Supplemental Materials for "Assessing dimensionality in dichotomous items when many subjects have all zero responses: An example from psychiatry and a solution using mixture models"

William F. Christensen, Department of Statistics, Brigham Young University
Melanie M Wall, Department of Psychiatry and Department of Biostatistics, Columbia University
Irini Moustaki, Department of Statistics, London School of Economics

Supplemental Appendix S1: Impact of all-zero cases on tetrachoric correlations

As noted in Section 4 of the article, the tendency towards unidimensionality when an inflated number of all-zero cases is added to the data is due to the resulting increased tetrachoric correlations between items. Specifically, even if groups of items exhibit large intragroup correlations but no correlation between items in different groups, adding a large number of all-zero cases will make all intra-item correlations exhibit strong positive correlation—the tell-tale sign of unidimensionality. Figure S1 illustrates the average eigenvalue profile for Scenarios 1, 4, 1a, and 3a of Table 1. The eigenvalues for Scenario 1 (moderate-severity items, ≈1% all-zero cases among the 4,000 pathological subjects, and no all-zero cases appended) exhibit the classic profile for a two-factor structure. However, in Scenario 1a (when 36,000 all-zero cases are added to the 4,000 pathological subjects measured on moderate-severity items), the eigenvalue structure strongly resembles the traditional profile for a single factor model. When the mix of items is 20,000 pathological cases plus 20,000 all-zero cases (as in Scenario 3a), the eigenvalue structure remains distorted, but the increased ratio of pathological to healthy subjects results in estimates for the second-highest eigenvalue that are large enough to yield a dimensionality assessment of two. We note that when an abundance of all-zero cases arises from a scenario where the items are very severe/difficult (as in Scenario 4), many of the dimensionality-assessment techniques (including the eigenvalue-based techniques) work just fine. That is, the eigenvalues for Scenario 4 (high-severity items, ≈20% all-zero cases among the 4,000 pathological subjects, and no all-zero cases appended) exhibit a structure that is quite similar to the eigenvalue structure for moderate-item-severity Scenario 1. It is only those scenarios where additional all-zero subjects are included who are healthy and the trait is not relevant that the presence of the inflated zeros is problematic. We conclude that the presence of additional all-zero subjects (as when healthy subjects are mixed with pathological subjects in community-based samples) will erroneously lead researchers to underestimate the dimensionality of the phenomena measured by the items. We also note that removing the all-zero cases will distort the eigenvalues for the data when high-severity items are the cause of all-zero cases. Note that the eigenvalues for Scenario 4 in Figure S2 differ from the eigenvalues for Scenario 4 in Figure S1.

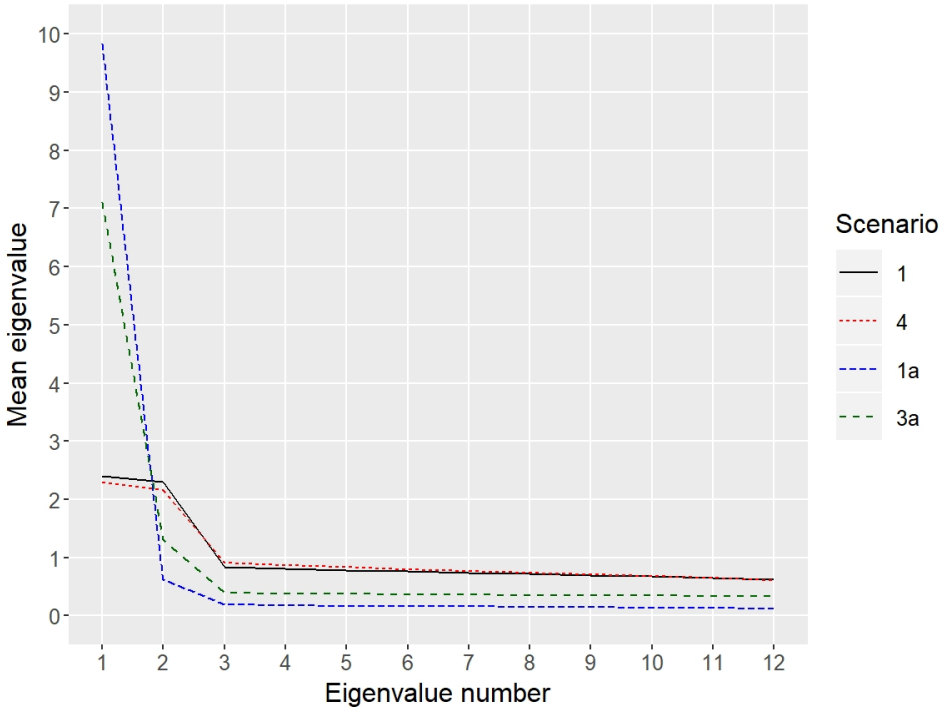

Figure S1: Mean eigenvalues for the tetrachoric correlation matrix when generating data according to Scenarios 1, 4, 1a, and 3a from Table 1 and using all observations to calculate tetrachoric correlations. Scenario 1 has moderate-severity items (≈1% all-zero cases) generated for 4000 pathological cases with no all-zero (i.e., "healthy") subjects added. Scenario 4 has high-severity items (≈20% all-zero cases) generated for 4000 pathological cases with no all-zero (i.e., "healthy") subjects added. Scenario 1a has moderate-severity items (≈1% all-zero cases among the 4000 pathological cases) with 36000 all-zero (i.e., "healthy") subjects added. Scenario 3a has moderate-severity items (≈1% all-zero cases) generated for 20000 pathological cases with 20000 all-zero (i.e., "healthy") subjects added.

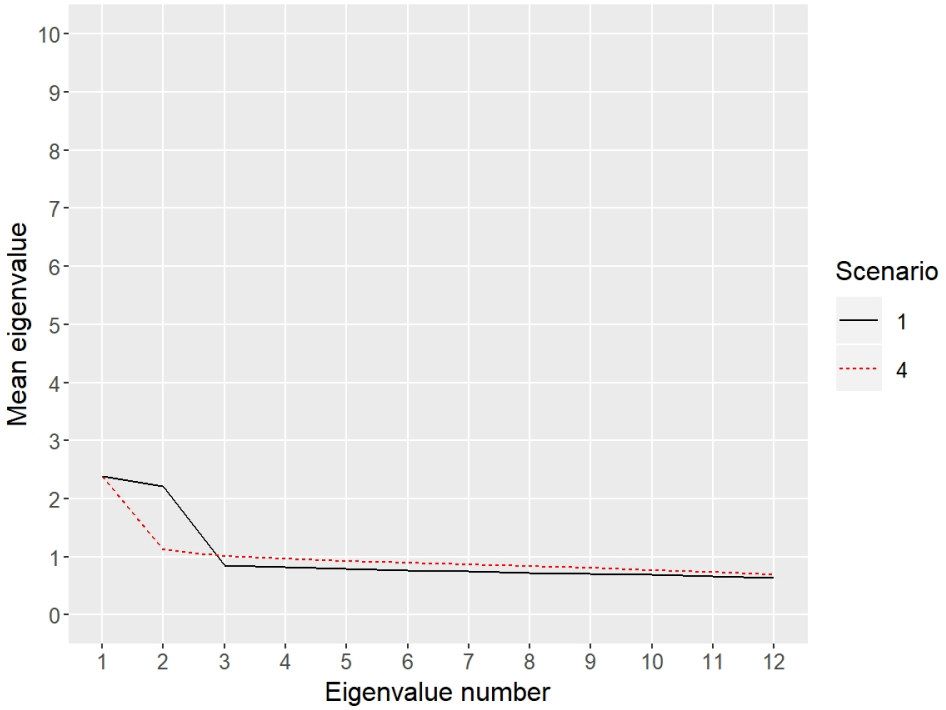

Figure S2: Mean eigenvalues for the tetrachoric correlation matrix when generating data according to Scenarios 1 and 4, and when correlations are calculated after removing the all-zero observations. Scenario 1 has moderate-severity items (≈1% all-zero cases) generated for 4000 pathological cases with no all-zero (i.e., "healthy") subjects added. Note that because the all-zero observations are removed, the results for Scenarios 1 and 1a in Table 1 will be identical. Scenario 4 has high-severity items (≈20% all-zero cases) generated for 4000 pathological cases with no all-zero (i.e., "healthy") subjects added.

Supplemental Appendix S2: Simulation Study with Small Sample Sizes

In this Supplement, we address several additional issues that could affect the performance of dimensionality-assessment tools when data include a mixture of pathological cases along with healthy cases (who exhibit no symptoms). Specifically, we consider scenarios in which: (i) factors are correlated, (ii) the data-generation model is truly unidimensional, and (iii) sample sizes are small. Data are generated in a manner similar to the simulations in Section 4 of the manuscript but with three different factor structures: one factor, two uncorrelated factors, and two correlated factors (cor(f_1, f_2) = 0.5). Additionally, the sample sizes are chosen to be much smaller than in the simulations in Section 4; the set of observed cases generated for each scenario was set to either 80 pathological plus 320 healthy cases or 200 pathological plus 200 healthy cases.

Results for these simulations are summarized in Table S1, which is structured in the same way as Table 1 in the article, but with one exception. With sample sizes much smaller in this simulation, we considered the use of the sample-size adjusted BIC (aBIC) in place of BIC. For use with the AZ-EFA fit, we report $aBIC_{AZ\text{-}EFA}$ in Table S1 because it performed better than $aBIC_{AZ\text{-}EFA}$. For the traditional factor analysis model (fit with maximum likelihood), we report BIC_{ML} because it outperformed $aBIC_{ML}$.

Comparing the third and fourth rows (where the two factors are uncorrelated) with the fifth and sixth rows (where the two factors are correlated), one can note that the presence or absence of correlation among the two generated factors has little impact on the dimensionality assessments. We also note that the unidimensionality of the true model (rows 1 and 2) is consistently identified ($\geq$ 97% of the time) for each of the dimensionality-assessment methods considered. The effect of smaller sample sizes proved to be more noteworthy. When the number of pathological cases was 200, both the "Rule of 1" and $aBIC_{AZ\text{-}EFA}$ were able to identify a two-factor model at least 97% of the time. However, regardless of whether we used $aBIC_{AZ\text{-}EFA}$ or $BIC_{AZ\text{-}EFA}$ to determine the dimensionality of the data, neither the AZ-EFA approach nor any of the other metrics considered were able to properly identify a two-factor model when the number of pathological cases was reduced to 80. That is, when the number of pathological cases among the observations is expected to be small (e.g., less than 200), the true dimensionality of a phenomenon is likely to be understated using any of these methods; AZ-EFA is a useful tool only when the data contain a moderate to large number of pathological subjects.

Supplemental Appendix S3: Factor loading estimates: traditional EFA vs AZ-EFA

We supplement our discussion of the simulations by examining the factor loading estimates of the traditional EFA fit with weighted least squares and the AZ-EFA. The X^2_{WLS} method with traditional EFA assessed the correct dimensionality between 92 and 99 percent of the time and the AZ-EFA method yielded a correct dimensionality assessment rate of 100 percent for each of the nine scenarios. The AZ-EFA approach yields not only the best assessments of item-set dimensionality, but also superior parameter estimates. We generally place particular interest in the factor loading estimates associated with a factor analysis model. Comparing the factor loading estimates for the maximum likelihood fit of both the standard factor analysis model and the AZ-EFA for Scenarios 1-6 (described in Table 1), we observe few differences in the estimated factor loadings. However, when additional all-zero observations are included in the sample—as in Scenarios 1a, 2a, and 3a from Table 1, the estimates of the factor loading estimates exhibit substantial bias. Figure S3 illustrates the factor loading estimates associated with both the two-factor standard factor analysis model and the two factor AZ-EFA when the data are generated according to Scenario 3a from Table 1. Note that even when half of the 40,000 observations in a simulated community-based sample are pathological, all of the unconstrained factor loadings associated with the standard factor analysis model are badly biased—factor loadings with true values of 1 have estimates that are tightly clustered around 2.94 and factor loadings with true values of 0 have estimates that are tightly clustered around −0.39. As the percent pathological among the 40000 observations drops to 10% (as in Scenario 1a), the bias in the factor loading estimates can become even worse. Consequently, we determine from these simulations that the all-zero observations should not be removed, and that using the AZ-EFA with BIC as the dimensionality assessment criterion is the superior approach for the factor analysis of dichotomous data with many all-zero observations.

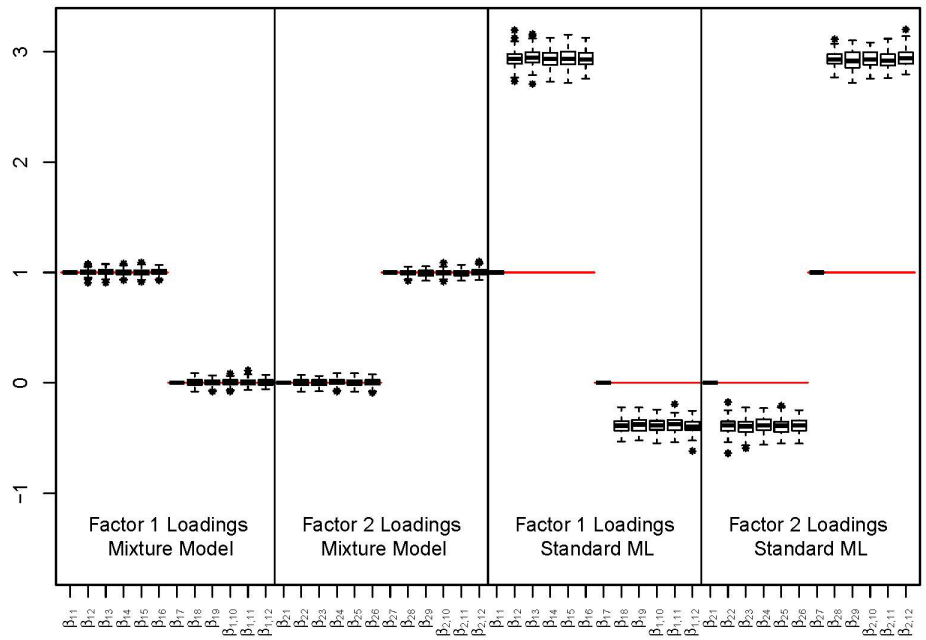

Figure S3: Estimated factor loadings (β_{ij}) when using maximum likelihood to estimate the parameters of the two-factor AZ-EFA (left side) and the standard two-factor analysis model (right side). Data were generated according to Scenario 3a in Table 1— moderate severity items ($\beta_0 = 0.25$) with the number of pathological subjects (n_p) equal to 20,000 and the number of nonpathological subjects equal to 20,000. Red lines indicate the actual factor loadings used to generate the data (see Model (2)).

Supplemental Appendix S4: Estimated Proportions for the Mixture Groups

The mixture proportions η_1 and η_2 represent the proportions of the sample that are associated with the degenerate ("healthy subject") and pathological groups, respectively. In order to evaluate the effectiveness of the AZ-EFA model, we provide in Table S2 the mean (and standard deviation) for η_1, expressed as a percentage. For each of the simulation scenarios discussed in Section 4 and summarized in Table 1 of the manuscript, we provide in Table S2 a description of the percentage of healthy cases in the sample, and the expected percentage of all-zero cases in the sample. Scenarios 1 through 6 have no healthy subjects added to the sample, but differing amounts of all-zero responses are expected, depending on the severity of the items; 1% of the pathological subjects are expected to have all-zero responses in Scenarios 1 through 3, while 20% of pathological subjects have all-zero responses in Scenarios 4 through 6. In contrast, the all-zero responses in Scenarios 1a, 2a, and 3a arise from each of the two mixture classes: a large portion from the presence of healthy subjects exhibiting no symptoms, and a smaller portion from the least severe cases among the pathological group.

As illustrated in Table S2, the mixing coefficient in the AZ-EFA model is remarkably well-estimated in each scenario, provided that the number of factors estimated is equal to (or greater than) the true number of factors (which is two in these simulations). When the fitted number of factors is two, the AZ-EFA approach consistently provides estimates of the healthy proportion (η_1) that is near zero for Scenarios 1 through 6; the mean value for the healthy proportion never exceeds 0.9%, even when 20% of the pathological cases yield all-zero responses. The estimation of the proportion in the healthy class remains accurate in Scenarios 1a, 2a, and 3a. Despite the fact that the expected proportion of all-zero responses is greater than the number of healthy subjects in the sample, the two-factor AZ-EFA fitted model has a mean value that is—rounded to the nearest tenth of a percent—equal the true representation of healthy subjects. Further, in all cases where the number of fitted factors is at least two, the variability in the estimates of η_1 is low. For example, in Scenario 3a the middle 95% of η_1 estimates are in the range (49.9%, 50.1%). The only examples of AZ-EFA inaccurately estimating the mixing proportion occurred when the estimated number of factors was insufficient for the data (i.e., when one factor was estimated in our simulations). The estimation of mixing proportions in the follow-up simulations in Appendix S2 was equally well-behaved.

Supplemental Appendix S5: Example Mplus code

```
TITLE:      Mplus code for all-zero inflated exploratory factor analysis (AZ-EFA) model
            with p=12 variables and q=2 factors;
DATA:       file="symptoms.txt";
VARIABLE:   names are x1 x2 x3 x4 x5 x6 x7 x8 x9 x10 x11 x12;
            classes = c (2);
            usevariables are x1-x12;
            categorical are x1-x12;
ANALYSIS:   type is mixture;
            algorithm = integration;
            algorithm = ODLL;
            stscale is 1;
            starts 50 10;
MODEL:
            %OVERALL%
            f1 by x1 x2-x6 x8-x12;
            f2 by x7 x2-x6 x8-x12;
                        ! x1 has a loading of 1 on f1 and a loading of 0 on f2
                        ! x7 has a loading of 0 on f1 and a loading of 1 on f2
            %c#1%
            f1 by x1@1 x2@1 x3@1 x4@1 x5@1 x6@1 x8@1
                  x9@1 x10@1 x11@1 x12@1;
            f2 by x7@1 x2@1 x3@1 x4@1 x5@1 x6@1 x8@1
                  x9@1 x10@1 x11@1 x12@1;
                        !In the healthy subgroup, all loadings are fixed to equal 1
            [x1$1@1 x3$1@1 x4$1@1 x5$1@1 x6$1@1 x7$1@1 x8$1@1
                  x9$1@1 x10$1@1 x11$1@1 x12$1@1];
            [x7$1@1 x2$1@1 x3$1@1 x4$1@1 x5$1@1 x6$1@1 x8$1@1
                  x9$1@1 x10$1@1 x11$1@1 x12$1@1];
                        !In the healthy subgroup, all thresholds are fixed to equal 1
            f1-f2 @0;
                        ! In the healthy subgroup, var(f1) = var(f2) = 0
            [f1-f2 @-100];
                        ! In the healthy subgroup, E(f1) = E(f2) = −100
            f1 with f2 @0;
                        ! In the healthy subgroup, f1 and f2 are uncorrelated
            %c#2%
            f1-f2@1;
                        ! In the pathological subgroup, var(f1) = var(f2) = 1
            [f1-f2 @0];
                        ! In the pathological subgroup, E(f1) = E(f2) = 0
            f1 with f2;
                        ! In the pathological subgroup, f1 and f2 are correlated
```

Table S1. Dimensionality assessments in follow-up simulation. For each scenario, the body of the table gives the percent of the cases where the method selects 1 factor, 2 factors, or 3+ factors. Results for the correct model are bolded. Each of the scenarios is characterized by the number of pathological cases in the data, the number of all-zero (i.e., non-pathological or healthy) cases added to the pathological sample, and the number of factors used to generate the simulated data—either 1 factor, 2 uncorrelated factors, or 2 correlated factors. All scenarios in this simulation had an expected percentage of all-zero cases equal to 1% (with $\beta_0 = 0.25$ in equation (1)). In contrast to the simulation in Table 1, for this simulation's much smaller sample sizes, the adjusted Bayesian Information Criterion ($aBIC_{AZ\text{-}EFA}$) is reported because it was superior to $BIC_{AZ\text{-}EFA}$ in assessing the dimensionality of the AZ-EFA model. For evaluating the traditional factor analysis with ML, the BIC_{ML} metric was superior to $aBIC_{ML}$.

			Eigenvalue-based methods						**Traditional FA model (fit with WLS)**									**Traditional FA model (fit with ML)**						**AZ-EFA (fit with ML)**		
			"Rule of 1" (# evals > 1)			**Parallel analysis**			**X^2_{WLS}**			**RMSEA**			**CFI**			**X^2_P (Pearson)**			**BIC_{ML}**			**$aBIC_{AZ\text{-}EFA}$**		
# of pathological cases	**# of all-zero cases added**	**# of factors**	**1**	**2**	**3+**	**1**	**2**	**3+**	**1**	**2**	**3+**	**1**	**2**	**3+**	**1**	**2**	**3+**	**1**	**2**	**3+**	**1**	**2**	**3+**	**1**	**2**	**3+**
80	320	1	**100**	0	0	**100**	0	0	**99**	0	1	**99**	0	0	**99**	0	0	**100**	0	0	**100**	0	0	**100**	0	0
200	200	1	**100**	0	0	**100**	0	0	**97**	1	2	**99**	0	0	**99**	0	0	**1**	0	99	**100**	0	0	**98**	2	0
80	320	2 (uncorr)	73	**27**	0	100	**0**	0	25	**73**	2	100	**0**	0	100	**0**	0	100	**0**	0	98	**2**	0	39	**60**	0
200	200	2 (uncorr)	0	**100**	0	0	**100**	0	1	**95**	4	23	**77**	0	99	**1**	0	0	**0**	100	16	**84**	0	0	**99**	0
80	320	2 (corr)	78	**22**	0	100	**0**	0	34	**66**	0	99	**1**	0	100	**0**	0	100	**0**	0	93	**7**	0	52	**46**	2
200	200	2 (corr)	1	**99**	0	0	**99**	1	2	**94**	4	38	**62**	0	96	**4**	0	0	**0**	100	24	**76**	0	1	**97**	2

Table S2. Average estimated mixing coefficient (η_1) for AZ-EFA approach when data are generated from the two-factor model. For each simulation scenario, the mean (standard deviation) for the estimated proportion of healthy subjects is given (rounded to the nearest tenth of a percent). Average values for the estimate of η_1 are reported for fits to the 1-, 2-, and 3-factor models, with values very close to the true percentage of healthy cases when the correct (2-factor) model is fit.

Scenario	% healthy cases in sample	% expected all-zero cases in sample	# of estimated factors 1	2	3
1	0%	1%	0.8% (0.2%)	0.1% (0.1%)	0.1% (0.1%)
2	0%	1%	0.8% (0.1%)	0.1% (0.1%)	0.1% (0.1%)
3	0%	1%	0.8% (0.1%)	0.0% (0.1%)	0.1% (0.1%)
4	0%	20%	12.2% (1.5%)	0.9% (0.9%)	0.7% (0.9%)
5	0%	20%	12.6% (0.5%)	0.7% (0.6%)	0.7% (0.6%)
6	0%	20%	12.4% (1.3%)	0.5% (0.4%)	0.7% (0.6%)
1a	90%	90.1%	90.1% (0.0%)	90.0% (0.0%)	90.0% (0.0%)
2a	70%	70.3%	70.2% (0.0%)	70.0% (0.0%)	70.0% (0.1%)
3a	50%	50.5%	50.4% (0.0%)	50.0% (0.1%)	49.9% (0.3%)