Intelligence 54 (2016) 37–54

Contents lists available at ScienceDirect

Intelligence

Comparison of methods for factor extraction for cognitive test-like data: Which overfactor, which underfactor?

Timothy Z. Keith [a,*], Jacqueline M. Caemmerer [a], Matthew R. Reynolds [b]

[a] University of Texas, United States
[b] University of Kansas, United States

ARTICLE INFO

Article history:
Received 15 April 2015
Received in revised form 10 November 2015
Accepted 10 November 2015
Available online xxxx

Keywords:
Exploratory factor analysis
Confirmatory factor analysis
Parallel analysis
Minimum average partial
Number of factors
Test interpretation

ABSTRACT

Research published in *Intelligence* showed that the number of factors measured by individual intelligence tests has increased dramatically over time (Frazier & Youngstrom, 2007). When "gold standard" methods (parallel analysis based on principal components analysis, PA-PCA, and minimum average partial, MAP) were applied to these same tests fewer factors emerged, leading the authors to conclude that tests were inappropriately overfactored, and that modern tests are measuring far fewer underlying constructs than they are intended to measure. The article was influential, with the findings cited commonly and used to guide subsequent analyses. Here, we tested a key assumption of Frazier and Youngstrom and subsequent research that has used these methods: whether MAP and PA-PCA are accurate in recovering the correct number of factors with cognitive-test-like data. MAP and PA-PCA were compared to other exploratory and confirmatory factor analytic techniques in their ability to recover the correct number of factors in simulated data. Data conformed to common values found in intelligence literature, and varied based on the number of tests per factor, magnitude of factor loadings and factor correlations, and sample size. Results showed that MAP and PA-PCA, in fact, underfactored under many realistic data conditions, meaning that they recovered too few factors. Confirmatory methods were more accurate. Among exploratory methods PA based on principal axis factoring (not principal components analysis) was most accurate, although all methods underfactored with few tests per factor and high factor correlations. These findings suggest that PA-PCA and MAP are not "gold standard" methods for determining the number of factors underlying intelligence data and that other methods are more accurate. We argue for the importance of formal and informal theory in factor analytic investigations of intelligence tests.

Factor analysis is inexorably linked to the development of intelligence tests and to intelligence theory. Early intelligence theorists, such as Spearman and Thurstone, were also the developers of factor analysis. That tradition continued throughout the 20th century, with researchers such as Carroll, Cattell, and Horn using (and developing) factor analysis in the development of intelligence theories. Carroll's three-stratum theory, for example, was primarily based on his factor analyses of 460 data sets of intelligence test results (1993). Cattell and Horn's extended Gf–Gc theory was likewise based, in large part, on factor analytic results (1966). Even now, intelligence theory is developed and revised based on factor-analytic findings. For example, Johnson and Bouchard's VPR model of intelligence was developed, in part, through factor analysis of various test batteries (Johnson & Bouchard, 2005a, 2005b). CHC Theory, the hybrid of three-stratum and extended Gf–Gc theory, also continues to be developed and revised based, in part, on factor analytic findings (Keith & Reynolds, 2010; Schneider & McGrew, 2012). Theorists have often used factor analysis to understand the constructs underlying tests; test developers have used that understanding to develop new tests or revise existing ones; and then researchers have, in turn, subjected those measures to further factor analysis.

The practical aspects of this symbiotic relation between factor analysis, intelligence theories, and intelligence tests can be illustrated via the development of the Wechsler Scales. Early factor analyses suggested that the Wechsler scales measured a third factor beyond Verbal and Performance constructs (Cohen, 1952); the WISC-III included new subtests in an effort to beef up measurement of this third factor. Factor analyses of the WISC-III, however, suggested that the new version of the test measured four, rather than three, underlying constructs (e.g., Keith & Witta, 1997). The WISC-IV cited CHC theory—which was based, in part, on factor analytic evidence—as one reason for changes in that measure from the previous version (Wechsler, 2003). Confirmatory factor analyses of the WISC-IV and WAIS-IV suggested, in turn, that these tests are measuring five factors, a finding which is even more consistent with CHC theory (Keith, Fine, Taub, Reynolds, & Kranzler, 2006; Weiss, Keith, Zhu, & Chen, 2013a; see, however, Canivez & Kush, 2013). The most recent iteration of the Wechsler Scales (the WISC-V) also cited CHC theory as a guiding theory and included scales reflecting five underlying factors consistent with CHC

* Corresponding author at: The University of Texas at Austin, Department of Educational Psychology, 1912 Speedway D5800, Austin, TX 78712-1289, United States.
E-mail address: tzkeith@austin.utexas.edu (T.Z. Keith).

http://dx.doi.org/10.1016/j.intell.2015.11.003
0160-2896/

theory (Wechsler, 2014). Again, intelligence theories, intelligence tests, and factor analysis have been and continue to be naturally linked.

Given this link, evidence that factor analyses conducted with intelligence measures are incorrect or misguided would have important implications for intelligence theory and the development and use of intelligence tests. Recent research would seem to provide that evidence. In particular, a 2007 article questioned whether researchers and test publishers are "overfactoring" recent tests, meaning that they are identifying more factors through factor analysis than really exist in the data (Frazier & Youngstrom, 2007). Frazier and Youngstrom (hereafter abbreviated as FY) reanalyzed data from fifteen individually administered intelligence test batteries to show that when what they deemed "gold-standard criteria" (p. 169) were used, modern tests measure far fewer factors than they are purported to measure, a finding with important implications for the validity of the tests and for the validity of the theory or theories underlying those tests.

Two conclusions from the FY research had the potential to shape factor analytic and test development practice. The first conclusion was that modern cognitive tests really measure fewer factors than they are designed to measure: "The present results indicate that recent commercial tests of cognitive ability are not adequately measuring the number of factors they are purported to measure by test developers" (Frazier & Youngstrom, 2007, p. 180). The second conclusion was that confirmatory factor analysis (CFA) methods may be less useful for an uncovering factor structure than exploratory factor analysis (EFA) methods. "The results of this study do not suggest that CFA is not a useful approach to examining the structure of cognitive abilities" (Frazier & Youngstrom, 2007, p. 180). Although the last sentence is a little confusing, the surrounding text does not seem supportive of the use of CFA as a method to determine the correct number of factors underlying intelligence test data. Subsequent email communication with the first author confirms the article's (qualified) support of EFA over CFA: "…EFA is probably a more clinically useful approach to sub-scale development because the factors tend to be more reliable…," with the caveat "that is not to say that CFA would not be useful theoretically…" (T. W. Frazier, personal communication, July 17, 2015).

The Frazier and Youngstrom research could be seen as a simple critique of modern and traditional factor analytic practice, one that demonstrates different findings using different factor analytic techniques. It has, however, influenced modern practice. The article is commonly cited by those conducting research on the validity of cognitive tests, and is often cited as evidence that:

1. Modern cognitive tests measure fewer factors than they are designed to measure (Jia & Jia, 2009; Major, Johnson, & Bouchard, 2011; Watkins & Beaujean, 2014). Specifically, these researchers argue that often more cognitive factors are retained "than data merit" (Nelson, Canivez, Lindstrom, & Hatt, 2007, p. 443). Other researchers are in agreement with Frazier and Youngstrom's conclusion that many modern cognitive tests should be considered measures of a single factor, rather than multiple factors (Dombrowski, Watkins, & Brogan, 2009).
2. The methods of parallel analysis (PA) and minimum average partial (MAP) test provide more accurate estimates of the true number of factors than do other methods, including CFA (Hoelzle, Nelson, & Smith, 2011; Mays, Kamphaus, & Reynolds, 2009; Nelson & Canivez, 2012; Tucker et al., 2011).
3. CFA methods are less accurate in determining the number of factors compared to EFA methods, or that CFA is overly relied on in the literature and should be done in conjunction with EFA (Canivez & Watkins, 2010; Dombrowski & Watkins, 2013; Mays et al., 2009; Nelson et al., 2007).

1. Conclusion and assumption: the accuracy of MAP and PA

Given the conclusions reached in the FY research, and the citations to this research by others, it is worth examining more closely exactly what that research did. FY argued and cited evidence that the methods of MAP and PA provide accurate estimates of the true number of factors underlying data. When those methods were used to determine the number of factors to extract from test data the findings suggested fewer factors than recent tests are designed to measure. In other words, the argument of FY was that *if* the methods of MAP and PA are more accurate, *then* modern tests are overfactored. It is worth noting that the contention that MAP and PA provide "gold-standard" (p. 169) evidence of the correct number of factors was an assumption and thus was never tested; rather, the authors cited two studies (Zwick & Velicer, 1982, 1986) and a review (Velicer, Eaton, & Fava, 2000) as evidence for their new gold standard.

1.1. Simulation research

If the issue of how many factors to extract in factor analysis had been settled objectively, this discovery would indeed be a major leap forward in factor analysis, but the issue is not nearly as settled as suggested in FY. PA (based on principal components analysis, or PA-PCA) is indeed often recommended by methodologists, and has been shown to be accurate in much simulation research (Mulaik, 2009, chap 8). Simulation research has also shown that this method may underfactor (return too few factors) under certain conditions (Beauducel, 2001b; Turner, 1998), however, and those conditions are often the exact conditions that are common to intelligence tests. According to Mulaik "my impression of these studies is that they have tended to use data sets…that have large loadings, uncorrelated factors, and relatively few common factors relative to the number of variables, and small number of variables" (2009; p. 189). Intelligence data, in contrast, often have high correlations among factors, more common factors, and some moderate-level loadings, quite different from the conditions in many simulation studies that have supported PA-PCA. Simulation research suggests that MAP may be even more likely to underestimate the number of factors (components) (Ruscio & Roche, 2012; Zwick & Velicer, 1986). Other methods or variations (beyond PA-PCA and MAP) have been better-supported with discrete or polytomous variables (Barendse, Oort, & Timmerman, 2014; Timmerman & Lorenzo-Seva, 2011).

Simulation research suggests a number of conditions that may lead to underfactoring in PA or MAP, or both. First, the number of tests per factor is related to the performance of both PA and MAP, with these and other methods tending to underfactor when there are fewer tests per factor (Crawford et al., 2010; Zwick & Velicer, 1986). Few simulation studies have included as few as two tests per factor in their simulations, however. Such conditions are recommended against by methodologists (perhaps more so for items as opposed to subtests/parcels), but are common in intelligence research because many tests include only two measures of some of the underlying constructs. Second, as noted by Mulaik (2009), the correlation among factors is also an important influence on the accuracy of PA-PCA. The higher the correlation among factors, the more likely PA-PCA is to return fewer factors than actually exist in the data (Crawford et al., 2010; Green, Levy, Thompson, Lu, & Lo, 2011). Third, the level of factor loadings may influence accuracy, with lower loadings resulting in less accurate findings (Crawford et al., 2010; Green et al., 2011). Last, sample size also appears to be an important influence on the number of factors suggested by various methods (Beauducel, 2001a; Crawford et al., 2010; Green et al., 2011). Variations among factor loadings (e.g., a mixture of higher and lower loadings) appear to have little effect on accuracy, however (Crawford et al., 2010).

In addition to data characteristics, methodological variations in PA have also been investigated. Both procedures recommended by FY were based on principal components analysis (PCA) rather than true factor analysis. PA based on principal axis factoring (PA-PAF)—a factor analytic method—is generally more accurate than PA based on PCA, and less likely to underfactor when factor correlations are high (Crawford et al., 2010; Green et al., 2011). With ordered polytomous variables, PA based on minimum-rank factor analysis has been supported