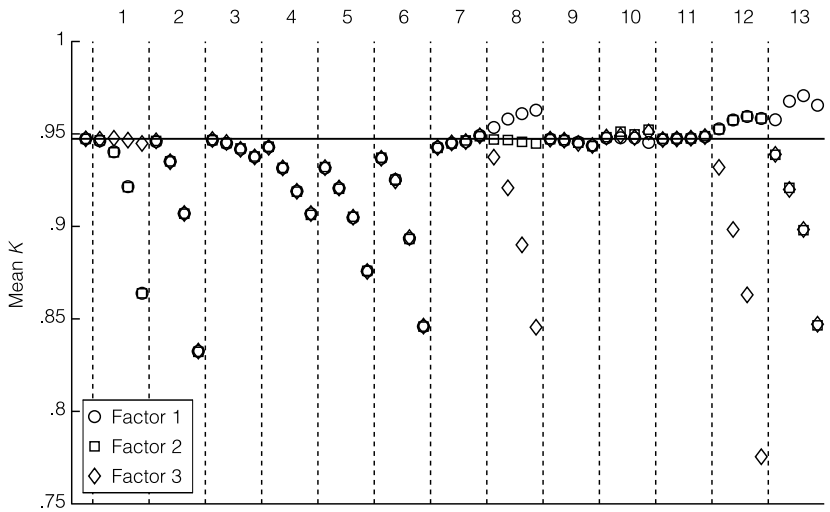
Figure 2 shows the mean *K*, mean *FSC*, the proportion of sample solutions exhibiting Heywood cases, and *ES* for each of the 13 groups.

Groups 1–2: Interfactor correlation. When one pair of factors was very strongly (.7) correlated or all factors were strongly (.5) correlated, mean *K* and mean *FSC* deteriorated considerably. Small interfactor correlation (up to .3) disturbed *K* and *FSC* to a far lesser extent. The proportion of sample solutions exhibiting Heywood cases increased slightly when all three factors were strongly correlated. Correlated factors negatively affected *ES* more than any other distortion did.

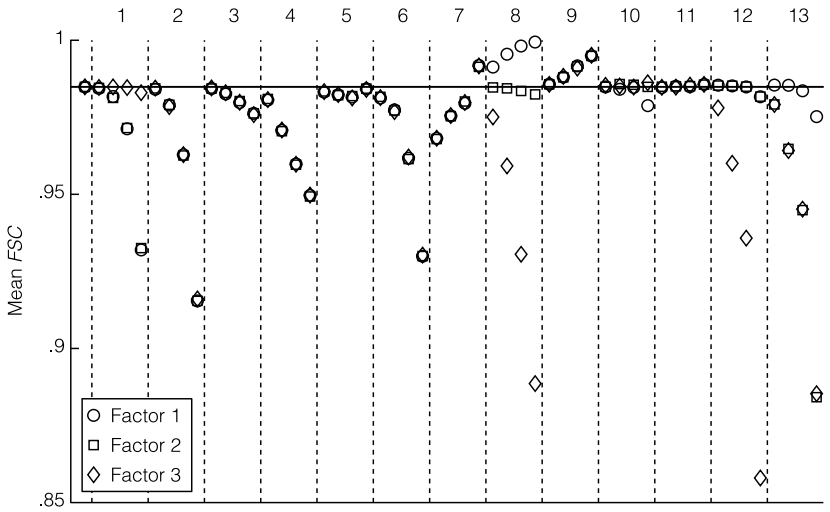
Groups 3–4: Model error. Model error slightly worsened factor recovery. This effect was seen in all four indices. Introducing a large model error (.2) across a more skewed distribution of minor factors ($\varepsilon = .25$ or .35) caused a relatively strong degradation of factor recovery as compared with the effect of a less skewed distribution.

Groups 5–6: Low loadings (.6) added or replacing high loadings. Adding low loading variables worsened mean *K*. Mean *FSC* slightly decreased for a small number of added low loadings but recovered for a larger (96) number of added low loadings. This can be explained by the fact that *K* takes into account all (low as well as high) loadings, whereas *FSC* is based on factor scores, obtaining the information from all the manifest variables. In other words, *FSC* benefited from (or at least stayed unaffected by) additional information. On the other hand, as an index of factor loadings similarity, *K* was more easily disturbed by the presence of variables of low quality. The deterioration of *K* was even more dramatic when low loading variables replaced high loading variables, whereas *FSC* degraded mainly when 18 out of the 24 variables had been replaced by low loadings. The proportion of sample solutions exhibiting Heywood cases reduced when low loading variables were added but increased when low loading variables replaced high loading variables. Appending low loadings influenced *ES* only slightly. However, this index degraded when low loadings replaced high loadings.

Group 7: Altering the number of variables. An increased number of variables slightly improved mean *K*, considerably improved mean *FSC*, and strongly suppressed the proportion of sample solutions exhibiting Heywood cases. In an additional test, driven more by theoretical interest rather than a realistic approach, factor recovery was estimated for 600 variables, a level at which mean *FSC* reached near unity (.997). In accordance with the first series of

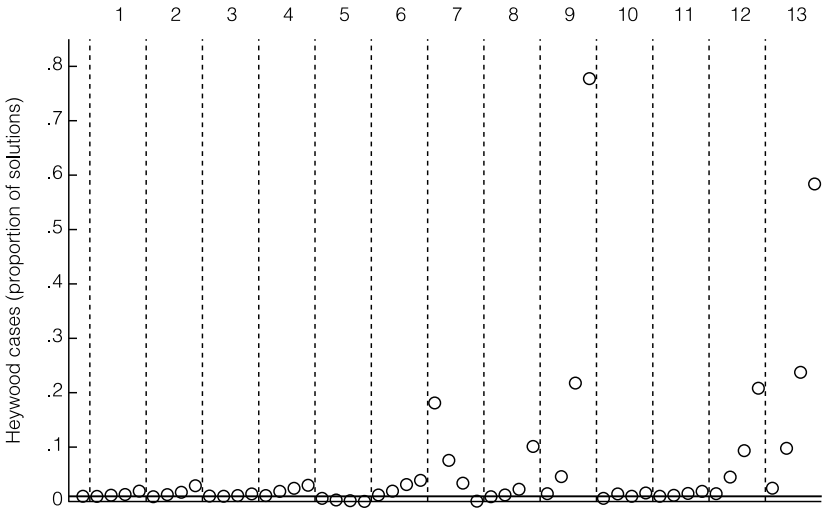


(a)

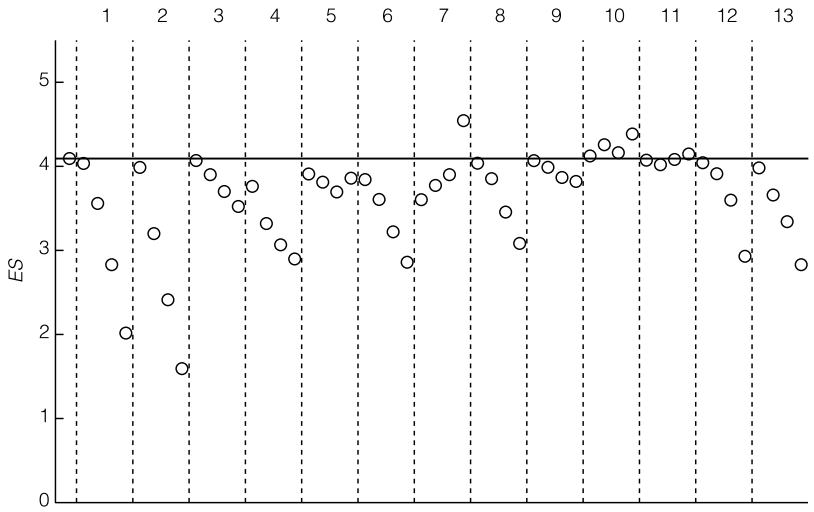


(b)

FIGURE 2 Factor recovery for the 13 investigated groups: (a) mean Tucker's congruence coefficient (*K*), (b) mean factor score correlation coefficient (*FSC*), (c) proportion of sample solutions exhibiting one or more Heywood cases, and (d) the Cohen's *d* effect size (*ES*) between the third and fourth eigenvalues. The horizontal line represents the average factor recovery for the baseline condition. (*continued*)



(c)



(d)

FIGURE 2 (*Continued*).

simulations, increasing the number of variables increased *ES*. The results of this group also showed that there was no discontinuity whatsoever with respect to factor recovery at the point where the number of variables exceeded the sample size.

Groups 8–9:. Unequal loadings between or within factors. For unequal loadings between factors (Group 8), the recovery of factors with high loadings improved (with the *FSC* of the first factor reaching .999 in the fourth condition), whereas the recovery of factors with low loadings deteriorated. Unequal loadings within factors (Group 9) strongly increased the likelihood of Heywood cases. However, this did not deteriorate factor recovery: mean *K* remained constant while mean *FSC* increased up to near unity (.995) in the fourth condition, which had a very large proportion of Heywood cases. *ES* decreased with unequal loadings between factors (Group 8) but was less sensitive to unequal loadings within factors (Group 9).

Group 10: Secondary loadings. Secondary loadings hardly influenced factor recovery. Only when the number and level of secondary loadings were the highest tested did mean *FSC* slightly decrease. Mean *K*, on the other hand, slightly increased when secondary loadings were high (.4). Moreover, secondary loadings were beneficial to *ES*.

Group 11: Random distortions of all loadings. Randomly distorted loadings hardly influenced factor recovery, signifying that the positive effect of the presence of high loadings compensated for the negative effect of random low loadings.

Groups 12–13: Unequal p/f (one or two weak factors). Low *p/f* had a negative effect on the recovery of the corresponding factor. The mean *FSC* of the worst condition ($p/f = 2$) was still higher than .85, so even for a very weak factor, factor recovery was not necessarily grossly wrong in a small *N* scenario. The recovery of the strong factors improved in terms of *K*. A low *p/f* increased the proportion of sample solutions exhibiting Heywood cases and weakened *ES*, diminishing the odds of correctly estimating the number of factors.

Group summary. The investigated baseline ($N = 17$, $\lambda = .8$, $f = 3$, $p = 24$) was noticeably robust against single small distortions. Each of the indices (mean *K*, mean *FSC,* Heywood cases, and *ES*) was sensitive to different distortions. The most serious degradation of factor recovery was caused by a highly unequal distribution of variables between factors (unequal *p/f*). In addition, *ES* was highly sensitive to interfactor correlations. Replacing high

with low loadings or having unequal loadings between factors also negatively influenced *ES*.

SUBSAMPLING STUDY

A subsampling study was carried out to investigate whether the findings of the simulation study are realistic and hold for actual data. An empirical dataset with a large sample size was used, acting as a population. The dataset consisted of 280,691 participants (mean age = 30.4, *SD* = 8.5 years, 54% women) who filled in the 44-item Big Five Inventory (Gosling et al., 2004; Srivastava et al., 2003). All selected participants indicated that they were filling in the inventory for the first time. Only participants who filled in the English version of the inventory, answering all items without giving identical answers to all 44 items, were included. Subsamples were drawn from the population sample, and factor recovery was assessed between the subsampling and the population solution.

Method

Factor recovery was estimated for a range of 25 subsample sizes spaced logarithmically between 10 and 200. For each subsample size, 10,000 random subsamples were drawn from the population and factor analyzed as in the simulation study. To investigate the role of the number of factors as a determinant of factor analytic performance, factor recovery was assessed when retaining from one up to five factors. Variables were selected so that each factor contained the 8 to 10 variables representing the corresponding personality trait. Table 5 summarizes the results of the five population patterns. Communalities were wide. Interfactor correlations were low, so these should hardly affect factor recovery, according to the simulations. Factor recovery was evaluated using mean *K*, mean *FSC*, the proportion of solutions exhibiting a Heywood case, and *ES*.

Results

Figure 3 shows the factor recovery results. For one extracted factor, a sample size around 13 and 17 was adequate for satisfactory *FSC* and *K* (= .95), whereas $f = 2$ required a sample size between 30 (for *FSC* = .95) and 50 (for *K* = .95). When retaining all factors of the Big Five ($f = 5$), a considerably larger sample size (80 to 140) was needed. For all numbers of factors, the proportion of solutions exhibiting a Heywood case was below .05 for sample sizes greater than 17. For one extracted factor, a sample size of 10 was sufficient for *ES* = 4. In contrast, when all five factors were retained, a larger sample size (140) was required to guarantee an adequate ES.

TABLE 5
Mean, Minimum, and Maximum of Primary Loadings, Secondary Loadings, Communalities, and Interfactor Correlations of the Population Solutions for the Investigated Number of Factors

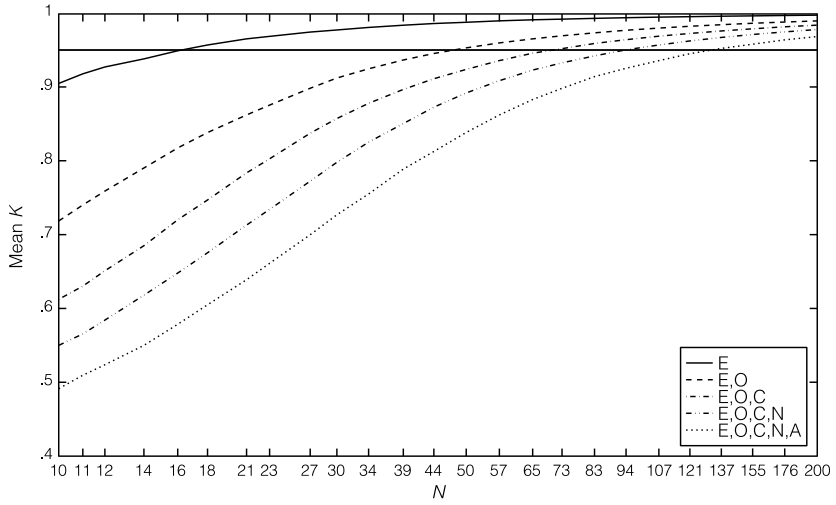
f	*p*	*Mean of Primary Loadings (Min–Max)*	*Mean of Secondary Loadings (Min–Max)*	*Mean Communalities (Min–Max)*	*Mean Interfactor Correlation (Min–Max)*
1 (E)	8	.64(.50–.77)	—	.42(.25–.59)	—
2 (E, O)	18	.58(.23–.79)	.10(.02–.28)	.37(.08–.60)	.17(.17–.17)
3 (E, O, C)	27	.58(.24–.80)	.11(.03–.27)	.38(.08–.62)	.14(.12–.16)
4 (E, O, C, N)	35	.59(.23–.81)	.13(.05–.28)	.40(.10–.63)	.14(.05–.24)
5 (E, O, C, N, A)	44	.57(.24–.80)	.15(.07–.27)	.39(.10–.62)	.12(.01–.20)

Note. All numbers are based on the absolute values of the pattern matrix and absolute values of the interfactor correlations. E = extraversion; O = openness; C = conscientiousness; N = neuroticism; A = agreeableness.

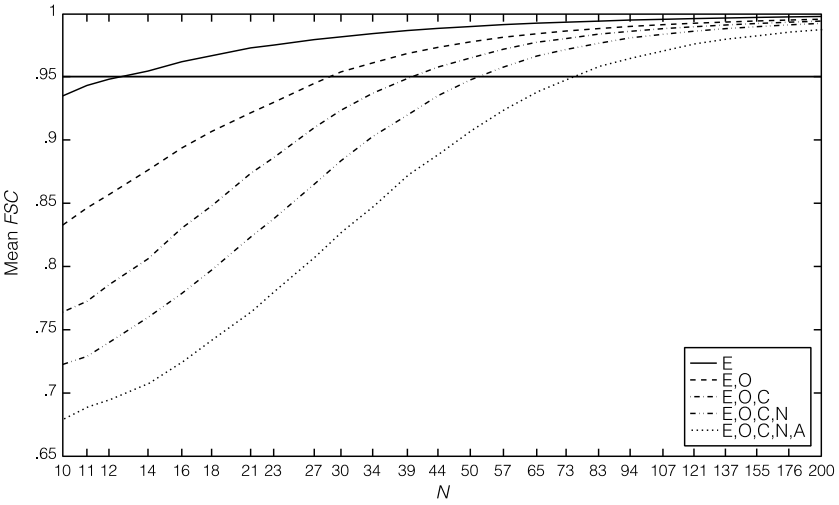
The subsampling study confirmed the findings of the simulation study with respect to the fact that a larger sample size is needed when extracting a larger number of factors. It should be noted that the subsampling study yielded moderately higher estimates of N compared with the simulations. For example, for (mean) $\lambda = .58$, $f = 3$, and $p = 27$, the subsampling study yielded an $N = 73$, whereas the respective value ($\lambda = .6$, $f = 3$, and $p = 24$) in the simulations (Table 2) was 55. This discrepancy can be attributed to the presence of model error as well as to the 5-point Likert scale data: to obtain reliable correlations between variables, Likert scale data require a larger sample size than continuous normally distributed data.

DISCUSSION AND RECOMMENDATIONS

The goal of this article is to offer a comprehensive overview of the conditions in which EFA can yield good quality results for small N. The simulations showed that, for the circumstances under which EFA is mostly applied (i.e., low to medium loadings, communalities, and a relatively large number of factors), a large sample size is required. However, when the data are well conditioned (i.e., high λ, low f, high p), EFA can yield reliable solutions for sample sizes well below 50. In some conditions, sample sizes even smaller than 10 (beyond the smallest sample size of previous simulation studies) were sufficient. For example, when $\lambda = .8$, $f = 1$, $p = 24$, and the structure was simple, $N = 6$ was adequate. A small sample solution ($N = 17$, $\lambda = .8$, $f = 3$, $p = 24$) was markedly robust against single small distortions. Weakly determined factors and

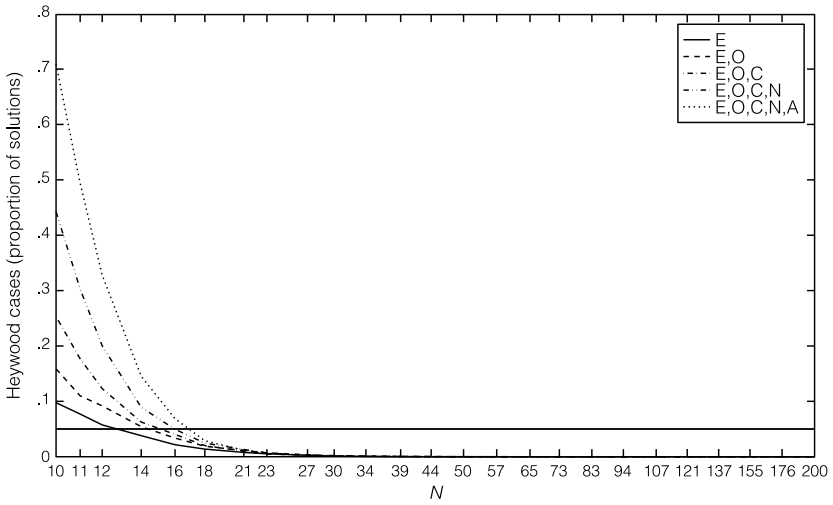


(a)

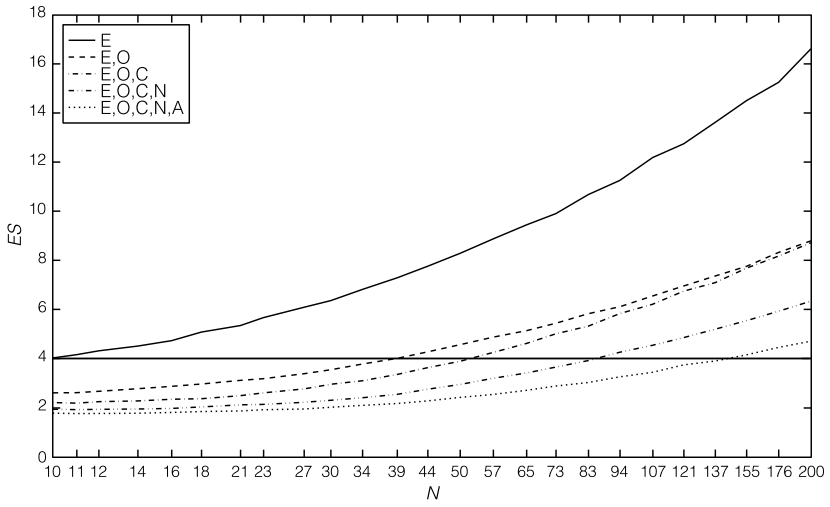


(b)

FIGURE 3 Subsampling factor recovery for the Big Five Inventory: (a) mean Tucker's congruence coefficient (*K*), (b) mean factor score correlation coefficient (*FSC*), (c) proportion of sample solutions exhibiting one or more Heywood cases, and (d) the Cohen's *d* effect size (*ES*) between the fth and $(f + 1)$th eigenvalues. The horizontal line represents a threshold for satisfactory factor recovery ($K = .95$, $FSC = .95$, Heywood cases $= .05$, $ES = 4$). Abbreviations: E = extraversion, O = openness, C = conscientiousness, N = neuroticism, A = agreeableness. (*continued*)



(c)



(d)

FIGURE 3 (*Continued*).

strong interfactor correlations negatively affected factor recovery, but even in the worst cases tested, factor recovery was still possible. The subsampling study confirmed the findings of the simulations with respect to the fact that a larger sample size is required when extracting a larger number of factors. For one extracted factor, a very small sample size (10–17) was adequate for satisfactory factor recovery.

An important issue when factor analyzing small samples is whether it is possible to correctly estimate the number of factors. The simulations showed that when the structure is simple, in most conditions, small sample sizes can guarantee an adequate $ES > 4$. However, when deviating from a simple structure, researchers should be extra cautious when deciding on the number of factors, particularly if these factors are correlated.

This article emphasizes that researchers should certainly not be encouraged to strive for small sample sizes. Large sample sizes are always beneficial and inevitably required when communalities are low. However, when factors are well defined or their number is limited, small sample size EFA can yield reliable solutions. Thus, a small sample size should not be the sole criterion for rejecting EFA. Inversely, if one prefers, subjecting a small sample to EFA can be worthwhile and may possibly reveal valuable latent patterns. Considering that models are useful unless they are grossly wrong (MacCallum, 2003) and a small sample size factor analytic model is not per definition grossly wrong, applying factor analysis in an exploratory phase is better than rejecting EFA a priori. Obviously, the reliability and theoretical soundness of the solutions should be very carefully assessed.

DEVIATIONS FROM A SIMPLE STRUCTURE

This study investigated factor recovery when deviating from a simple structure, a situation most likely to occur in real data but which had not previously been systematically investigated. Past studies usually focused on one kind of distortion and on sample sizes larger than 50.

A number of studies (Boomsma & Hoogland, 2001; Gagné & Hancock, 2006; Gerbing & Anderson, 1987; Marsh et al., 1998) introduced an *ifc* of .3 in their simulation studies without, however, investigating the effect of different levels of *ifc*. Anderson and Gerbing (1984) examined two levels of *ifc* but only with respect to improper solutions. It is surprising that the effect of *ifc* has not been exhaustively studied yet, considering that interfactor correlations are often present in psychological models. The current simulations investigated a range of *ifc* and revealed that a small *N* solution was able to sustain small interfactor correlations; factor recovery deteriorated considerably, however, when factors were strongly correlated.

Model error is usually considered as having no or only a small effect on factor recovery (MacCallum et al., 2001). The present results showed that when model error was small it was indeed of only little influence. A large model error, however, had a strong negative effect on factor recovery, particularly when its distribution was skewed in favor of the earlier factors in the sequence.

Gagné and Hancock (2006) found that when replacing higher with lower loadings (in an equal manner between factors), the number of improper solutions increased; appending loadings, on the other hand, was beneficial. These findings are in agreement with the present simulations. Additionally, the present study showed that, when replacing high with low loadings, all four investigated indices deteriorated. When appending low loadings, on the other hand, the indices exhibited various tendencies, indicating the importance of assessing factor recovery by means of more than one index.

Past studies showed that unequal loadings within factors may cause an increased number of improper solutions (Anderson & Gerbing, 1984) and a less rapid improvement of factor recovery as *N* increases (Velicer & Fava, 1998). The current simulations showed that in the presence of unequal loadings within factors *K* and *ES* remained unaffected, whereas *FSC* increased up to near unity despite an increased likelihood of solutions exhibiting a Heywood case. The robustness of the small *N* solution to unequal loadings with respect to *K* and *ES* and the improvement of *FSC* are of interest because such conditions resemble real data.

Beauducel and Wittmann (2005) investigated factor recovery in the presence of secondary loadings and found that secondary loadings did not influence absolute indices of model fit but negatively affected incremental indices. The present study showed that factor recovery was robust against secondary loadings and that *ES* can even improve. This is an important result because small secondary loadings are inevitably present in real data and researchers consistently use the presence of secondary loadings as a reason to disregard items during the process of scale construction. It should be noted, however, that scale developers may still prefer simple structures to ensure that individual items do not end up in multiple subscales and subscale correlations do not risk becoming inflated.

Beauducel (2001) and Briggs and MacCallum (2003) investigated patterns with weak factors but both studies focused more on comparing the performance of various factor analytic methods rather than examining factor recovery. Ximénez (2006) investigated the recovery of weak factors using CFA and found that the recovery of weak factors may be troublesome if their loadings are small and the factors orthogonal. The present simulations investigated the effects of weak factors by means of unequal loadings between factors as well as by means of unequal *p/f*. When loadings were unequal, weak factors did not inhibit the recovery of the strong factors. For unequal *p/f*, the recovery of factors with low *p/f* deteriorated considerably. On the other hand, the recovery of the

strong factors improved in terms of *K* (even if *p/f* of the strong factor was unchanged).

In summary, this study investigated the effect of a wide range of single small distortions on EFA performance and showed that although a small *N* solution was relatively robust against distortions, factor recovery deteriorated strongly when factors were correlated or weak.

THE EFFECTS OF *p*, *f*, *p/f*, and *p/N*

The simulations showed that an increased *p* improved factor recovery, raising *K*, *FSC,* and *ES* and reducing Heywood cases. An increased *p* was particularly beneficial for low λ patterns. A large number of references in the literature consider *p/f* a strong factor analytic determinant. The present simulations confirm that *p/f* is an important criterion; lowering *p/f* had a negative influence on factor recovery. When *p/f* was equal between factors, however, *p* and *f* had clearly distinct effects on the quality of the factor solutions (see Figure 1); therefore the *p/f* ratio should not be considered a comprehensive measure. For example, in a simple structure and for the same level of loadings (.8), two factors and 12 variables (i.e., $p/f = 6$) required a minimum sample size of 11, whereas with eight factors and 48 variables this minimum increased to 47. MacCallum et al. (2001) described a similar effect. They noticed that the effect of overdetermination on their empirical data was considerably weaker than in their Monte Carlo data. The difference was that in the empirical study the nature and number of factors were kept constant while the number of variables varied, whereas in the Monte Carlo study the number of variables was kept constant while the number of factors varied. The present study indicates that when *p/f* is equal, one should evaluate *p* and *f* separately instead of their ratio.

In some simulation conditions and in the subsampling study, the number of variables exceeded the sample size. Many factor analytic studies (e.g., Aleamoni, 1976), statistical packages, and factor analysis guidelines claim that the number of variables should never exceed the sample size. Contrary to this popular belief, Marsh and Hau (1999) reported no discontinuities in their simulation results when surpassing the $p = N$ barrier and suggested that there might be nothing special about such a barrier. The present simulations and the subsampling study concur with this view for all the investigated ranges of *p* and *N*. In fact, increasing the number of variables was beneficial, including when $p > N$. Moreover, in recent work, Robertson and Symons (2007) proved that $p > N$ is valid for maximum likelihood factor analysis. This method usually considers $p > N$ impossible because the covariance matrix turns nonpositive definite. Besides, as Bollen (2002) noted,

> Resolution of this indeterminacy is theoretically possible under certain conditions ... (a) when the sample size (N) goes to infinity, (b) when the number of observed variables goes to infinity, and (c) when the squared multiple correlation for the latent variable goes to one and the predictors are observed variables. (p. 616)

In other words, increasing the number of variables originates from the same striving for reducing factor indeterminacy as increasing the sample size. This is of importance for small sample sizes: when increasing N is not possible, one can attempt to append good quality variables, no matter if such a strategy may lead to a $p > N$ condition.

The simulations show that adding low loading variables considerably affected Tucker's congruence coefficient (*K)*. This may then imply that only variables expected to load highly on a factor should be considered. Such a recommendation is only partially true as it could lead to the pitfall of "bloated specific" factors because highly loading variables can be also highly redundant (Boyle, 1991). Such variables lead to factors that are internally consistent but have low validity because they mask the presence or power of other factors and contaminate the entire factor structure.[5] In fact, the selected variables should be such that they assure validity while being sufficiently diverse. The present simulations show that the *FSC* considerably improved when many variables were added, even when those variables had low factor loadings. In conclusion, we recommend increasing the number of variables as much as possible but only as long as this does not undermine the overall quality of the set.

INDICES FOR ASSESSING FACTOR RECOVERY

Indices used to evaluate the quality of factor solutions were *K*, *FSC*, Heywood cases, and *ES*. These indices exhibited varying tendencies and were sensitive to different determinants and distortions. The difference in the behavior of *K* and *FSC* is attributed to their inherent nature. As an index of factor loadings similarity, *K* is influenced both by high and low loadings. *FSC,* on the other hand, is an index of similarity of factor scores that are a weighted sum of the manifest variables. *FSC* monotonically increases with p because it benefits from added information in general. We conclude that *K* and *FSC* evaluate

[5]Cronbach's α was calculated for two conditions of the first simulation series (low loadings: $\lambda = .2$, $f = 2$, $p = 24$, $N = 1{,}438$ and high loadings: $\lambda = .9$, $f = 2$, $p = 24$, $N = 6$), Although factor recovery was identical in those two conditions (see Table 2), average Cronbach's a among variables loading on the factor was .332 for the low loadings and .968 for the high loadings. This demonstrates that high internal consistency is not necessary for good factor recovery. A more detailed discussion of this issue can be found in Boyle (1991).

different aspects of factor recovery and recommend using them complementarily.

A number of studies have discussed the effects of sample size and *p/f* on the proportion of sample solutions exhibiting a Heywood case (or improper solutions; e.g., Gerbing & Anderson, 1987; Marsh et al., 1998; Velicer & Fava, 1998). According to Velicer and Fava, Heywood cases are more likely to occur when the sample size is small, *p/f* is limited, and the loadings are low. Boomsma and Hoogland (2001) noticed that high factor loadings can also lead to Heywood cases. In the present study, the Heywood cases occurred indeed when loadings were high. However, Heywood cases were not detrimental to factor recovery, as high *K* and high *FSC* could still be obtained in the unscreened solutions. This agrees with MacCallum et al. (1999) and MacCallum et al. (2001), who carried out their simulations twice, once by screening out samples that yielded Heywood cases and again with unscreened data, and showed that there was virtually no difference in the results.

The present study not only included the Heywood cases but also used them as an index of factor recovery. Similarly, Briggs and MacCallum (2003) studied the behavior of Heywood cases when comparing different methods of factor analysis. Gagné and Hancock (2006) used nonconvergent and Heywood cases as a primary index of model quality. Based on the results of the present study, we recommend using the proportion of solutions exhibiting Heywood cases as an additional index because it offers valuable information about the effect of determinants and distortions.

An important question when factor analyzing small samples is whether the sample will consistently yield a correct decision as to the number of factors. Considering that none of the current methods for determining the number of factors is infallible (Fabrigar et al., 1999), *ES* was used to represent the size of the gap between the fth and $(f + 1)$th eigenvalues. When making the simplifying assumption of normally distributed independent eigenvalues with equal standard deviations, an $ES = 4$ corresponds to a maximum of 95.5% correct classifications. To illustrate the eigenvalue gap size in real data, Figure 4 shows the scree plot of the subsampling study for $N = 39$ and $f = 2$. Here, between the second and third eigenvalue, *ES* was 4.04. Applying the threshold at the optimal location (2.34) allowed for 96.5% correct estimations of $f = 2$. A caveat is in order: *ES* does not identify the most appropriate number of factors, nor does it tell where the "large gap" or the "elbow" can be found in the scree plot. Rather, *ES* is a between-samples measure.

DECIDING THE NUMBER OF FACTORS

Deciding the "correct" number of factors has been the subject of many studies.

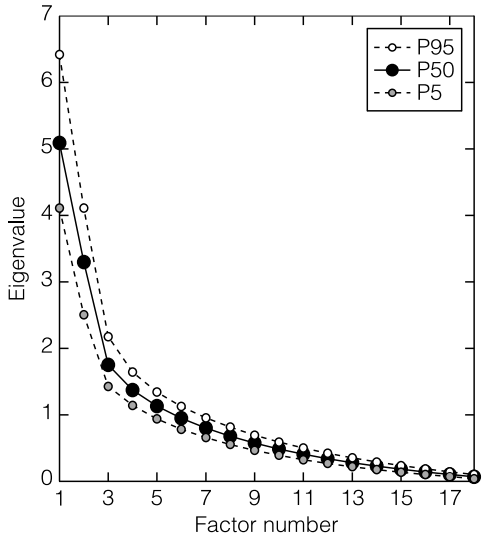


FIGURE 4 Scree plot of the subsampling study for $N = 39$ and $f = 2$ with the 5th, 50th, and 95th percentiles of the eigenvalues. The Cohen's d effect size (ES) between the second and third eigenvalue was 4.04. Applying the threshold at the optimal location (2.34) allowed for 96.5% correct estimations of $f = 2$. This figure is based on 100,000 repetitions.

As Bentler (2000) indicated,

> Inevitably, due to the variety of possible criteria and methods of evaluating this question, in any empirical application there will never be precise unanimity among all researchers. This does not worry me too much because various models always fit in degrees ... and perhaps there may not even be a "true" model. (p. 86)

In other words, it is better to think in terms of "most appropriate" than "correct" number of factors. Yet, even when the common factor model holds exactly in the population and $ES > 4$ (such as in most current simulations), automatically estimating the correct number of factors is a challenge. We made several attempts to estimate the correct number of factors in the first series of simulations by using Velicer's Minimum Average Partial (O'Connor, 2000), a Bayesian Information Criterion (Hansen, Larsen, & Kolenda, 2001), an automatic scree test (Zhu & Ghodsi, 2006), and parallel analysis (O'Connor, 2000; data not shown). Each of these methods was effective in many conditions, but none was successful in all conditions. A directly related topic is the effect of over- and underextraction (e.g., Fava & Velicer, 1992, 1996): although it has been reported that the effect of overextraction can be stronger when N is small and λ low (Lawrence & Hancock, 1999), one may question whether factor misspecification is a small

N problem per se or a matter of well- or ill-conditioned data. More research is needed on the strengths and weaknesses of procedures to determine the most appropriate number of factors.

STUDY LIMITATIONS

This study is not free of caveats or limitations. First, the simulation sample matrices were generated by a method described by Hong (1999), which produces normally distributed data and uses certain assumptions to generate model error (e.g., distribution of minor factors). Hong's method is a state-of-the-art procedure for introducing model error and interfactor correlations in a correlation matrix, based on the more commonly used Tucker-Koopman-Linn model (Tucker et al., 1969). Because of the normally distributed data, the simulations may have provided somewhat overoptimistic values for the minimum sample size compared with empirical data, as was also found in the subsampling study. Moreover, as Table 2 shows, the estimated minimum sample size would have been higher had factor recovery been assessed by using the 5th percentile of Tucker's congruence coefficient instead of its mean.

Second, although the simulated conditions corresponded to the ranges of determinants and distortions in psychological batteries, the conditions were of course far from exhaustive and might not be fully representative for all small sample conditions. For instance, one may conceive of structures that include combinations of distortions. Nonetheless, that factor recovery is possible in the presence of small distortions remains important for real applications.

Third, all correlation matrices were subjected to principal axis factoring and all loading matrices to oblique direct quartimin rotation. Different model fit procedures and rotations can have different effects on factor recovery. It is also possible that differently distorted matrices may have different favorable rotations. Those are issues that deserve further investigation.

Fourth, it should be noted that not just the factor recovery determines the quality of the factor analytic solution. As in any statistical analysis, the nature of both the sample and the variables involved remains among the most critical decisions (Fabrigar et al., 1999). A sample insufficiently representative of the population will distort the factor structure. Redundant variables can lead to bloated specific factors, obscuring the presence of more important factors. Irrelevant variables can also lead to spurious common factors (Fabrigar et al., 1999) and additional model error. Moreover, when the sample size is small, one should expect the standard error of loadings to be larger, which involves the risk of spurious high loadings.

Fifth, generalizing the findings to CFA should be done with great care. Although the communalities in CFA are usually higher due to variable selection

(MacCallum et al., 2001), particular caution should be taken with respect to misspecification and stronger sources of model error. The presence of model error may alter the minimum required sample size for CFA. However, as MacCallum et al. (2001) noted, one can expect to find similar tendencies and determinants for CFA.

Finally, one may doubt that real data can satisfy the constraints of high communalities and loadings or few factors. Moderate to weak communalities ranging between .4 and .7 (Costello & Osborne, 2005) or moderate to weak loadings ranging between .3 and .5 (Lingard & Rowlinson, 2006) are more common in behavioral or social data. The Big Five dataset of the subsampling study showed how indispensable a sufficiently large sample size is in such circumstances. However, cases involving high loadings do exist, for example, in neuroscience or psychosomatic research (e.g., Bailer, Witthöft, & Rist, 2006; Gaines, Shapiro, Alt, & Benedict, 2006; Yuasa et al., 1995; with loadings up to .90 or .95). One-factor structures are not uncommon in scientific literature either, such as in psychometrics, psychiatry, or epidemiology (e.g., general intelligence factor, self-concept, general distress factor, metabolic syndrome factor). Animal behavior and behavioral genetics (Preacher & MacCallum, 2002) as well as evolutionary psychology (Lee, 2007) often offer data with high communalities and few factors. Outside the field of behavioral sciences, physics and chemistry can feature data with high reliability. Paradoxically, when high quality data are likely to occur, researchers seem to think there is no need to resort to latent structures and prefer deductive reasoning and mathematical modeling instead. The question is whether dismissing EFA in those cases is not accompanied by a weaker representation of the reality (Haig, 2005) when neglecting the latent pattern of the data. EFA is indeterminate by nature, but so is the empirical world.

ACKNOWLEDGMENTS

The research of J. C. F. de Winter has been funded by the Dutch Ministry of Economic Affairs, under its Innovation Program on Man-Machine Interaction, IOP MMI. We are grateful to Samuel D. Gosling for providing the dataset of the Big Five Inventory for the subsampling study.

REFERENCES

Acito, F., & Anderson, R. D. (1980). A Monté Carlo comparison of factor analytic methods. *Journal of Marketing Research, 17,* 228–236.

Aleamoni, L. M. (1976). The relation of sample size to the number of variables in using factor analysis techniques. *Educational and Psychological Measurement, 36,* 879–883.

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.

Arrindell, W. A., & Van der Ende, J. (1985). An empirical test of the utility of the observations-to-variables ratio in factor and components analysis. *Applied Psychological Measurement, 9,* 165–178.

Bailer, J., Witthöft, M., & Rist, F. (2006). The chemical odor sensitivity scale: Reliability and validity of a screening instrument for idiopathic environmental intolerance. *Journal of Psychosomatic Research, 61,* 71–79.

Bearden, W. O., Sharma, S., & Teel, J. E. (1982). Sample size effects on chi square and other statistics used in evaluating causal models. *Journal of Marketing Research, 19,* 425–430.

Beauducel, A. (2001). On the generalizability of factors: The influence of changing contexts of variables on different methods of factor extraction. *Methods of Psychological Research Online, 6,* 69–96.

Beauducel, A., & Wittmann, W. W. (2005). Simulation study on fit indexes in CFA based on data with slightly distorted simple structure. *Structural Equation Modeling, 12,* 41–75.

Bentler, P. M. (2000). Rites, wrong, and gold in model testing. *Structural Equation Modeling, 7,* 82–91.

Bernaards, C. A., & Jennrich, R. I. (2005). Gradient projection algorithms and software for arbitrary rotation criteria in factor analysis. *Educational and Psychological Measurement, 65,* 676–696.

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. *Annual Review of Psychology, 53,* 605–634.

Boomsma, A. (1982). The robustness of LISREL against small sample sizes in factor analysis models. In K. G. Jöreskog & H. Wold (Eds.), *Systems under indirect observation: Causality, structure, prediction (part 1)* (pp. 149–173). Amsterdam: North-Holland.

Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling: Present and future. A festschrift in honor of Karl Jöreskog* (pp. 139–168). Lincolnwood, IL: Scientific Software International.

Boyle, G. J. (1991). Does item homogeneity indicate internal consistency or item redundancy in psychometric scales? *Personality and Individual Differences, 12,* 291–294.

Briggs, N. E., & MacCallum, R. C. (2003). Recovery of weak common factors by maximum likelihood and ordinary least squares estimation. *Multivariate Behavioral Research, 38,* 25–56.

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

Cattell, R. B. (1978). *The scientific use of factor analysis in behavioral and life sciences.* New York: Plenum.

Comrey, A. L. (1973). *A first course in factor analysis.* New York: Academic.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research & Evaluation, 10,* 1–9.

Everett, J. E. (1983). Factor comparability as a means of determining the number of factors and their rotation. *Multivariate Behavioral Research, 18,* 197–218.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods, 4,* 272–299.

Fava, J. L., & Velicer, W. F. (1992). The effects of overextraction on factor and component analysis. *Multivariate Behavioral Research, 27,* 387–415.

Fava, J. L., & Velicer, W. F. (1996). The effects of underextraction in factor and component analysis. *Educational and Psychological Measurement, 56,* 907–929.

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

Gaines, J. J., Shapiro, A., Alt, M., & Benedict, R. H. B. (2006). Semantic clustering indexes for the Hopkins Verbal Learning Test-Revised: Initial exploration in elder control and dementia groups. *Applied Neuropsychology, 13,* 213–222.

Gerbing, D. W., & Anderson, J. C. (1987). Improper solutions in the analysis of covariance structures: Their interpretability and a comparison of alternate respecifications. *Psychometrika, 52,* 99–111.

Geweke, J. F., & Singleton, K. J. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association, 75,* 133–137.

Gorsuch, R. L. (1974). *Factor analysis.* Philadelphia: Saunders.

Gosling, S. D., Vazire, S., Srivastava, S., & John, O. P. (2004). Should we trust web-based studies? A comparative analysis of six preconceptions about internet questionnaires. *American Psychologist, 59,* 93–104.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Haig, B. D. (2005). Exploratory factor analysis, theory generation, and scientific method. *Multivariate Behavioral Research, 40,* 303–329.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Grablowsky, B. J. (1979). *Multivariate data analysis.* Tulsa, OK: Pipe Books.

Hansen, L. K., Larsen, J., & Kolenda, T. (2001). Blind detection of independent dynamic components. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 5,* 3197–3200.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research: Common errors and some comment on improved practice. *Educational and Psychological Measurement, 66,* 393–416.

Hong, S. (1999). Generating correlation matrices with model error for simulation studies in factor analysis: A combination of the Tucker-Koopman-Linn model and Wijsman's algorithm. *Behavior Research Methods, Instruments, & Computers, 31,* 727–730.

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.

Knapp, T. R., & Sawilowsky, S. S. (2005). Letter to the editor. *Journal of Nursing Measurement, 12,* 95–96.

Lawrence, F. R., & Hancock, G. R. (1999). Conditions affecting integrity of a factor solution under varying degrees of overextraction. *Educational and Psychological Measurement, 59,* 549–579.

Lee, J. J. (2007). A *g* beyond Homo sapiens? Some hints and suggestions. *Intelligence, 35,* 253–265.

Lingard, H., & Rowlinson, S. (2006). Letter to the editor. *Construction Management and Economics, 24,* 1107–1109.

Lorenzo-Seva, U., & Ten Berge, J. M. F. (2006). Tucker's congruence coefficient as a meaningful index of factor similarity. *Methodology, 2,* 57–64.

MacCallum, R. C. (2003). 2001 presidential address: Working with imperfect models. *Multivariate Behavioral Research, 38,* 113–139.

MacCallum, R. C., & Tucker, L. R. (1991). Representing sources of error in the common factor model: Implications for theory and practice. *Psychological Bulletin, 109,* 502–511.

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

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

Marsh, H. W., & Hau, K. T. (1999). Confirmatory factor analysis: Strategies for small sample sizes. In R. H. Hoyle (Ed.), *Statistical issues for small sample research* (pp. 251–284). Thousand Oaks, CA: Sage.

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.

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

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers, 32,* 396–402.

Pennell, R. (1968). The influence of communality and N on the sampling distributions of factor loadings. *Psychometrika, 33,* 423–439.

Preacher, K. J., & MacCallum, R. C. (2002). Exploratory factor analysis in behavior genetics research: Factor recovery with small sample sizes. *Behavior Genetics, 32,* 153–161.

Preacher, K. J., & MacCallum, R. C. (2003). Repairing Tom Swift's electric factor analysis machine. *Understanding Statistics, 2,* 13–43.

Robertson, D., & Symons, J. (2007). Maximum likelihood factor analysis with rank-deficient sample covariance matrices. *Journal of Multivariate Analysis, 98,* 813–828.

Sapnas, K. G., & Zeller, R. A. (2002). Minimizing sample size when using exploratory factor analysis for measurement. *Journal of Nursing Measurement, 10,* 135–154.

Srivastava, S., John, O. P., Gosling, S. D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology, 84,* 1041–1053.

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

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

Ximénez, C. (2006). A Monte Carlo study of recovery of weak factor loadings in confirmatory factor analysis. *Structural Equation Modeling, 13,* 587–614.

Yuasa, S., Kurachi, M., Suzuki, M., Kadono, Y., Matsui, M., Saitoh, O., et al. (1995). Clinical symptoms and regional cerebral blood flow in schizophrenia. *European Archives of Psychiatry and Clinical Neuroscience, 246,* 7–12.

Zeller, R. A. (2006). *Statistical tools in applied research.* Retrieved July 10, 2008, from http://www.personal.kent.edu/∼rzeller/Ch.%2010.pdf

Zhu, M., & Ghodsi, A. (2006). Automatic dimensionality selection from the scree plot via the use of profile likelihood. *Computational Statistics and Data Analysis, 51,* 918–930.