Sample Size for Parallel Analysis and Not-So-Common Criteria for Dimensions in Factor Analysis: Modifying the Eigenvalue > 1 Kaiser Rule

Pornchanok Ruengvirayudh
Loyola Marymount University

Gordon P. Brooks
Ohio University

This study investigated the performance (i.e., consistency and correctness) of parallel analysis (PA) compared to other criteria for selecting the number of dimensions (i.e., components or factors) to extract, which is a step prior to exploratory factor analysis (EFA). Common criteria included PA using original/unreduced correlation matrices with 1s on the diagonal (PA_{COR}); PA using reduced correlation matrices with squared multiple correlations on the diagonal (PA_{SMC}); minimum average partial (MAP); and Kaiser's eigenvalue-greater-than-one rule or average root (K1). Not-So-Common criteria included indicator function (IND), imbedded error (IE), modified average roots (MARs), and broken stick (BS). These same criteria were studied from existing real test data (Experiment I) and generated data (Experiment II) to expand the generalizability of the results to real-world data and to explore more conditions with more variables, respectively. Ultimately, we provide guidelines on employing these criteria to yield the best results under studied conditions.

Factor analysis, specifically EFA and principal component analysis (PCA), is a process of data reduction aimed to reduce a large number of variables/items into a small number of dimensions (typically called components or factors) to make the data more manageable or interpretable, while maintaining as much information in the data as possible (Meyers, Gamst, & Guarino, 2013; Warner, 2013). EFA processes require applied researchers to specify the number of dimensions (if not using the K1 rule by default in SPSS, for example) to extract so that the EFA process can be completed. Selecting the correct number of dimensions is crucial to the validity of the results and decisions made.

There are many criteria used to determine the number of dimensions to feed into factor analysis. Probably, the most popular criterion is K1 rule. The logic of the K1 rule is that any retained dimensions should explain more variance than the average standardized variable, which is 1.0 (Jackson, 1991; Kaiser, 1960; Loehlin, 2011; Warne & Larsen, 2014). However, many researchers have criticized the cut-off value of 1.0 as being arbitrary and too inflexible. Consequently, K1 is not typically recommended (Jackson, 1991; van der Eijk & Rose, 2015).

Several studies investigated multiple numbers attempting to identify the appropriate cut-off value. Jolliffe (1972) recommended a modified cut-off value, or modified average root, of 70%. In the pilot testing process of this present study (Experiment I), using 70% of the average root as recommended yielded too many components. Therefore, other modified cut-off values (i.e., 1.20 and 1.40 in Experiment I and more cut-off values in Experiment II) were also investigated to compare with the average root and 70% of the average root. The cut-off value of 1.4 was studied for unidimensionality by Smith and Miao (1994). From their simulations, they found that the eigenvalue of the second component (the first component is pertinent to Rasch dimension), which is the largest component at the noise level, was never higher than 1.40. Therefore, they recommended that 1.4 be used as a cut-off value to determine whether a second factor exists (Smith & Miao, 1994). Chou and Wang (2010) studied a cut-off value of 1.5 from PCA on standardized residuals is performed to check for determining unidimensionality in item response theory (IRT), especially in Rasch models. Ultimately, however, Chou and Wang recommended against such a rule. As multiple modified cut-off values were used in this current study, this criterion is called modified average roots and abbreviated as MARs.

Versions of Horn's (1965) parallel analysis have been studied by many researchers and have been found to be useful in many conditions (e.g., Buja & Eyuboglu, 1992; Hubbard & Allen, 1987; Humphreys & Ilgen, 1969). PA could be viewed as an adaptation (sample-based) of the K1 rule (population-based) modified by taking sampling variability into account in order to improve the performance of the K1 rule (Timmerman & Lorenzo-Seva, 2011; Zwick & Velicer, 1986). PA is based on the comparison of eigenvalues of actual data to those of randomly generated data. A dimension is retained if the observed eigenvalue is greater than the random eigenvalue in the same ordinal position (Timmerman & Lorenzo-Seva, 2011; Zwick & Velicer, 1986). Because PCA uses original/unreduced correlation matrices (COR)

but EFA uses reduced correlation matrices (SMC), scholars debate on which type of input correlation matrices should be used for PA when used with EFA (e.g., Hill & Ray, 2014). For the number of randomly generated data sets, Horn suggested 50 data sets due to the limitations of computational power at that time (Glorfeld, 1995). After that, many scholars used at least 100 random data sets (e.g., Garrido, Abad, & Ponsoda, 2013; Green, Levy, Thompson, Lu, & Lo, 2012). Only a few scholars used a higher number than 100 (e.g., 1,000 in Debelak & Tran, 2013; 1,000 in Weng & Cheng, 2005; and 5,000 in Glorfeld, 1995). For threshold, Horn suggested the mean or the average of eigenvalues of random data although it tends to yield too many dimensions (Buja & Eyuboglu, 1992; Glorfeld, 1995). Some scholars used 50th percentile (i.e., median) that is usually slightly different from the mean (Debelak & Tran, 2013). To avoid overfactoring, a more stringent threshold (e.g., 95^{th}, 99^{th} percentile) is typically recommended to yield more accurate results (Buja & Eyuboglu, 1992; Cota, Longman, Holden, Fekken, & Xinaris 1993; Glorfeld, 1995).

Multiple average partial is based on the use of successive partial correlation matrices (Velicer, 1976). MAP starts from a complete PCA, then the first component is partialed out of the correlations among variables, and the average squared partial correlation is computed. This process is conducted for all possible partial correlation matrices and then the estimated number of components equals the step number yielding the smallest average squared partial correlation coefficient (O'Connor, 2000).

The imbedded error is the error that remains in the reproduced data matrix in exploratory factor analysis (Malinowski, 1977b). The number of dimensions is equal to the ordinal position of the variable that yields the minimum IE value (Jackson, 1991).

The indicator function is a function of the ordinal position of the variable, total number of variables, eigenvalues, and sample sizes, which are the same variables used in IE but calculated differently (Malinowski, 1977a). The number of dimensions is equal to the ordinal position of the variable that yields the minimum IND value (Jackson, 1991). Malinowski (1977a) recommended IND over IE, especially as sample sizes increase.

BS is one variant of PA as a "quick-and-dirty version of Horn's technique" (Jackson, 1991, p. 47). The name broken stick comes from the analogy of a stick that is broken at random into p segments (like, perhaps, a scree plot). Then, "the expected length of the kth longest segment" is calculated by a formula (Jolliffe, 1986) and a dimension will be retained if its proportion explained is larger than the computed chance value corresponding to each variable.

Methods and Data Source

This study used two sources of data: existing real test data (in Experiment I) and generated data (in Experiment II). In Experiment I, we used existing real test data (i.e., applied Monte Carlo study), which is advantageous in that the results are more applicable to real-world data, and practitioners can make use of provided guidelines in their applied research. In Experiment II, which was the extension of Experiment I, using generated data (i.e., true Monte Carlo study) is advantageous in that more variables could be varied, thus resulting in more studied conditions.

Experiment I: Methods and Data Source

In Experiment I, existing real test data were two (i.e., Mathematics as well as Career and Technology) out of eight subjects of the Ordinary National Educational Test (O-NET) data. The data were national examination scores of twelfth-grade students from Thailand in 2013, with permission granted from the National Institute of Educational Testing Service (Public Organization), Thailand or NIETS and IRB approval (Ruengvirayudh, 2018). Based on preliminary analyses of item analysis, exploratory factor analysis, and confirmatory factor analysis, those two subjects (i.e., Math and Tech) were manipulated into six factor structures based on the number of items (i.e., Math9, Math21, Math40, Tech15, Tech23, and Tech40; the number following the name of the subject refers to the number of items in that subject), the number of factors (i.e., one, two, and completely unknown), and the loading strength (i.e., moderately strong, acceptable, and as it was for 40 items), respectively (see Table 1).

Test items used in the study were binary (i.e., scored either right or wrong), so this study used only binary items. The data (approximately 400,000 cases) were treated as the population level because a Monte Carlo simulation process needs a population to repeatedly sample from. The R scripts in the R program were written to repeatedly sample from the data. Sample sizes were 100, 200, 300, 400, 500, 600, 700, 800,

Table 1. Summary of Factor Loading Criteria for Six Reduced Factor Structures

	Math			Tech		
	Math9	**Math21**	**Math40**	**Tech15**	**Tech23**	**Tech40**
Factor structure	One-factor	One-factor	Completely unknown	Two-factor (10 and 5 items)	Two-factor (12 and 11 items)	Completely unknown
Loading strength	Moderately strong	Acceptable	-	Moderately strong	Acceptable	-
Loading criteria	≥ .40	≥ .30	As it was	1L ≥ .40 & 2L < .20	1L ≥ .40 & 2L < .30 1L ≥ .30 & 2L < .20	As it was

Note. For the loading criteria, all loadings were at least .40 in Math9 and at least .30 in Math21. In Tech15, moderately strong loading strength refers to loading of .40 or higher in the first factor and loading less than .20 in the second factor. In Tech23, acceptable loading strength refers to loading of .40 or higher in the first factor and loading less than .30 in the second factor, or loading of .30 or higher in the first factor and loading less than .20 in the second factor. Math40 and Tech40 were a 40-item factor structure with loadings from the original data.

900, and 1,000 to investigate the results with a wider range and larger sample sizes. The increment of 100 was used to clearly see the change and diminishing returns in the results as a trend. The number of replications was 1,500 in the final study.

A computer program was written in R, the statistical programming language. The core of the program was tested and the output was verified to the extent possible. For PA and MAP, we tested and used the R Psych package functions. We programmed and tested K1, IND, IE, BS, and MARs by ourselves. Sample size and factor structure elaborated above were varied for all criteria. For PA, three more variables were varied: threshold (mean, 75^{th}, 90^{th}, and 95^{th} percentiles); type of input correlation matrices (original/unreduced correlation matrices with 1s on the diagonal or COR and reduced correlation matrices with squared multiple correlations on the diagonal or SMC); and number of randomly generated data sets (100, 300, 500, 700, 900, and 1000). Based on the literature review, the number of randomly generated data sets has not been found studied. For MARs, the cut-off values included 0.7, 1 (i.e., K1), 1.2, and 1.4. The only dependent variable of this study was the number of dimensions (i.e., components or factors).

Results of Experiment I

The data analysis included the mean of 1,500 samples as well as comparison statistics for precision (Standard Error (SE), Root Mean Squared Error (RMSE), Mean Absolute Value (MAE), and Relative Efficiency (RE)) and accuracy (percentage of correctness and bias). The data were analyzed step by step to select a few methods yielding the most precise results. Tables 2 and 3 show the mean and SE of the number of dimensions resulting from the methods in the final phase across 10 sample sizes from the Math9 data. The other factor structures were analyzed in the same manner.

As the true number of dimensions is unknown due to the use of existing real test data, SE, the average distance of a set of data from its mean, is the best statistic for Experiment I to inform the consistency or precision of the studied criteria because it is the only statistic that is not calculated based on the number of dimensions in the population level. Thus, SEs were more focused than the other statistics to compare the consistency of these methods.

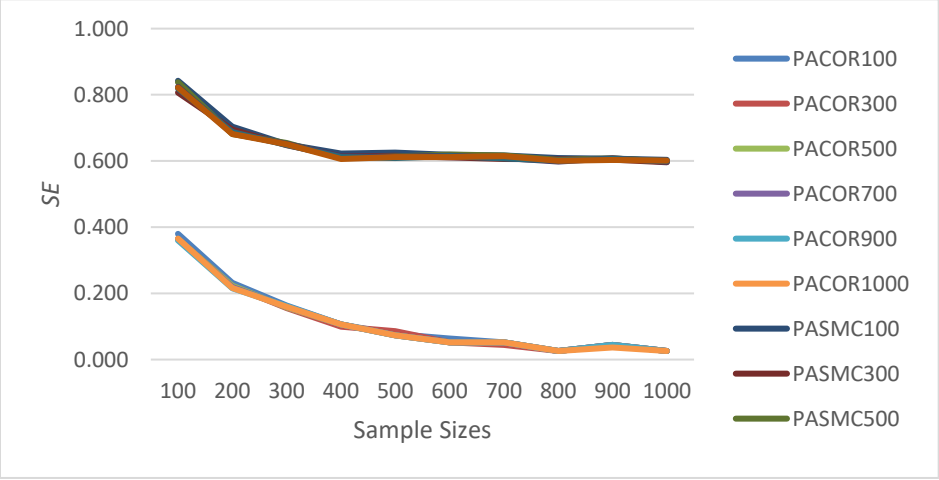
In the complete results, there were three groups of the criteria having more than one method: PA_{COR}, PA_{SMC}, and K1 and MARs. The best methods (with the lowest SE meaning that method produced the most precise results) of each group were: PA_{COR95}, PA_{SMC95}, and $MAR_{1.4}$. These criteria were chosen for further analysis below. Overall, PA_{COR} performed better than PA_{SMC}. Interestingly and unexpectedly, the differences in SE when varying the number of randomly generated data sets were minimal. Similar trends were found in the results from both PA_{COR} and PA_{SMC} although *SE* from PA_{COR} was consistently lower than

Table 2. Mean of the Number of Dimensions Resulting from Final Methods across 10 Sample Sizes from the Math9 Data

Criteria		*n*									
No.	**Methods**	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
1	PA_{CORMU}	1.421	1.161	1.070	1.051	1.021	1.019	1.011	1.013	1.002	1.003
2	PA_{COR75}	1.279	1.106	1.039	1.028	1.014	1.009	1.007	1.009	1.001	1.002
3	PA_{COR90}	1.185	1.067	1.023	1.014	1.009	1.004	1.005	1.004	1.001	1.001
4	PA_{COR95}	1.144	1.051	1.016	1.010	1.007	1.003	1.003	1.003	1.001	1.001
5	PA_{COR100}	1.136	1.055	1.026	1.011	1.006	1.004	1.003	1.001	1.002	1.001
6	PA_{COR300}	1.125	1.052	1.023	1.010	1.007	1.003	1.002	1.001	1.002	1.001
7	PA_{COR500}	1.125	1.053	1.024	1.011	1.005	1.003	1.003	1.001	1.002	1.001
8	PA_{COR700}	1.127	1.051	1.025	1.011	1.005	1.003	1.003	1.001	1.002	1.001
9	PA_{COR900}	1.125	1.049	1.025	1.011	1.005	1.003	1.003	1.001	1.002	1.001
10	$PA_{COR1000}$	1.125	1.049	1.025	1.011	1.005	1.003	1.003	1.001	1.001	1.001
11	PA_{SMCMU}	3.039	2.541	2.337	2.202	2.122	2.065	2.113	2.049	2.037	2.061
12	PA_{SMC75}	2.401	2.055	1.986	1.853	1.824	1.801	1.851	1.805	1.837	1.836
13	PA_{SMC90}	1.955	1.707	1.683	1.603	1.575	1.603	1.647	1.620	1.646	1.641
14	PA_{SMC95}	1.762	1.544	1.527	1.482	1.476	1.507	1.538	1.519	1.544	1.552
15	PA_{SMC100}	1.673	1.545	1.487	1.467	1.491	1.523	1.490	1.517	1.557	1.586
16	PA_{SMC300}	1.636	1.521	1.487	1.460	1.481	1.515	1.471	1.508	1.551	1.575
17	PA_{SMC500}	1.652	1.516	1.483	1.451	1.473	1.516	1.478	1.509	1.548	1.571
18	PA_{SMC700}	1.640	1.521	1.479	1.451	1.479	1.513	1.471	1.498	1.550	1.573
19	PA_{SMC900}	1.640	1.515	1.479	1.454	1.467	1.514	1.471	1.501	1.547	1.569
20	$PA_{SMC1000}$	1.637	1.520	1.479	1.450	1.470	1.508	1.476	1.507	1.547	1.572
21	MAP	1.007	1.001	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
22	BS	0.601	0.630	0.609	0.621	0.637	0.660	0.660	0.663	0.691	0.687
23	$MAR_{0.7}$	5.194	5.517	5.707	5.841	5.921	5.966	6.063	6.129	6.143	6.195
24	Average Root (K1)	2.785	2.247	1.945	1.725	1.531	1.453	1.361	1.273	1.219	1.191
25	$MAR_{1.2}$	1.729	1.189	1.041	1.013	1.003	1.001	1.000	1.001	1.000	1.000
26	$MAR_{1.4}$	1.157	1.005	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
27	IE	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
28	IND	1.095	1.002	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

Note. Only the 95th percentile was used as a threshold when varying the number of randomly generated data sets in PA_{COR} and PA_{SMC}.

Figure 1. SE of PA_{COR} and PA_{SMC} by the Number of Randomly Generated Data Sets from the Math9 Data



Note. COR refers to PA using original/unreduced correlation matrices. SMC refers to PA using reduced correlation matrices. The number following the type of input correlation matrices refers to the number of randomly generated data sets.

SE from PA_{SMC} in the Math9, Tech15, Math21, and Tech23 data sets. It should be noted, also, that PA_{COR} was consistently more accurate than PA_{SMC} when comparing the number of dimensions to the researcher-determined correct number of components (see Table 2). Those minimal differences were slightly larger in PA_{SMC} than in PA_{COR}. Therefore, the results showed that varying the number of randomly generated data sets did not have an impact on SE (see Figure 1).

For IE, the results showed that the mean of the number of dimensions is 1.000 across all 10 sample sizes from 100 to 1,000 in all six factor structures in the final study. As a consequence, IE results would not be focused. However, the results of IE did vary when sample sizes were smaller (i.e., 50 and 75) in a pilot study. Furthermore, K1 and BS also performed poorly. That is, K1 overfactored, and BS required larger sample sizes although it required low sample sizes (n = 100) in a few conditions (i.e., many items in the data).

To compare the consistency of the effective methods, line graphs in Figures 2a, 2b, 2c, 2d, 2e, and 2f display the SEs of the number of dimensions resulting from the six best methods (i.e., PA_{COR95}, PAS_{MC95}, MAP, BS, $MAR_{1.4}$, and IND) across 10 sample sizes in all six factor structures (i.e., Math9, Tech15, Math21, Tech23, Math40, and Tech40).

Discussion of Experiment I

The performance of the six best methods were analyzed based on combinations of the mean, *SE*, *RMSE*, *MAE*, *RE*, percentage of correctness, and bias. The overall results showed that the criteria (including methods) performing the best to worst were: MAP, IND, $PA_{COR95,}$ $MAR_{1.4}$, BS, and IE. MAP performed well when there was only one to two dimensions, even when sample sizes were small ($n \leq 300$). MAP performed poorly when there were unequal numbers of items in each dimension. PA_{COR95} with 100 randomly generated data sets performed the best in most conditions. Using 100 or 1,000 randomly generated data sets yielded trivially different results. Thus, using 100 randomly generated data sets is recommended with PA_{COR95}. PA_{COR95} performed poorly when there were many items in the data. Instead, $MAR_{1.4}$, which performed much better than K1, is recommended for use in these conditions where MAP or PA performed poorly. Furthermore, MAP and PA usually yield the same results based on the results of simulations studies. When inaccurate results occurred, MAP tended to underfactor, whereas PA tended to overfactor. Therefore, researchers are recommended to run MAP and PA together to compare the results (O'Connor, 2000; Zwick & Velicer, 1986). When there were many items in the data, IND and BS performed well depending on the sample size. BS is recommended when only 100 cases are available. When at least 200 cases are available, IND is recommended. For these reasons, we recommended using a combination of these criteria based on conditions of the data available to applied researchers. Finally, IE is not recommended.

Also, larger sample sizes tremendously improved the performance of these criteria across all factor structures. In general, sample sizes of 200, 300, and 400 were recommended to be absolute minimum, acceptable, and desirable sample sizes to yield consistent results. For the impact of factor structures, unequal numbers of items in each dimension and a large number of items in the data showed a negative impact on the precision of the factor extraction results.

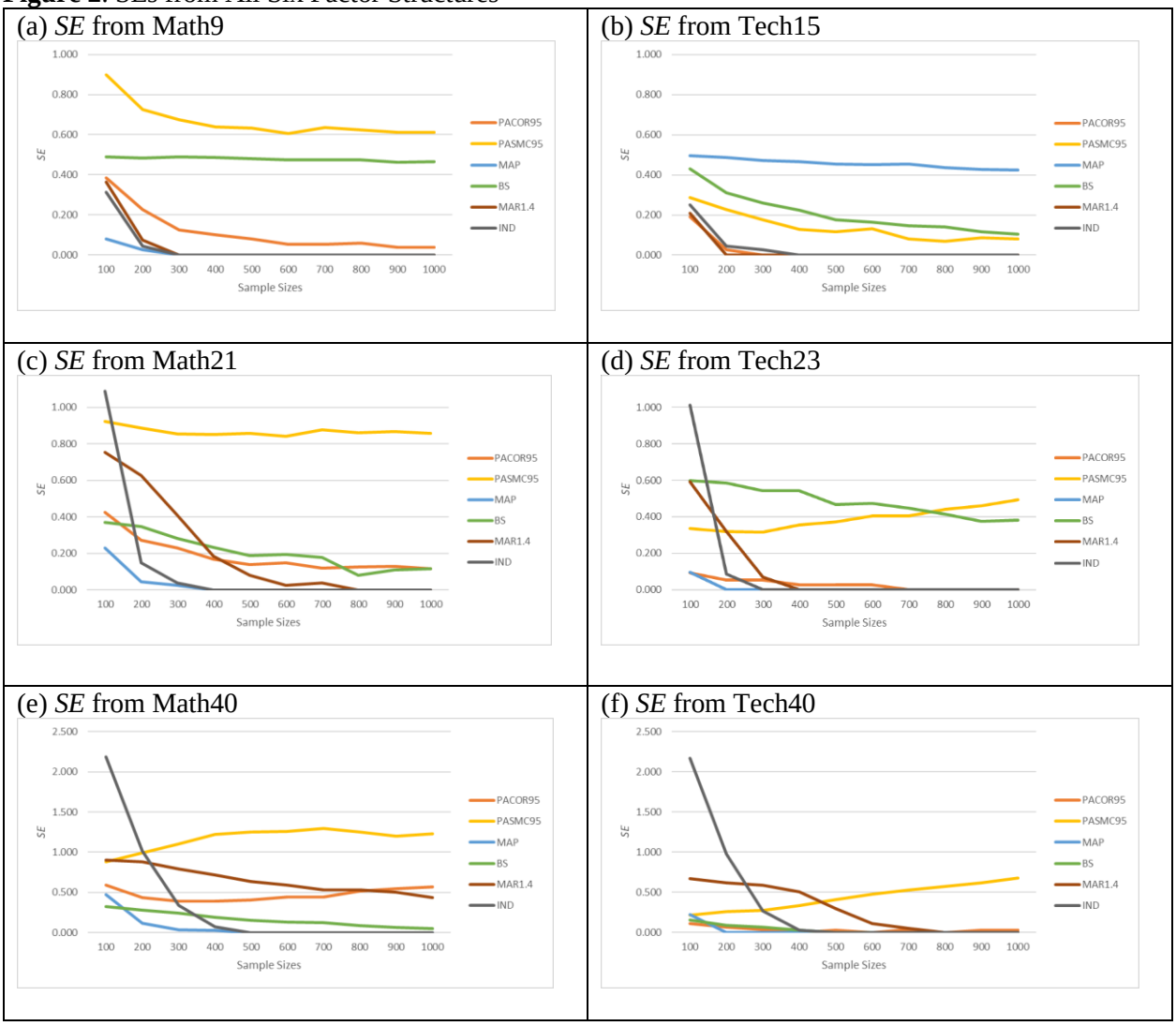
Experiment II

Experiment I using existing real test data (i.e., applied Monte Carlo study) was a doctoral dissertation (Ruengvirayudh, 2018). Even though the results are more applicable to real-world data, fewer variables could be varied and with limitations. Experiment II, used generated data (i.e., true Monte Carlo study where true population values are known), which makes it a true experiment study. That is, more variables could be varied to explore more studied conditions. We use both types of data, real and generated, to investigate dimensionality criteria because each type of data has its own limitations. Therefore, using both data would offset weaknesses of each.

Methods and Data Source of Experiment II

In Experiment II, we created a program in R to generate data for a variety of conditions. We systematically varied sample size from 100 to 600 with the increment of 100 (i.e., 100, 200, 300, 400, 500, and 600) and the number of variables from 6 to 36 with the increment of 6 (i.e., 6, 12, 18, 24, 30, and 36). We also varied the number of underlying dimensions or factors in the population data together with the number of variables in these patterns: (a) for 6 variables: 1, 2, and 3 dimensions (i.e., we generated data

Figure 2. SEs from All Six Factor Structures



that had 1, 2, and 3 underlying dimensions); (b) for 12, 24, and 36 variables: 1, 2, 3, 4, and 6 dimensions; and (c) for 18 and 30 variables: 1, 2, 3, and 6 dimensions. In Experiment I, the number of variables and the number of underlying dimensions could not be varied.

We varied the size of the correlations among factors in the population data and grouped them into three conditions: orthogonal (uncorrelated factors), relatively orthogonal (small correlation among factors), and oblique (relatively strong correlation among factors). This variable could not be varied in Experiment I. For MARs in Experiment II, we varied cut-off values from 1.1 to 1.6 with the increment of 0.1 (i.e., 1.1, 1.2, 1.3, 1.4, 1.5, and 1.6).

Examples of population correlation matrices and their associated population factor loadings were provided in Tables 4 to 6. In these examples, the orthogonal condition had a correlation between factors of $r < .001$. The relatively orthogonal condition had a correlation between factors of $r = .288$, and the oblique condition had a correlation between factors of $r = .473$. Therefore, there were 6 sample sizes, 26 combinations of variables and dimensions, and 3 factor correlations, which resulted in a total of 468 studied conditions.

We generated normally distributed, correlated data to represent a population of 200,000 cases for each studied condition (a combination of sample size, number of variables, number of dimensions, and correlation among factors). We generated data using a population correlation matrix that satisfied the correlations required for each condition. Then, we randomly selected 2,000 samples from each condition's population data. For each sample, we used the criteria for determining the number of dimensions described

Table 3. SE of the Number of Dimensions Resulting from Final Methods across 10 Sample Sizes from the Math9 Data

Criteria		*n*									
No.	**Methods**	**100**	**200**	**300**	**400**	**500**	**600**	**700**	**800**	**900**	**1000**
1	PA_{CORMU}	0.659	0.395	0.260	0.224	0.147	0.135	0.103	0.120	0.045	0.058
2	PA_{COR75}	0.537	0.325	0.193	0.165	0.118	0.096	0.081	0.093	0.037	0.045
3	PA_{COR90}	0.437	0.259	0.151	0.118	0.093	0.063	0.068	0.063	0.037	0.037
4	PA_{COR95}	0.384	0.227	0.126	0.100	0.081	0.052	0.052	0.058	0.037	0.037
5	PA_{COR100}	0.380	0.232	0.163	0.106	0.077	0.063	0.052	0.026	0.045	0.026
6	PA_{COR300}	0.364	0.222	0.155	0.100	0.085	0.052	0.045	0.026	0.045	0.026
7	PA_{COR500}	0.364	0.223	0.157	0.106	0.073	0.052	0.052	0.026	0.045	0.026
8	PA_{COR700}	0.363	0.219	0.159	0.106	0.073	0.052	0.052	0.026	0.045	0.026
9	PA_{COR900}	0.360	0.215	0.161	0.106	0.073	0.052	0.052	0.026	0.045	0.026
10	$PA_{COR1000}$	0.366	0.217	0.159	0.106	0.073	0.052	0.052	0.026	0.037	0.026
11	PA_{SMCMU}	1.555	1.231	1.003	0.946	0.901	0.815	0.796	0.737	0.736	0.718
12	PA_{SMC75}	1.236	1.014	0.872	0.827	0.797	0.735	0.738	0.685	0.695	0.683
13	PA_{SMC90}	0.983	0.833	0.760	0.701	0.691	0.657	0.675	0.655	0.644	0.641
14	PA_{SMC95}	0.900	0.725	0.676	0.638	0.634	0.607	0.635	0.625	0.611	0.611
15	PA_{SMC100}	0.842	0.704	0.651	0.623	0.626	0.619	0.617	0.609	0.607	0.603
16	PA_{SMC300}	0.807	0.695	0.648	0.614	0.616	0.610	0.606	0.605	0.604	0.596
17	PA_{SMC500}	0.838	0.682	0.655	0.609	0.613	0.619	0.617	0.603	0.608	0.599
18	PA_{SMC700}	0.824	0.684	0.652	0.614	0.612	0.616	0.609	0.599	0.608	0.597
19	PA_{SMC900}	0.819	0.688	0.648	0.612	0.609	0.615	0.608	0.601	0.606	0.600
20	$PA_{SMC1000}$	0.822	0.682	0.651	0.606	0.612	0.612	0.615	0.601	0.603	0.601
21	MAP	0.081	0.026	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
22	BS	0.490	0.483	0.488	0.485	0.481	0.474	0.474	0.473	0.462	0.464
23	$MAR_{0.7}$	0.780	0.779	0.773	0.756	0.761	0.769	0.769	0.738	0.719	0.745
24	Average Root (K1)	0.697	0.645	0.579	0.554	0.534	0.514	0.486	0.455	0.421	0.393
25	$MAR_{1.2}$	0.573	0.395	0.199	0.112	0.058	0.037	0.000	0.026	0.000	0.000
26	$MAR_{1.4}$	0.364	0.073	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
27	IE	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
28	IND	0.313	0.045	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Note. Only the 95th percentile was used as a threshold when varying the number of randomly generated data sets in PA_{COR} and PA_{SMC}.

above. For some criteria, multiple options were explored (e.g., PA with the mean, 75th, 90th, and 95th percentiles and thresholds as well as several MAR calculations).

In Experiment II we knew the population characteristics; therefore, we were able to calculate a variety of common Monte Carlo statistics, including bias, RMSE, and correctness. We tested the R code at the individual sample level, across multiple samples, and for piloted conditions. For PA and MAP, we tested and used the R Psych package functions. We programmed and tested K1, IND, IE, BS, and MARs by ourselves. In Experiment II, we focused on the correctness results. The results below focused only on the single most correct version of each criterion from their families (e.g., only PA_{COR95} and PA_{SMC95}). We found no discernible benefit to PA with 900 randomly generated data sets, so the results were for PA methods with only 100 randomly generated data sets (see Ruengvirayudh, 2018).

Results of Experiment II

The overall results for the correctness criterion we used in this study were shown in Table 7. Specifically, it shows the proportion of all 468 studied conditions where the criteria reached 90% correctness. That is, how well the criteria identified the correct number of dimensions at least 1,800 out of 2,000 samples across each condition. Table 7 shows that PA_{COR95} was 90% correct across the most total conditions. This table also shows that PA_{COR95} was most frequently 90% correct for both the orthogonal and relatively orthogonal conditions, but was third to MAP and $MAR_{1.6}$ for the oblique conditions. K1, IE, and BS performed poorly in most conditions, so they were not considered further.

Table 4. Example Orthogonal and Relatively Orthogonal Population Correlation Matrices for 12 Variables (or Items) with 2 Dimensions

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
V1	---	.16	.22	.18	.34	.34	.06	.06	.05	.05	.09	.02
V2	.20	---	.28	.35	.23	.19	.10	.08	.04	.02	.07	.08
V3	.20	.20	---	.30	.22	.34	.08	.10	.10	.06	.09	.08
V4	.20	.20	.20	---	.26	.34	.05	.08	.03	.08	.02	.04
V5	.20	.20	.20	.20	---	.27	.04	.06	.04	.06	.02	.07
V6	.20	.20	.20	.20	.20	---	.01	.05	.08	.04	.10	.07
V7	.00	.00	.00	.00	.00	.00	---	.33	.26	.32	.35	.21
V8	.00	.00	.00	.00	.00	.00	.20	---	.19	.27	.21	.26
V9	.00	.00	.00	.00	.00	.00	.20	.20	---	.36	.33	.24
V10	.00	.00	.00	.00	.00	.00	.20	.20	.20	---	.35	.17
V11	.00	.00	.00	.00	.00	.00	.20	.20	.20	.20	---	.24
V12	.00	.00	.00	.00	.00	.00	.20	.20	.20	.20	.20	---

Note. Orthogonal values in lower triangular half; relatively orthogonal values in upper triangular highlighted half; correlation between relatively orthogonal factors is r = .288; correlation between orthogonal factors is r < .001.

Table 5. Example Oblique and Relatively Orthogonal Population Correlation Matrices for 12 Variables (or Items) with 2 Dimensions

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12
V1	---	.16	.22	.18	.34	.34	.06	.06	.05	.05	.09	.02
V2	.25	---	.28	.35	.23	.19	.10	.08	.04	.02	.07	.08
V3	.29	.32	---	.30	.22	.34	.08	.10	.10	.06	.09	.08
V4	.26	.35	.32	---	.26	.34	.05	.08	.03	.08	.02	.04
V5	.34	.29	.28	.31	---	.27	.04	.06	.04	.06	.02	.07
V6	.34	.27	.34	.34	.31	---	.01	.05	.08	.04	.10	.07
V7	.11	.20	.16	.10	.09	.01	---	.33	.26	.32	.35	.21
V8	.12	.15	.20	.16	.13	.10	.34	---	.19	.27	.21	.26
V9	.09	.07	.20	.06	.08	.15	.31	.27	---	.36	.33	.24
V10	.10	.03	.12	.16	.12	.09	.34	.31	.34	---	.35	.17
V11	.19	.14	.17	.04	.05	.19	.35	.28	.34	.35	---	.24
V12	.04	.16	.16	.08	.14	.14	.28	.30	.30	.26	.30	---

Note. Oblique values in lower triangular half; relatively orthogonal values in upper triangular highlighted half; half; correlation between relatively orthogonal factors is r = .288; correlation between oblique factors is r = .473.

Table 6. Example Factor Loadings for Correlation Matrices with 12 Variables and 2 Dimensions

	Orthogonal		**Relatively Orthogonal**		**Oblique**	
Item	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**	**Factor 1**	**Factor 2**
1	.45		.47		.60	
2	.45		.45		.51	.06
3	.45		.52		.57	
4	.45		.57		.59	
5	.45		.50		.59	
6	.45		.59		.49	
7		.45		.59		.53
8		.45		.46		.52
9		.45		.52	.10	.52
10		.45		.58		.59
11		.45		.59		.56
12		.45		.40		.60

Note. The factor loadings in the table are after a Promax Rotation.

Because it was apparent that the overall results might not hold across all conditions consistently, we created Figure 3 to show the average number of correct samples for each sample size for each factor correlation condition. Several criteria did not work well when $N \leq 200$, even if they worked well with larger sample sizes. The same criteria identified in the overall results (Table 7) could be seen generally to have the most correctness across sample sizes. One curious finding was that PA_{SMC95} generally became less correct as the sample sizes increased, except for the orthogonal conditions.

The criteria broken down across the number of variables were shown in Figure 4. Because we determined that smaller sample sizes produced less than desirable results, these figures show the averages only when $N \geq 300$. Again, we could see the relative consistency of the correctness results for PA_{COR95} that it did not perform quite as well when the number of variables increased. Alternatively, MAP, $MAR_{1.6}$, and IND generally improved their correctness as the number of variables increased. However, these criteria had lower than desirable correctness when the number of variables was only 6 or 12.

Because it became obvious that there were interactions among the conditions, we created Figures 5 to 7 to show examples of the results for the better criteria broken down by both sample sizes and number of variables. One of the results that became particularly clear was that most of the criteria failed as the structure became more diffuse (i.e., fewer variables per dimension, such as Figure 5e) as compared to more concentrated structures (i.e., more variables per dimension, such as Figure 5a), especially when sample sizes were smaller.

Finally, we explored the accuracy of combinations of criteria in terms of proportion of correctness across the relatively orthogonal conditions. Tables 8 and 9 show the results when combining criteria and only when $N \geq 200$. Table 10 shows that PA_{COR95} & MAP & $MAR_{1.6}$ produced 95% correctness (i.e., they agreed on the correct number of dimensions in at least 95% of those agreed-upon samples if they agreed at all) in 73 of 86 conditions (85%) when $N \geq 300$. Note that there were 18 conditions with more diffuse structures where PA_{COR95} & MAP & $MAR_{1.6}$ never agreed. When excluding these three most diffuse conditions where no criteria worked consistently well (e.g., 3 or fewer variables per dimension), agreement on the combined criteria reached 95% correctness for 97% of the 84 conditions.

Discussion of Experiment II

Despite the popularity of confirmatory factor analysis (CFA), there are still many circumstances when applied researchers will want to perform EFA or PCA as data reduction techniques. Therefore, determining the correct number of factors to extract in such analyses will continue to be a critical matter. Generally speaking, and across many of the conditions studied, PA using the 95th percentile (PA_{COR95}) was most frequently correct, followed by MAP, $MAR_{1.6}$, and IND. However, we found that using agreement among a combination of criteria could produce results that challenged any individual criterion. In particular, our recommendation is that applied researchers use a combination of PA_{COR95} & MAP & $MAR_{1.6}$, with $N \geq 300$. When the combination does not agree, then use one of the better individual criteria. However, when the structure is diffuse (e.g., only two or three items per factor), then applied researchers need much larger sample sizes or may not find a strong solution at all.

Conclusions and Scholarly Significance

Determining the correct number of dimensions in the data is crucial to validity in many fields, such as checking dimensionality in item response theory (IRT), identifying the number of sub-categories in an exam, end-of-semester evaluation form, survey data, and so on. Choosing criteria to determine the number of dimensions correctly and consistently is a critical step prior to performing exploratory factor analysis (EFA). Based on our results from existing real test data in Experiment I and generated data in Experiment II, we found new specific and general findings for common and not-so-common criteria.

Our results also suggest several other new, specific findings. First, all criteria suffer as the number of dimensions among the variables increased (e.g., two or three variables per dimension, such as 12 total items with 4 or 6 dimensions), especially with smaller numbers of variables and small sample sizes (e.g., 6 items with 3 dimensions). Therefore, we recommend that researchers use extreme caution when performing EFA or PCA with a relatively small number of variables or items. However, it should be noted that, based on the real data we used in Experiment I, we used loadings that most would consider moderate. If loadings are much stronger than we used, then perhaps these criteria will perform better with a large number of dimensions relative to the number of items.

Table 7. Proportion of 120 Sample Size, Number of Variables, and Number of Dimensions Conditions, where N ≥ 200 and Variable-to-Dimension Ratio ≥ 3, and where Criteria Reached 90% Correctness (out of 2,000 Samples for Each Condition) Sorted by Total Percentage Correct.

	Orthogonal	Relatively Orthogonal	Oblique	Total
Parallel Analysis (PA_{COR95})	.958	.625	.600	.728
Modified Average Root ($MAR_{1.6}$)	.742	.633	.717	.697
Minimum Average Partial	.583	.650	.633	.622
Indicator Function	.733	.525	.442	.567
Parallel Analysis (PA_{SMC95})	.833	.058	.158	.350
Broken Stick	.208	.333	.333	.291
Imbedded Error	.250	.250	.250	.250
Kaiser's Criterion	.092	.058	.117	.089

Figure 3. Average Correctness of Criteria across Sample Sizes for All Conditions

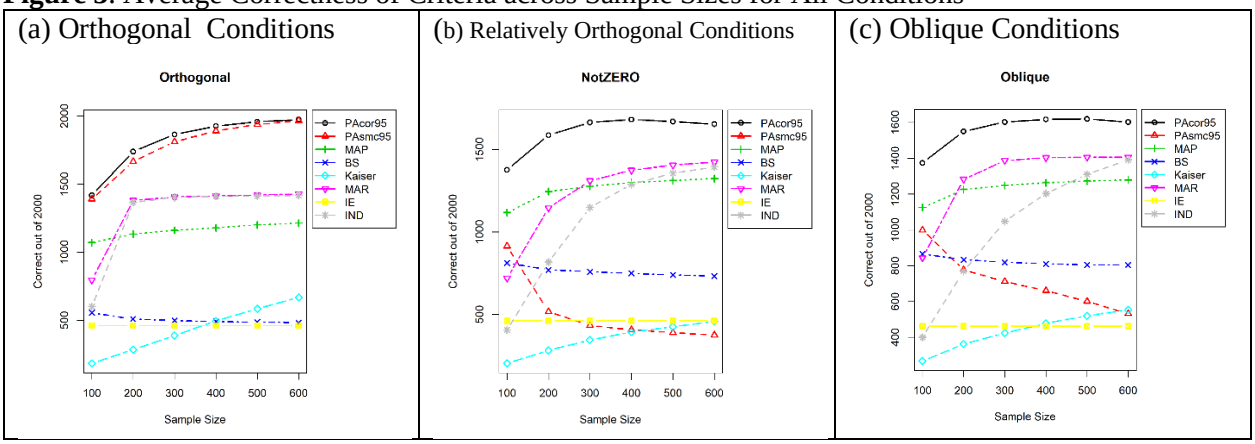
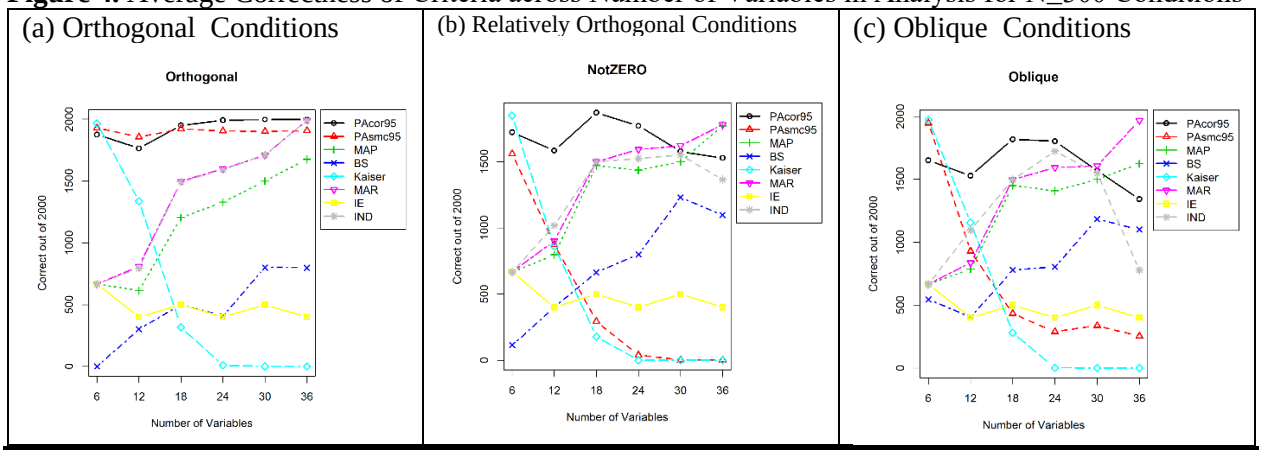


Figure 4. Average Correctness of Criteria across Number of Variables in Analysis for N≥300 Conditions



Second, PA_{COR95} outperformed PA_{SMC95} in terms of correctness in almost all conditions, arguing in favor of using the original (unreduced) correlation matrices with 1s diagonal even for PA used with EFA. That is, we recommend using the original correlation matrix for PA for both PCA and EFA (i.e., we do not recommend using the reduced correlation matrix with SMCs on the diagonal for PA used with EFA). Further, PA_{COR95} performed the best in most conditions, followed by MAP, $MAR_{1.6}$, and IND. MAP performed quite well across many conditions, and outperformed PA_{COR95} when there were fewer dimensions in the data, even when sample sizes were small (N ≤ 300). Using 100 or 1,000 randomly generated data sets in PA yielded trivially different results. Thus, using 100 randomly generated data sets is recommended with PA_{COR95}.

Third, we recommended that applied researchers use a combination of PA_{COR95} & MAP & $MAR_{1.6}$ or just PA_{COR95} & MAP, with $N \geq 300$ whenever possible. When these methods agreed, there was a high level of correctness for the resulting choice of number of dimensions. When the combination does not agree, then use one of the better individual criteria (usually PA_{COR95}, but perhaps MAP if conditions are right). We also recommend exploring dimensionality both above and below the number recommended by any criteria used to look for the most theoretically defensible factor results. For example, if the criterion recommends 3 dimensions, then the researcher should examine both the 2 dimensional and 4 dimensional solutions to see if one makes more sense theoretically.

Fourth, one relatively unknown approach, IND, and one new approach, $MAR_{1.6}$, performed admirably across many conditions. Indeed, $MAR_{1.6}$ worked well enough that we would recommend that 1.6 become the default eigenvalue for comparison in statistical programs (e.g., SPSS, where eigenvalues greater than 1 is the default). $MAR_{1.6}$ is easy to perform by changing a default eigenvalue of 1.0 (i.e., K1) to 1.6 as a minimum eigenvalue to determine a factor, and IND formula is not terribly complicated to calculate.

Fifth, a general conclusion about sample size is that determining the number of dimensions for EFA may require larger sample sizes than most scholars have recommended for EFA itself. We found that $N \geq 300$ is a desired minimum for correctness in most conditions and most criteria. No criteria performed well across conditions with N = 100. Unlike many EFA sample size recommendations, we found that a minimum of N = 300 is more important than a cases-per-variable rule (e.g., 10 or 20 cases per variable or item). Furthermore, we also found that sample sizes must be larger when the structure in the variables is more diffuse. Nevertheless, no criteria performed well across conditions with a variable-to-dimension ratio of 2:1 (e.g., 6 items and 3 factors). Applied researchers cannot know true structure from sample data, but if the researchers decide on several dimensions for a smaller number of variables, they should have a large number of cases to have more confidence in their decision.

Figure 5. Correctness of Criteria for *V*=12 Variables across All Sample Sizes and Number of Dimensions (*D*) for the Relatively Orthogonal Conditions**.**

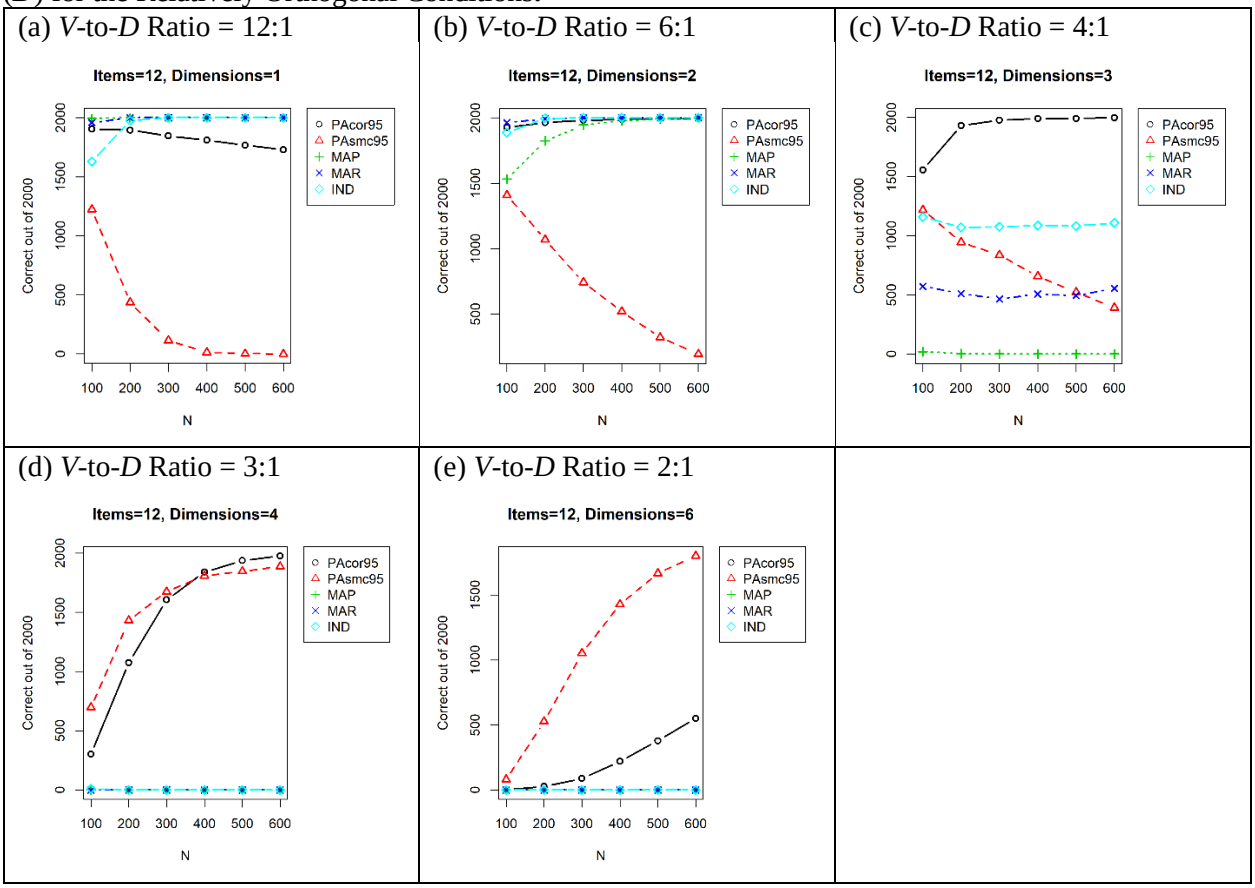


Figure 6. Correctness of Criteria for $V = 24$ Variables across All Sample Sizes and Number of Dimensions (D) for the Relatively Orthogonal Conditions

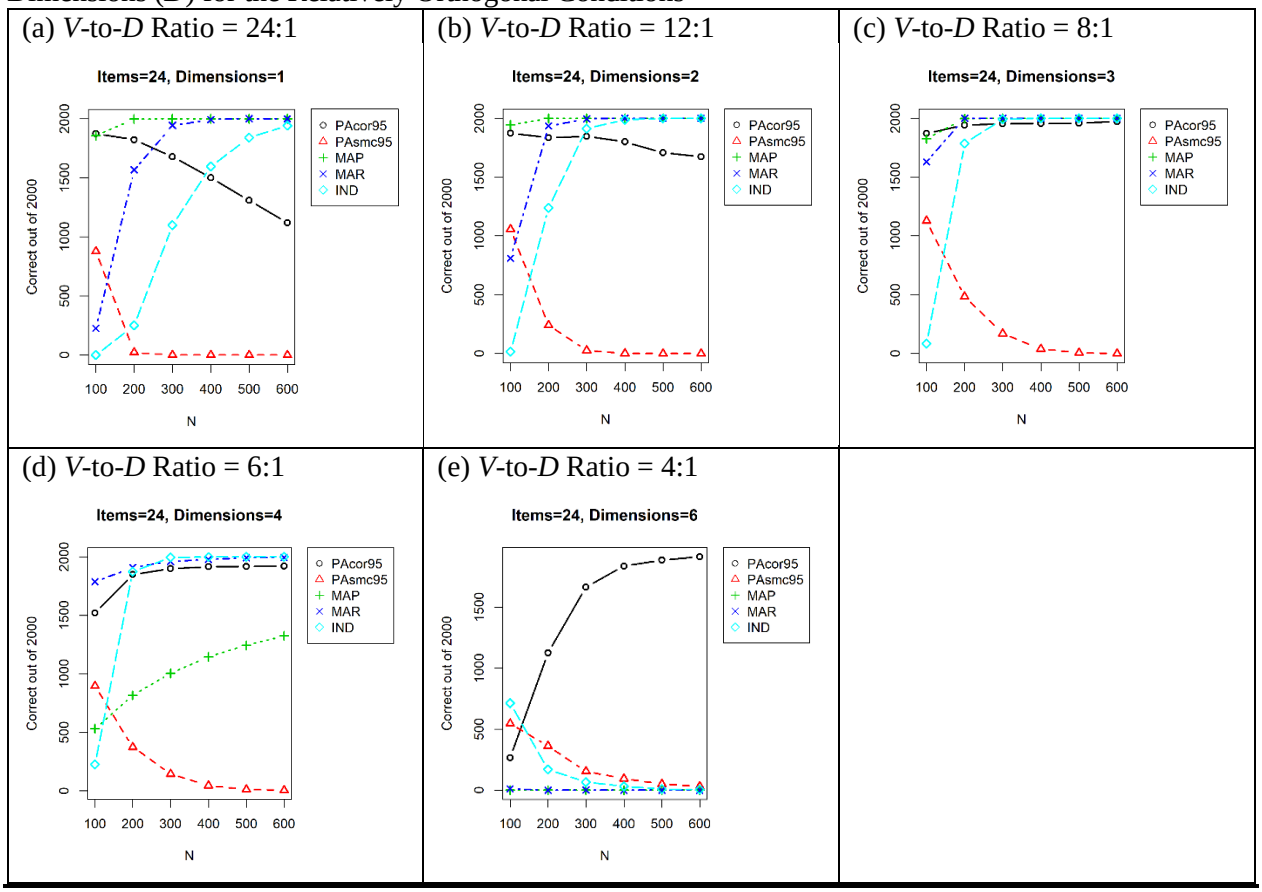


Table 8. Proportion of 120 Sample Size, Number of Variable, and Number of Dimensions Conditions, where N ≥ 200 and Variable-to-Dimension Ratio ≥ 3, and where Average Bias Less than 0.5 Reached 90% (out of 2,000 Samples for Each Condition) Sorted by Total Percentage

	Orthogonal	Relatively Orthogonal	Oblique	Total
Parallel Analysis (PA_{COR95})	.917	.842	.842	.867
Modified Average Root ($MAR_{1.6}$)	.750	.708	.742	.733
Minimum Average Partial	.608	.683	.667	.653
Indicator Function	.750	.583	.567	.633
Parallel Analysis (PA_{SMC95})	.917	.117	.258	.431
Broken Stick	.258	.358	.442	.353
Imbedded Error	.250	.250	.250	.250
Kaiser's Criterion	.167	.142	.183	.164

Table 9. Proportion of 120 Sample Size, Number of Variable, and Number of Dimensions Conditions, where N ≥ 200 and Variable-to-Dimension Ratio ≥ 3, and where Average RMSE Less than 0.5 Reached 90% (out of 2,000 Samples for Each Condition) Sorted by Total Percentage

	Orthogonal	Relatively Orthogonal	Oblique	Total
Parallel Analysis (PA_{COR95})	.967	.842	.725	.845
Modified Average Root ($MAR_{1.6}$)	.750	.683	.725	.719
Minimum Average Partial	.583	.658	.650	.630
Indicator Function	.750	.583	.508	.614
Parallel Analysis (PA_{SMC95})	.900	.075	.192	.389
Broken Stick	.225	.333	.375	.311
Imbedded Error	.250	.250	.250	.250
Kaiser's Criterion	.133	.108	.158	.133

Figure 7. Correctness of Criteria for $V = 36$ Variables across All Sample Sizes and Number of Dimensions (D) for the Relatively Orthogonal Conditions.

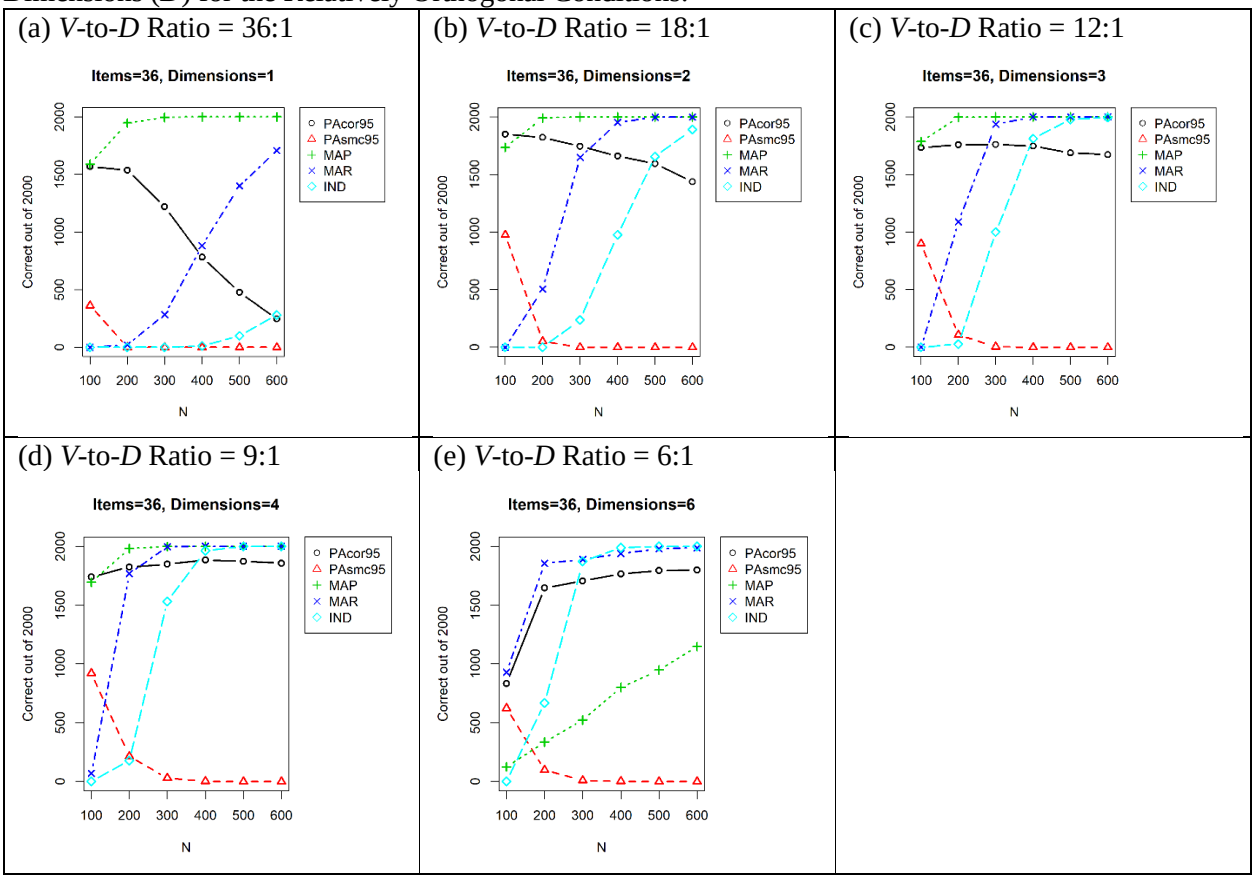


Table 10. Percentage of the Best Combination Criteria across Conditions where the Agreement of Criteria Reached 95% Correctness

	When N ≥ 300	When N ≥ 200	Total
PA_{COR95} & MAP & $MAR_{1.6}$	73 of 86 (85%)	90 of 111 (81%)	97 of 136 (71%)
PA_{COR95} & MAP	73 of 87 (84%)	91 of 112 (81%)	101 of 138 (73%)
PA_{COR95} & MAP & IND	72 of 86 (84%)	89 of 110 (81%)	97 of 127 (76%)
MAP & $MAR_{1.6}$	71 of 100 (71%)	86 of 126 (68%)	92 of 151 (61%)
MAP & IND	71 of 102 (70%)	88 of 127 (69%)	95 of 144 (66%)

Note. First number represents the number of conditions where agreement was at least 90% correct; second number represents the total number of conditions with non-zero agreement; number in parentheses represents the percentage of conditions with non-zero agreement that resulted in at least 90% correctness.

References

Buja, A., & Eyuboglu, N. (1992). Remarks on parallel analysis. *Multivariate Behavioral Research, 27*(4), 509-540. https://doi.org/10.1207/s15327906mbr2704_2

Chou, Y.-T., & Wang, W.-C. (2010). Checking dimensionality in item response models with principal component analysis on standardized residuals. *Educational and Psychological Measurement, 70*(5), 717–731. https://doi.org/10.1177/0013164410379322

Cota, A. A., Longman, R. S., Holden, R. R., Fekken, G. C., & Xinaris, S. (1993). Interpolating 95th percentile eigenvalues from random data: An empirical example. *Educational and Psychological Measurement, 53*(3), 585-596. https://doi.org/10.1177/0013164493053003001

Debelak, R., & Tran, U. S. (2013). Principal component analysis of smoothed tetrachoric correlation matrices as a measure of dimensionality. *Educational and Psychological Measurement, 73*(1), 63-77. https://doi.org/10.1177/0013164412457366

Garrido, L. E., Abad, F. J., & Ponsoda, V. (2013). A new look at Horn's parallel analysis with ordinal variables. *Psychological Methods, 18*(4), 454-474. https://doi.org/10.1037/a0030005

Glorfeld, L. W. (1995). An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement, 55*(3), 377-393. https://doi.org/10.1177/0013164495055003002

Green, S. B., Levy, R., Thompson, M. S., Lu, M., & Lo, W. J. (2012). A proposed solution to the problem with using completely random data to assess the number of factors with parallel analysis. *Educational and Psychological Measurement, 72*(3), 357-374. https://doi.org/10.1177/0013164411422252

Hill, B. D., & Ray, C. M. (2014, April). *A misunderstanding and misuse of parallel analysis.* Paper presented at the annual meeting of the American Educational Research Association, Philadelphia, PA.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185. https://doi.org/10.1007/BF02289447

Hubbard, R., & Allen, S. J. (1987). An empirical comparison of alternative methods for principal component extraction. *Journal of Business Research, 15*(2), 173-190. https://doi.org/10.1016/0148-2963(84)90047-X

Humphreys, L. G., & Ilgen, D. R. (1969). Note on a criterion for the number of common factors. *Educational and Psychological Measurement, 29*(3), 571-578. https://doi.org/10.1177/001316446902900303

Jackson, J. E. (1991). *A user's guide to principal components*. Wiley series in probability and mathematical statistics. John Wiley & Sons.

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. I: Artificial data. *Journal of the Royal Statistical Society. Series C (Applied Statistics), 21*(2), 160-173.

Jolliffe, I. T. (1986). *Principal component analysis*. Springer series in statistics. New York, NY: Springer-Verlag.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement, 20*(1), 141-151. https://doi.org/10.1177/001316446002000116

Loehlin, J. C. (2011). *Latent variable models: An introduction to factor, path, and structural equation analysis* (4th ed.). Routledge.

Malinowski, E. R. (1977a). Determination of the number of factors and the experimental error in a data matrix. *Analytical Chemistry, 49*(4), 612-617. https://doi.org/10.1021/ac50012a027

Malinowski, E. R. (1977b). Theory of error in factor analysis. *Analytical Chemistry, 49*(4), 606-612. https://doi.org/10.1021/ac50012a026

Meyers, L. S., Gamst, G., & Guarino, A. J. (2013). *Applied multivariate research: Design and interpretation* (2nd ed.). Sage.

NIETS. (2013). รูปแบบข้อสอบ สาระและมาตรฐานการเรียนรู้ ข้อสอบ O-NET ป.6 ม.3 ม.6 ปีการศึกษา 2556. Content regarding topics in each subject in the 2013 O-NET examination for sixth, ninth, and twelfth grades. http://www.niets.or.th/th/content/view/1735

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*(3), 396-402. https://doi.org/10.3758/BF03200807

Raîche, G. (2005). Dichotomous & polytomous category information. *Transactions of the Rasch Measurement SIG American Educational Research Association, 19*(1), 1005-1012.

Raîche, G., Walls, T. A., Magis, D., Riopel, M., & Blais, J.-G. (2013). Non-graphical solutions for Cattell's scree test. *Methodology, 9*(1), 23-29. https://doi.org/10.1027/1614-2241/a000051

Ruengvirayudh, P. (2018). *A Monte Carlo study of parallel analysis, minimum average partial, indicator function, and modified average roots for determining the number of dimensions with binary variables in test data: Impact of sample size and factor structure* (Unpublished doctoral dissertation, Ohio University). http://rave.ohiolink.edu/etdc/view?acc_num=ohiou151516919677091

Smith, R. M., & Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice Volume 2* (pp. 316-327). Ablex Publishing Corporation.

Timmerman, M. E., & Lorenzo-Seva, U. (2011). Dimensionality assessment of ordered polytomous items with parallel analysis. *Psychological Methods, 16*(2), 209-220. https://doi.org/10.1037/a0023353

van der Eijk, C., & Rose, J. (2015). Risky business: Factor analysis of survey data – Assessing the probability of incorrect dimensionalisation. *PLoS ONE, 10*(3), 1-31. https://doi.org/10.1371/journal.pone.0118900

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41*(3), 321-327. https://doi.org/10.1007/BF02293557

Warne, R. T., & Larsen, R. (2014). Evaluating a proposed modification of the Guttman rule for determining the number of factors in an exploratory factor analysis. *Psychological Test and Assessment Modeling, 56*(1), 104-123. https://psycnet.apa.org/record/2014-19867-006

Warner, R. M. (2013). *Applied statistics: From bivariate through multivariate techniques* (2nd ed.). Sage.

Weng, L. J., & Cheng, C. P. (2005). Parallel analysis with unidimensional binary data. *Educational and Psychological Measurement, 65*(5), 697-716. https://doi.org/10.1177/0013164404273941

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99*(3), 432-442. https://doi.org/10.1037/0033-2909.99.3.432.

Send correspondence to:
Pornchanok Ruengvirayudh
Loyola Marymount University
Email: porsche.ruengvirayudh@lmu.edu