Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 17, Number 15, November 2012 ISSN 1531-7714

Replication Analysis in Exploratory Factor Analysis: What it is and why it makes your analysis better

Jason W. Osborne, *Old Dominion University*
David C. Fitzpatrick, *North Carolina State University*

Exploratory Factor Analysis (EFA) is a powerful and commonly-used tool for investigating the underlying variable structure of a psychometric instrument. However, there is much controversy in the social sciences with regard to the techniques used in EFA (Ford, MacCallum, & Tait, 1986; Henson & Roberts, 2006) and the reliability of the outcome. Simulations by Costello and Osborne (2005), for example, demonstrate how poorly some EFA analyses replicate, even with clear underlying factor structures and large samples. Thus, we argue that researchers should routinely examine the stability or volatility of their EFA solutions to gain more insight into the robustness of their solutions and insight into how to improve their instruments while still at the exploratory stage of development.

Exploratory factor analysis (EFA) is a widely-used technique fraught with controversy, debate, and misconception. For example, there is lingering debate over the best extraction techniques to use, when particular rotation techniques are appropriate, how to decide the number of factors to extract and interpret, how large of a sample is sufficient for a good solution, whether results of an EFA can be used in a "confirmatory" fashion to test hypotheses, how generalizable EFA results can be generalized, and so forth (e.g., Costello & Osborne, 2005; Henson & Roberts, 2006; MacCallum, Widaman, Preacher, & Hong, 2001; Tabachnick & Fidell, 2001).

Of course, exploratory factor analysis was created so that researchers could *explore* the structure of their data. It was not meant to serve the same purpose as confirmatory factor analyses or inferential analyses, despite the fact that EFA is often (incorrectly) used for that purpose.[1] Access to statistical computing power has allowed EFA to proliferate and expand, affording researchers easy access to complex analyses of the psychometric properties of their instruments. Indeed, this access to computing power has also allowed development of more complex estimation procedures (e.g., maximum likelihood) and the spreading use of *confirmatory* factor analysis and latent variable modeling. Today, exploratory and confirmatory factor analyses are among the most common types of analyses reported in social science journals (Osborne, Costello, & Kellow, 2008).

However, researchers should keep in mind the exploratory nature of EFA. It is, by nature, quirky, temperamental, valuable, and interesting. Exploratory factor analysis, like all models that fit weighted linear combinations to data, takes advantage of all the information in the interrelationships between variables, whether those interrelationships are representative of the population or not. In other words, as with prediction in multiple regression (e.g., Osborne, 2000; Osborne, 2008), EFA tends over-fit a model to the data—in other words, when the same model is applied to a new sample, the model is rarely as good a fit. It is often close, but sometimes wildly different, as with multiple regression. We as readers have no way of knowing whether the results reported are likely to be the former or the latter, and we believe that it would benefit the literature to report this more nuanced information.

Why is replication important in EFA, and what determines replicability?

Many authors reporting the results of EFA use confirmatory language despite the exploratory nature of the analyses. We need to re-emphasize in our discipline that EFA is *not* a mode for testing of hypotheses or *confirming*

[1] Indeed, Henson & Roberts (2006) reported that approximately one-third of EFAs in high quality, measurement related journals should have been performed as CFAs instead but were not.

ideas (e.g., Briggs & Cheek, 1986; Floyd & Widaman, 1995), but rather for exploring the nature of scales and item inter-relationships. EFA merely presents a solution based on the available data.

These solutions are notoriously difficult to replicate, even under abnormally ideal circumstances (exceptionally clear factor structure, very large sample to parameter ratios, strong factor loadings, and high communalities; (Costello & Osborne, 2005; Osborne, et al., 2008). As mentioned already, many point estimates and statistical analyses vary in how well they will generalize to other samples or populations (which is why we are more routinely asking for confidence intervals for point estimates). But EFA seems particularly problematic in this area.

To underscore this point, we will remind readers of two previously-published findings (Costello & Osborne, 2005; Osborne, et al., 2008). The first is that the robustness and accuracy of EFA benefits from large samples. Traditional rules of thumb for inferential statistics have advised having ten participants per group or per variable minimum (e.g., Cohen & Cohen, 1983), and some authors (e.g., Baggaley, 1983; P. T. Barrett & Kline, 1981; Guadagnoli & Velicer, 1988) have attempted to provide similar guidance in EFA (although these rules of thumb fail to recognize that the number of parameters estimated are far greater than simply the number of items in a scale). For example, Comrey and Lee (Comrey & Lee, 1992) suggested that sample sizes of 300 are "good," 500 are "very good," and 1000 are "excellent." Yet is a sample of 300 still "good" if a scale has 300 items and five subscales, producing 1500 parameter estimates? Authors such as Stevens (2002) have provided recommendations ranging from 5-20 participants per scale item, with Jöreskog and Sörbom (1996) encouraging at least *10 participants per parameter estimated.*

In EFA it is not just the number of items that matters, but the number of parameters estimated that matters in producing an accurate, replicable result. An EFA with 10 items that extracts two factors produces 20 estimated parameters (not counting eigenvalues, communalities, etc.). Given 20 items and three extracted factors, the EFA estimates 60 parameters, again just for the factor loadings. Having strong factor loadings, communalities, and sample size all benefit the robustness of an EFA (Osborne & Costello, 2004).

Costello and Osborne (2005) examined the effect of sample size on an instrument with a very clear, strong factor structure (e.g., the Rosenberg Self-View Inventory (Rosenberg, 1965)) with respect to aspects of factor analysis that matter: whether items were assigned to the correct factor, and when they were, how much the factor loadings varied. After hundreds of simulations with real educational data varying sample size, EFAs were found to be relatively unstable. At a sample size of 10 participants per item, only 60% of the analyses reproduced the expected factor structure, and at 20 participants per item, only 70% reproduced the factor structure of this instrument that is noted for a strong, clear factor structure. The Rosenberg SVI is unusually strong in terms of factor loading and structure. Obviously with more complex or less strong factor structures, replication gets increasingly problematic.

The second piece of information is that surveys of the literature indicate that many (or most) studies reporting EFAs fail to reach thresholds that indicate strong probabilities of replication. For example, Henson and Roberts' (2006) survey of *Educational and Psychological Measurement*, *Journal of Educational Psychology*, *Personality and Individual Differences*, and *Psychological Assessment* indicates a median sample size of 267 for reported EFAs, mean participant:item ratio of 11, and a median of 60 parameters (20 items x 3 factors) estimated. Extrapolating the simulations reported from Costello and Osborne to those findings from Henson and Roberts, we would expect less than 60% of these EFAs to replicate basic factor structure (number of factors extracted, which items load on which factor). Another survey by (Osborne, et al., 2008) notes that the majority of EFAs reported in prominent social science journals (63.2%) are performed with samples that consist of less than 10 participants per item. Similarly, two-thirds of EFAs reported in these studies surveyed fail to provide the expectation that they will be at least 60% replicable under the best circumstances.

We find this troubling, and you should too. Of course, we are extrapolating from our simulations, and the details of each EFA are unique. We find the current situation troubling, but in reality have no specific information about how replicable we should expect particular factor structures to be because direct tests of replicability are almost never published. As Thompson (1999) and others note, replication is a key foundational principle in science.

Let's bring replication to EFA.

Authors can (and, we argue, should) directly estimate the replicability of their exploratory factor analyses reported in scientific journals. Authors (e.g., Thompson, 2004) have introduced replicability procedures for EFA, similar to those procedures considered best practices in validation of prediction equations in multiple regression (Osborne, 2000, 2008). Although few authors perform the procedure, it has intuitive appeal.

Specifically, since the goal of EFA is usually to infer or explore the likely factor structure of an instrument when used within a particular population, it is important to know whether a solution (or evident factor structure) within a particular data set is likely to be observed within another,

similar data set.[2] The lowest threshold for replicability should be replicating the same basic factor structure (same number of factors extracted, same items assigned to each factor) within a similar sample. A more rigorous threshold for replicability would be seeing the same number of factors extracted, the same items assigned to the same factors, and the same range of magnitudes of factor loadings (within reason). Stronger replicability gives researchers more confidence that a particular scale will behave as expected in data subsets or a new sample.

The EFA replication procedures we demonstrate will allow researchers to provide readers information about the extent to which their EFAs meet these reasonable and basic expectations for replicability.

Replication or cross-validation in the literature. In the clinical literature, the use of factor scores (weighted averages of items based on factor loadings) is a contentious issue as factor loadings (and as noted above, even factor structure) often vary across groups, thus leading identical patient or participant responses to vary considerably across samples where factor loadings differ . Thus, for example, Floyd and Widaman (1995) suggest cross-validation procedures for factor scores, similar to those recommended for regression prediction equations. This recommendation highlights the importance of knowing how well a solution within one sample – even a very large, representative sample—generalizes.

Similarly, Briggs and Cheek (1986) argued almost three decades ago that one of the critical concerns to personality psychologists (and personality measurement) should be replicability of factor structure, demonstrating replicability issues within a commonly used Self-Monitoring scale.

One high-profile application of EFA replication techniques was an ambitious attempt by Costa and McRae (1997) to examine whether the commonly-held Five Factor Model of personality generalized across six different translations of their revised NEO personality inventory. In this application, strong replication across cultures and languages including English, German, Portuguese, Hebrew, Chinese, Korean, and Japanese samples not only confirmed the goodness of the translations of the instrument, but the universality of the five factor model.

[2] As a field, we have traditionally referred to scales as "reliable" or "unidimensional", but methodologists since Lord and Novick (1968) caution that *instruments* do not have reliability, only *scores from particular samples* do (see also Wilkinson and the Task Force on Statistical Inference, 1999). Despite this, we should have a reasonable expectation for instruments to have the same basic structure across samples if we are to have any rational basis for the science of measurement within the social sciences.

Procedural aspects of replicability analysis

For those familiar with shrinkage analyses and cross-validation of prediction equations in multiple regression, these procedures and suggestions will hopefully feel familiar. Replicability analyses in EFA (e.g., Thompson, 2004) can be conducted in two different ways: via *internal* or *external* replication. In internal replication, the researcher splits a single data set into two samples via random assignment. In external replication, the researcher uses two separately gathered datasets. In brief, replicability analysis occurs as follows:

1. EFA is conducted on each sample by extracting a fixed number of factors using a chosen extraction method (i.e., maximum likelihood) and rotation method (i.e., oblimin or varimax).

2. Standardized factor loadings are extracted from the appropriate results for each sample (e.g., pattern matrix if using an oblique rotation), creating a table listing each item's loading on each factor within each sample.

3. Factor loadings and structures are then compared.

Unfortunately, references on this topic do not go into depth as to how researchers should perform this comparison and what the criteria is for strong vs. weak replication, and how to summarize or quantify the results of the replication.

Quantifying Replicability in Exploratory Factor Analysis

Researchers since the early 1950s have been proposing methods of quantifying and summarizing this sort of analysis. We start this part of the discussion with the reminder that invariance analysis in confirmatory factor analysis should be considered the gold standard for attempting to understand whether an instrument has the same factor structure across different groups (randomly constituted or otherwise).

For authors not at the stage of development where invariance analysis via CFA is appropriate, scholars since the 1950s have been proposing methods of summarizing replication analyses in EFA (and criticizing other proposals). Our position is that since replication with EFA is also exploratory, and preliminary to a more rigorous CFA analysis, simple summary measures are to be preferred.

One method of summarizing replication analyses include a family of coefficients first presented by Kaiser, Hunka, and Bianchini (1971). This "similarity coefficient" utilized the cosines between the unrotated and rotated axes, but had faulty assumptions (and therefore are invalid from a mathematical point of view, ten Berge (1996); see also Barrett (1986)) and could yield similarity coefficients that

indicate strong agreement when in fact there was little agreement. Thus, they are inappropriate for this purpose.

Tucker (1951) and Wrigley and Neuhaus (1955) have presented congruence coefficients that seem less problematic (ten Berge, 1986) but are also controversial (c.f., P. Barrett, 1986). For example, Tucker's (1951) Congruence Coefficient examines the correlations between factor loadings for all factor pairs extracted. Yet as Barrett (1986) correctly points out, these types of correlations are insensitive to the magnitude of the factor loadings, merely reflecting the patterns. [3] For our purposes, which is to examine whether the factor structure and magnitude of the loadings are generally congruent, this insensitivity to magnitude of loadings is problematic. We prefer a more granular analysis that examines (a) whether items are assigned to the same factors in both analyses, and (b) whether the individual item factor loadings are roughly equivalent in magnitude—the former being the basic threshold for successful replication, the latter being a more reasonable, stronger definition of replication.

Assessing whether the basic factor structure replicated. As mentioned above, regardless of whether the researcher is performing *internal* (a single sample, randomly split) or *external* (two independently gathered samples) replication, the researcher needs to perform the same EFA procedure on both, ideally specifying the same number of factors to be extracted, the same extraction and rotation procedures, etc. Researchers should then identify the strongest loading for each item (i.e., which factor does that item "load" on), and confirm that these are congruent across the two analyses. For example, if item #1 has the strongest loading on Factor 1, and item #2 has the strongest loading on factor #2, that pattern should be in evidence in both analyses. If any items fail this test, we would consider these analyses to fail to meet the most basic threshold of replicability: structural replicability. There is therefore little reason to expect factor structure to replicate in any basic way in future samples.

If there are a small percentage of items that seem volatile in this way, this replication analysis may provide important information—that these items might need revision or deletion. Thus, replication can also serve important exploratory and developmental purposes. If a large number of problematic items are observed, this represents an opportunity for the researcher to revise the scale substantially before releasing it into the literature, where this volatility might be problematic.

[3] We could go on for many more pages summarizing various historical approaches to summarizing congruence. For the sake of parsimony we will simply refer the readers to the above-cited resources that give thorough coverage of the issues.

Assessing strong replication in EFA. If a scale passes the basic test of having items structurally assigned to same factors, the other important criterion for strong replication is confirming that the factor loadings are roughly equivalent in magnitude. We believe that because we are still in exploration mode, simple metrics serve our goal well. We advocate for simply subtracting the two standardized (rotated) factor loadings for congruent items, and squaring the difference. Squaring the difference has two benefits: eliminating non-important negative and positive values (if one loading is .75 and one is .70, subtracting the first from the second produces a -0.05, and subtracting the second from the first produces a 0.05, yet the direction of the difference is unimportant—only the magnitude is important) and highlighting larger differences. Researchers can then quickly scan the squared differences, and either confirm that all are small and unimportant, or identify which items seem to have large differences across replication analyses.

An example of replication analysis.

Replication analysis is demonstrated on a scale developed by the first author, the School Perceptions Questionnaire (SPQ; (Osborne, 1997)) within a data set consisting of 1908 participants from several community colleges around the USA. The SPQ is a scale of 13 questions designed to measure identification with academics (also called selective valuing or domain identification in the self-concept literature; (for a recent article on this concept, see Osborne & Jones, 2011) . Appendix 1 lists the SPQ Scale questions.

To demonstrate this technique, we used *internal replicability analysis*, randomly splitting the original sample into two samples that were then analyzed separately using specific extraction and rotation guidelines based on prior analyses of the scale. In this example we report a two-factor solution (the factor structure suggested by previous research on the scale) as well as 3- and 4-factor solutions to demonstrate how mis-specification of a factor model can quickly become evident through replication analysis.

2-factor replication analysis. Replication of this scale fails to meet the initial criterion, structural replication. Specifically, looking at the factor loadings in Table 1, you can see Question 12 has the highest factor loading on Factor #2 in the first analysis and on Factor #1 in the second analysis. This item is probably not a good one, and would benefit from revision or deletion. All other items have their strongest loading on congruent factors, so if we delete Question 12, we would say that the factor structure of the scale meets the basic level of replication. The next step is to look at the squared differences in the factor loadings. These range from 0.0000 to 0.01, indicating that the largest difference between the standardized factor loadings is |.10|-

- which is not bad. We would suggest that once the squared differences achieve a magnitude of .04—indicating a difference of |.20| -- that is when a researcher may begin to consider factor loadings volatile.

Table 1: 2 Factor SPQ Replicability Analysis, Maximum Likelihood Extraction, Oblimin Rotation with 250 max iterations

	Sample 1			Sample 2			
	Comm-unality	Factor Load		Comm-unality	Factor Load		Squared Diff
	Extract	1	2	Extract	1	2	
Question 1	0.42	0.66	0.07	0.36	0.60	0.01	0.0036
Question 2	0.29	0.51	-0.11	0.29	0.51	-0.07	0.0000
Question 3	0.26	-0.24	0.41	0.23	-0.18	0.39	0.0004
Question 4	0.33	-0.16	0.52	0.36	-0.08	0.57	0.0025
Question 5	0.44	0.64	-0.10	0.48	0.71	0.04	0.0049
Question 6	0.28	0.08	0.54	0.19	0.03	0.44	0.0100
Question 7	0.12	0.06	0.35	0.14	0.02	0.38	0.0009
Question 8	0.26	0.52	0.09	0.31	0.58	0.09	0.0036
Question 9	0.39	-0.44	0.36	0.39	-0.50	0.24	0.0036
Quest 10	0.28	0.54	0.13	0.27	0.54	0.06	0.0000
Quest 11	0.50	0.71	0.00	0.54	0.74	0.01	0.0009
Quest 12	0.35	-0.34	0.42	0.38	-0.45	0.31	*Failed*
Quest 13	0.15	0.10	0.40	0.22	0.07	0.49	0.0081
Eigen Val		2.76	1.60		3.06	1.66	
Min	0.12			0.14			
Max	0.50			0.54			

3-factor replication analysis. As mentioned above, this should replicate poorly as a 3-factor solution is not a strong solution for this scale. As you can see in Table 2, problems are immediately obvious. Even with such a large sample, two of the thirteen items failed to replicate basic structure—in other words, they loaded on non-congruent factors. Further, Question 8 is problematic because it is not clear what factor to assign it to in the first analysis (it loads 0.32 on both factors 1 and 3), whereas in the second analysis it loads strongly on Factor 1, so it could be argued that three of the thirteen items failed basic structural replication. Beyond these three, the squared differences for the loadings were within reasonable range (0.0000-0.0225) except for Question 8, which had a 0.0529, reflecting a large change in factor loading from 0.32 to 0. 55. This would be a second red flag for this item, if the researcher decided to let the issue of structural replication pass.

4-factor replication analysis. As you can see in Table 3, the basic structural replication fails dramatically – and unsurprisingly—with ten of the thirteen items loading on noncongruent factors. Of the other three, one changes from 0.99 to -0.58, which represents a massive shift in magnitude, another shifts from -0.52 to 0.33, again a relatively large shift, and the final one shifts modestly from 0.44 to 0.37. In almost every way, this analysis demonstrates everything that can go wrong with a replication analysis.

Table 2: 3 Factor SPQ Replicability Analysis, Maximum Likelihood Extraction, Oblimin Rotation with 250 max iterations

	Sample 1				Sample 2				
	Comm-unality	Factor Load			Comm-unality	Factor Load			Squared Diff
	Extract	1	2	3	Extract	1	2	3	
Question 1	0.45	0.43	0.13	0.39	0.39	0.57	-0.01	0.18	.0196
Question 2	0.57	0.13	0.05	0.70	0.47	0.45	-0.11	0.41	*failed*
Question 3	0.36	0.01	0.32	-0.45	0.36	-0.11	0.45	-0.32	*failed*
Question 4	0.34	-0.04	0.47	-0.24	0.35	-0.06	0.57	-0.10	.0100
Question 5	0.45	0.60	-0.13	0.10	0.47	0.69	0.04	0.01	.0081
Question 6	0.30	0.01	0.55	0.05	0.18	0.02	0.43	0.06	.0144
Question 7	0.15	-0.04	0.39	0.12	0.14	0.00	0.36	0.08	.0009
Question 8	0.27	0.32	0.14	0.32	0.34	0.55	0.07	0.17	.0529
Question 9	0.39	-0.39	0.36	-0.13	0.45	-0.54	0.25	0.18	.0225
Quest 10	0.31	0.57	0.08	-0.02	0.27	0.54	0.06	-0.03	.0009
Quest 11	0.60	0.76	-0.06	0.00	0.56	0.76	0.02	-0.13	.0000
Quest 12	0.38	-0.37	0.45	-0.02	0.52	-0.51	0.32	0.31	*failed*
Quest 13	0.16	0.17	0.35	-0.15	0.21	0.08	0.48	-0.05	.0169
Eigen Val		2.45	1.46	1.84		3.09	1.69	0.58	
Min	0.15				0.14				
Max	0.60				0.56				

Appropriately large samples make a difference. In Table 4 we replicate the two-factor analysis presented in Table 1 but with two random samples of N=100 each, much smaller than the almost N=1000 samples in Table 1. In this analysis, you can see two of the thirteen items loaded on non-concordant factors (interestingly, not the originally-troublesome Question 12), and two more items had troublingly large differences in factor loadings. Question 1 loaded 0.77 in the first analysis and 0.56 in the second analysis. As you can see from the communality estimates, that led to a large decrease in the communality for this item—and a squared difference of over 0.04. Additionally, Question 7 had a loading of 0.82 in the first analysis and 0.39 in the second analysis, again leading to a large change in communality and a squared difference of 0.1849. Thus, even if a researcher deleted the two troublesome items, two others showed non-replication of magnitude of factor loading. As previous authors have noted, EFA is a large-sample procedure, and replications with relatively small samples may lead to more volatility than one would see with larger samples. With over 900 in each sample, this scale seems reliably replicable, but with only 100 in each sample there are some serious questions about replicability.

Table 3: 4 Factor SPQ Replicability Analysis, Random 50% sample, Maximum Likelihood Extraction, Oblimin Rotation with 250 max iterations

	Sample 1					Sample 2					
	Comm-unality	Factor Load				Comm-unality	Factor Load				Squared difference
	Extract	1	2	3	4	Extract	1	2	3	4	
Qu. 1	0.47	0.60	0.21	0.11	0.22	0.39	-0.05	0.43	0.01	0.30	*failed*
Qu. 2	0.58	0.59	0.18	-0.22	0.39	0.50	-0.03	0.23	-0.02	0.58	*failed*
Qu. 3	0.36	-0.45	0.14	0.34	-0.13	0.34	0.06	0.04	0.35	-0.40	*failed*
Qu. 4	0.35	-0.40	0.18	0.38	0.12	0.37	0.00	-0.04	0.55	-0.16	*failed*
Qu. 5	0.46	0.64	0.10	0.12	-0.14	0.48	-0.07	0.62	0.00	0.10	*failed*
Qu. 6	0.29	-0.20	0.44	0.19	0.13	0.23	-0.04	-0.06	0.48	0.07	*failed*
Qu. 7	0.73	-0.17	0.80	-0.19	-0.14	0.16	0.04	-0.03	0.39	0.08	*failed*
Qu. 8	0.27	0.44	0.23	0.05	0.13	0.34	-0.14	0.34	0.14	0.30	*failed*
Qu. 9	0.41	-0.58	0.06	0.14	0.23	1.00	0.99	0.05	-0.02	0.00	.1681
Qu. 10	0.35	0.45	0.09	0.37	0.00	0.35	0.08	0.63	-0.04	-0.03	*failed*
Qu. 11	0.59	0.68	0.15	0.27	-0.17	0.58	-0.16	0.68	-0.02	-0.04	*failed*
Qu. 12	0.39	-0.52	0.14	0.13	0.30	0.43	0.33	-0.32	0.25	0.10	.0361
Qu. 13	0.18	-0.10	0.16	0.37	0.09	0.21	0.03	0.09	0.44	-0.09	.0049
Eigen		2.58	1.13	1.23	2.17		2.24	2.67	1.46	1.45	
Min	0.18					0.16					
Max	0.73					1.00					

Table 4: 2 Factor SPQ Replicability Analysis, Sample N= 100, Maximum Likelihood Extraction, Oblimin Rotation

	Sample 1			Sample 2			Squared difference
	Comm	Factor Load		Comm	Factor Load		
	Extract	1	2	Extract	1	2	
Question 1	0.55	0.77	0.12	0.31	0.56	0.06	.0441
Question 2	0.42	0.62	-0.07	0.39	0.57	-0.17	.0025
Question 3	0.29	-0.05	0.52	0.35	-0.07	0.57	.0025
Question 4	0.27	-0.34	0.30	0.35	-0.02	0.58	*failed*
Question 5	0.56	0.68	-0.18	0.37	0.62	0.07	.0036
Question 6	0.32	-0.02	0.56	0.30	0.02	0.55	.0001
Question 7	0.62	0.24	0.82	0.15	0.06	0.39	.1849
Question 8	0.34	0.61	0.21	0.33	0.56	-0.03	.0025
Question 9	0.40	-0.49	0.28	0.40	-0.35	0.46	*failed*
Quest 10	0.21	0.46	0.02	0.32	0.58	0.08	.0144
Quest 11	0.46	0.64	-0.11	0.49	0.71	0.03	.0049
Quest 12	0.50	-0.42	0.46	0.24	-0.31	0.34	.0144
Quest 13	0.19	-0.09	0.40	0.35	0.16	0.60	.0400
Eigen		2.76	1.60		3.06	1.66	
Min	0.12			0.14			
Max	0.50			0.54			

It is also useful to point out again that merely deleting items that are troublesome may not be ideal. A researcher performing the analyses in Table 4 first (with small samples) would delete two items that showed fine replicability in Table 1 (larger samples), and would retain the one troublesome item. Thus, researchers should ensure they have large, generalizable samples prior to performing ANY exploratory factor analysis.

Conclusion

In the 21st century, exploratory factor analysis remains a commonly-used (and commonly misused) technique despite the more rigorous and useful confirmatory techniques that are widely available. While we do not assert that EFA with replication is a viable substitute for confirmatory factor analysis, there are times when EFA is appropriate or necessary, and we believe that replication should be a promiment part of any of these analyses. Without a reasonable likelihood of replicability, researchers have little reason to use a particular scale.

Although authors have been presenting methods for summarizing replication in EFA for half a century and more, most summarization techniques have been flawed and/or less informative than ideal. In the 21st century, with CFA invariance analysis as the gold standard for assessing

generalizability and replicability, replication within EFA has an important role to play—but a different role than half a century ago before CFA was widely available. Today, replication is a starting point, as is EFA. It adds value to EFA analyses in that it helps indicate the extent to which these models are likely to generalize to the next data set, and also in helping to further identify volatile or problematic items. This information is potentially helpful in the process of developing and validating an instrument, as well as for potential users of an instrument that has yet to undergo CFA invariance analysis.

We urge readers to take that brief additional step of performing and reporting replication results as a routine practice, and to further move forward (obviously, with new samples) to confirmatory factor analysis when the time is right to present the scale for broad usage within the research or practitioner community.

References

Baggaley, A. R. (1983). Deciding on the ratio of number of subjects to number of variables in factor analysis. *Multivariate Experimental Clinical Research, 6*(2), 81-85.

Barrett, P. (1986). Factor comparison: An examination of three methods. *Personality and Individual Differences, 7*(3), 327-340.

Barrett, P. T., & Kline, P. (1981). The observation to variable ratio in factor analysis. *Personality study and group behavior, 1*, 23-33.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality, 54*(1), 106-148. doi: 10.1111/j.1467-6494.1986.tb00391.x

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*: Lawrence Erlbaum.

Costello, A. B., & Osborne, J. W. (2005). Exploratory Factor Analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(7), 1-9.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological assessment, 7*(3), 286.

Ford, J. K., MacCallum, R. C., & Tait, M. (1986). The application of exploratory factor analysis in applied psychology: A critical review and analysis. *Personnel Psychology, 39*(2), 291-314.

Guadagnoli, E., & Velicer, W. F. (1988). relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265-275.

Henson, R. K., & Roberts, J. K. (2006). Use of exploratory factor analysis in published research. *Educational and Psychological Measurement, 66*(3), 393-416.

Jöreskog, K. G., & Sörbom, D. (1996). *LISREL 8 user's reference guide*: Scientific Software.

Kaiser, H., Hunka, S., & Bianchini, J. (1971). Relating factors between studies based upon different individuals. *Multivariate Behavioral Research, 6*(4), 409-422.

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.

McCrae, R. R., & Costa Jr, P. T. (1997). Personality trait structure as a human universal. *American psychologist, 52*(5), 509.

Osborne, J. W. (1997). Identification with Academics and Academic Success Among Community College Students. *Community College Review, 25*(1), 59-67.

Osborne, J. W. (2000). Prediction in Multiple Regression. *Practical Assessment, Research & Evaluation, 7*, n2.

Osborne, J. W. (2008). Creating valid prediction equations in multiple regression: Shrinkage, Double Cross-Validation, and Confidence Intervals around prediction. In J. w. Osborne (Ed.), *Best practices in quantitative methods.* (pp. 299-305). thousand oaks, CA: Sage Publishing.

Osborne, J. W., & Costello, A. B. (2004). Sample size and subject to item ratio in principal components analysis. *Practical Assessment, Research & Evaluation, 9*(11), 8.

Osborne, J. W., Costello, A. B., & Kellow, J. T. (2008). Best Practices in Exploratory Factor Analysis. In J. W. Osborne (Ed.), *Best Practices in Quantitative Methods* (pp. 205-213). Thousand Oaks, CA: Sage Publishing.

Osborne, J. W., & Jones, B. D. (2011). Identification with Academics and Motivation to Achieve in School: How the Structure of the Self Influences Academic Outcomes. *Educational Psychology Review*, 1-28.

Rosenberg, M. (1965). *Society and the adolescent self-image*. Princeton, NJ: Princeton University press.

Stevens, J. (2002). *Applied multivariate statistics for the social sciences*: Lawrence Erlbaum.

Tabachnick, B., & Fidell, L. (2001). Principal components and factor analysis. *Using multivariate statistics*, 582-652.

ten Berge, J. M. F. (1986). Rotation to perfect congruence and the cross validation of component weights across populations. *Multivariate Behavioral Research, 21*(1), 41-64.

ten Berge, J. M. F. (1996). The Kaiser, Hunka and Bianchini factor similarity coefficients: a cautionary note. *Multivariate Behavioral Research, 31*(1), 1-6.

Thompson, B. (1999). Five Methodology Errors in Educational Research: The Pantheon of Statistical Significance and Other Faux Pas. In B. thompson (Ed.), *Advances in Social Science Methodology* (Vol. 5, pp. 23-86). Stamford, CT: JAI Press.

Tucker, L. R. (1951). A method for synthesis of factor analysis studies: Educational Testing Service Princeton NJ.

Wrigley, C., & Neuhaus, J. O. (1955). The matching of two sets of factors. *American psychologist, 10*, 418-419.

Appendix 1

Items in the School Perceptions Questionnaire (SPQ) Scale

1. Being a good student is an important part of who I am.
2. I feel that the grades I get are an accurate reflection of my abilities.
3. My grades do not tell me anything about my academic potential.*
4. I don't really care what tests say about my intelligence.*
5. School is satisfying to me because it gives me a sense of accomplishment.
6. If the tests we take were fair, I would be doing much better in school.*
7. I am often relieved if I just pass a course.*
8. I often do my best work in school.
9. School is very boring for me, and I'm not learning what I feel is important.*
10. I put a great deal of myself into some things at school because they have special meaning or interest for me.
11. I enjoy school because it gives me a chance to learn many interesting things.
12. I feel like the things I do at school waste my time more than the things I do outside school.*
13. No test will ever change my opinion of how smart I am.

Note. All items measured on a scale of 1 to 5 (1=strongly disagree and 5 = Strongly Disagree). * Indicates that item is reverse coded.

Citation:

Osborne, Jason W. & David C. Fitzpatrick (2012). Replication Analysis in Exploratory Factor Analysis: What it is and why it makes your analysis better. *Practical Assessment, Research & Evaluation*, 17(15). Available online: http://pareonline.net/getvn.asp?v=17&n=15.

Corresponding Author:

Jason W. Osborne
Old Dominion University
Darden College of Education, Room 120
Norfolk, VA, 23529
919-244-3538
Jxosborn [at] odu.edu