Available online at www.sciencedirect.com
ScienceDirect
Intelligence 35 (2007) 169–182
INTELLIGENCE

Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring?

Thomas W. Frazier [a,b,*], Eric A. Youngstrom [c,d]

[a] Department of Psychology, John Caroll University, 20700 North Park Blvd., University Heights, OH 44118, USA
[b] Department of Psychiatry and Psychology, Cleveland Clinic Foundation, USA
[c] Department of Psychology, Case Western Reserve University, USA
[d] Department of Psychiatry, University Hospitals of Cleveland, USA

Received 5 May 2006; received in revised form 26 June 2006; accepted 6 July 2006
Available online 9 August 2006

Abstract

A historical increase in the number of factors purportedly measured by commercial tests of cognitive ability may result from four distinct pressures including: increasingly complex models of intelligence, test publishers' desires to provide clinically useful assessment instruments with greater interpretive value, test publishers' desires to include minor factors that may be of interest to researchers (but are not clinically useful), and liberal statistical criteria for determining the factor structure of tests. The present study examined the number of factors measured by several historically relevant and currently employed commercial tests of cognitive abilities using statistical criteria derived from principal components analyses, and exploratory and confirmatory factor analyses. Two infrequently used statistical criteria, that have been shown to accurately recover the number of factors in a data set, Horn's parallel analysis (HPA) and Minimum Average Partial (MAP) analysis, served as gold-standard criteria. As expected, there were significant increases over time in the number of factors purportedly measured by cognitive ability tests ($r = .56$, $p = .030$). Results also indicated significant recent increases in the overfactoring of cognitive ability tests. Developers of future cognitive assessment batteries may wish to increase the lengths of the batteries in order to more adequately measure additional factors. Alternatively, clinicians interested in briefer assessment strategies may benefit from short batteries that reliably assess general intellectual ability.
© 2006 Elsevier Inc. All rights reserved.

Keywords: Cognitive ability test; Factor analysis

Commercial tests of cognitive abilities have become increasingly complex. The Wechsler Adult Intelligence Scales (WAIS, WAIS-R, and WAIS-III) and the Wechsler Child Intelligence Scales (WISC, WISC-R, WISC-III, and WISC-IV) are good examples. These scales have moved from purportedly measuring two aspects of intelligence, verbal and performance abilities (WISC; Wechsler, 1949; WAIS; Wechsler, 1955; WISC-R; Wechsler, 1974; WAIS-R; Wechsler, 1981), to measuring four cognitive abilities: Verbal Comprehension, Perceptual Organization or Perceptual Reasoning, Freedom from Distractibility or Working Memory, and Processing Speed (WAIS-III; Wechsler, 1997b; WISC-IV; Wechsler, 2003). Unfortunately, the added complexity of commercial ability tests has not resolved controversies regarding the structure of these batteries. The Wechsler adult intelligence scales are a good example.

Factor analytic studies of the revised Wechsler Adult Intelligence Scale (WAIS-R) disagreed about the true

* Corresponding author. Department of Psychology, John Caroll University, 20700 North Park Blvd., University Heights, OH 44118, USA.
E-mail address: tfrazier@jcu.edu (T.W. Frazier).

0160-2896/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2006.07.002

structure of the test with some authors espousing one factor solutions (O'Grady, 1983), others two factors (Silverstein, 1982), and still others three factors (Naglieri & Kaufman, 1983; Parker, 1983). Based upon findings from the latter studies, authors of the most recent revision, the WAIS-III, decided to add three subtests in an attempt to clarify potential second, third, and fourth factors (Wechsler, 1997a,b). According to the test authors, exploratory and confirmatory factor analyses of the WAIS-III standardization data support the proposed four-factor model. However, a recent confirmatory factor analytic study of the WAIS-III using similar methods to those employed by the authors of the test, has suggested the presence of fewer than four factors (Ward, Ryan, & Axelrod, 2000).

Controversies regarding factor structure have not been limited to the Wechsler scales. Several studies have debated the structure of the 15-subtest Stanford–Binet Fourth-Edition (SB-IV). Some research has supported the four factors proposed by test authors with occasional minor alterations (Boyle, 1989). Other studies have suggested that two and three factor solutions more parsimoniously represent the structure of the test (Gridley & McIntosh, 1991; Kline, 1989). The title of Thorndike's (1990) article investigating the structure of the SB-IV "Would the real factors of the Stanford–Binet Fourth-Edition please come forward?" best captures the existing confusion regarding the true factor structure of this instrument. At present, consensus has not been reached for either the SB-IV or the WAIS-III.

Clearly, the increasing complexity of cognitive assessment batteries has not resolved controversy regarding the structure of these tests. Yet large, elaborate test batteries, such as the new Woodcock–Johnson Psycho-Educational Battery Third-Edition (WJ-III) with 20 ability and 23 achievement subtests, continue to be marketed to clinicians. This observation leaves open the question of what is driving the movement toward longer, factorially complex cognitive ability batteries. The present paper proposes that several forces have influenced this trend including: increasingly complex theories of intelligence (Carroll, 1993; Vernon, 1950), commercial test publishers' desire to provide assessment instruments with greater interpretive value to clinicians, publishers' desire to include minor ability factors that may only be of interest to researchers, and heavy reliance on liberal statistical criteria for determining the number of factors measured by a test. The latter hypothesis is evaluated empirically in the present study by comparing several statistical criteria for determining the number of factors present in current and historically relevant cognitive ability batteries.

1. Theories of intelligence

Spearman (1904) developed one of the first theories of the structure of intelligence. He proposed that intelligence can best be described by a general ability factor and specific factors represented by each subtest used to measure intellectual ability. A number of other researchers, and later Spearman himself, thought that this model was too simple (for a detailed review see Carroll, 1993). As a result, more complex models of intelligence were developed. Some of these models include a hierarchical structure with g, general ability at the top of the hierarchy, primary or intermediary factors in the middle stages of the hierarchy, and specific factors attributed to each subtest at the bottom of the hierarchy (Cattell, 1963).

Vernon (1950) detailed what is probably the first hierarchical theory of intelligence. His work proposed two correlated primary abilities comprised of verbal/educational and spatial/mechanical abilities. Around the same time Cattell (Cattell, 1963) posited a two-factor theory of intelligence with the two factors representing crystallized (Gc) and fluid (Gf) intellectual abilities (see also work by Horn, 1968). More recent theories, some of which are elaborations on Cattell's Gf/Gc theory, have suggested the presence of larger numbers of primary abilities. Carroll's (1993) three stratum theory is an example of one popular and extensively researched extension of Gf/Gc theory. His theory proposes a general ability factor at the top of the hierarchy, 8 broad ability factors in the middle of the hierarchy, and approximately 70 narrow abilities at the bottom of the hierarchy.

Other models of intellectual ability have proposed factors that are intended to be relatively uncorrelated with each other, and thus a general ability factor was not assumed (Guilford, 1967; Thurstone, 1938). These models have often been called into question based upon the finding that typically the measures of distinct abilities in these models show sizeable correlations amongst themselves and with standard IQ tests, implying a general cognitive ability factor. However, the lack of empirical support for uncorrelated intellectual abilities has generally not diminished the popularity of multiple intelligence theories. Gardner's (1983) theory of multiple intelligences is probably the most widely known. His theory and others like it have become popular presumably because they offer the "politically correct" possibility that an individual who would not be viewed as intelligent based upon more traditional models of intelligence could be seen as possessing other non-traditional intelligences. The popularity of theories espousing a number of separate, minimally correlated or uncorrelated, cognitive abilities may suggest that commercial tests have become more complex

due to a desire to measure these additional abilities. Assessing diverse abilities could theoretically provide a richer evaluation of the individual, allowing for more detailed and helpful clinical recommendations. In sum, the large number of increasingly complex theoretical models of the structure of intelligence may have led test developers to try to measure the cognitive abilities described by these models.

2. Commercial pressure on test development

Psychological and psycho-educational assessments have become a big business with psychologists playing an expanding role in diagnosis and treatment planning. As psychological assessments have become increasingly profitable endeavors, so have developing and marketing psychological tests. This is best exemplified by the numerous psychological test publications in existence from The Psychological Corporation, Western Psychological Services, and other publishers. Cognitive ability batteries, along with personality measures, play a large role in most psychological assessments and have been consistently found to be among the most frequently used psychological tests (Lees-Haley, Smith, Williams, & Dunn, 1996). The extensive use of cognitive ability batteries in psychological assessment, an increased market for psychological assessments in general, a desire to create tests that are marketable to both clinicians and researchers, and the desire to increase the reliability of IQ measures may create a pressure on publishers to market ability tests that measure everything that other tests measure and more. This, in turn, forces other ability test publishers to try to keep pace. If publishers do not market their instruments as offering more information than previous versions or other competing instruments, the instruments will likely be viewed by clinicians and researchers as outdated or insufficient. More specifically, if test batteries do not include subtests, or groups of subtests, that clinicians feel improve their ability to make decisions and provide recommendations, they are unlikely to be successfully marketed to clinicians. Similarly, there may be a pressure on test developers to include minor, poorly measured, additional factors into new test revisions in order to market them to researchers who are interested in examining multiple distinct of abilities at the group level, even if these additional factors are not clinically useful.

Commercial pressure to market tests with additional subtests may exist even in the absence of actual research to substantiate the validity of additional subtests or studies indicating that inclusion of these subtests allows for the reliable measurement of additional factors. In fact, very few studies have supported the reliability and validity of subtest or profile interpretation (Glutting, Watkins, & Youngstrom, 2003). Glutting and colleagues (2003) reviewed several failed attempts to develop subtest profiles with predictive validity using the Wechsler Intelligence Scales for Children-Third-Edition (WISC-III). The conclusion of this review was that, in general, subtest profiles provided very little incremental validity beyond estimates of general ability, and that subtest profiles were poor predictors of various diagnoses. These findings indicate that measurement of additional factors may not be clinically useful. Rather, inclusion of additional factors may simply increase the cost to psychologists, who buy newer assessment batteries to keep up with the standard of practice. Additionally, longer, more elaborate batteries are likely to increase the cost to society when insurance companies and individuals are charged for longer psychological assessments.

3. Statistical criteria for determining the number of factors

Empirical studies evaluating the structure of intelligence and the structure of commercial measures of cognitive ability have focused on three basic types of factor analytic techniques (Carroll, 1993). These include principal components analysis (PCA), exploratory factor analysis (EFA), and confirmatory factor analysis (CFA). PCA and EFA are conceptually similar and share methods for determining the number of factors/components to retain, therefore these two methods will be discussed jointly.

PCA and EFA have been used extensively to examine the structure of human cognitive abilities and the number of factors measured by commercial ability tests. Both techniques summarize the relationships between sets of variables, letting the data drive the analysis. The principal components obtained from PCA are simply linear combinations of the original measured variables. Thus, PCA has been described as being strictly a data reduction technique, with no distinction made between common and unique sources of variance in the measured variables. EFA, on the other hand, is thought by some to be a more appropriate technique for identifying latent constructs (Fabrigar, Wegener, MacCallum, & Strahan, 1999; Gorsuch, 1983; Widaman, 1993), since EFA methods parse unique and common sources of variance. Methodologists have debated the relative utility of PCA and EFA, with some favoring PCA (Velicer & Jackson, 1990) and others EFA (Widaman, 1993). Although both may possess advantages in some situations, they frequently produce similar results (Fabrigar et al., 1999). For this reason, the terms factor and component will be

used interchangeably, although in the strictest sense they refer to different analyses.

In both PCA and EFA there are two basic analytic decisions that a researcher must make, 1) determining the number of factors to retain and 2) choosing a method for rotating the retained factors to simple structure. Deciding on the appropriate number of factors to retain is probably the most important decision made in EFA and PCA since simple structure cannot be achieved if too few or too many factors are subjected to rotation. It is this decision, and the implications of this decision for cognitive ability test development and interpretation, that is the focus of the empirical analyses of the present paper.

3.1. PCA/EFA decision rules

There are a number of different decision rules available for determining the number of factors to retain. These rules are derived from PCA or EFA analyses and often result in retention of different numbers of factors. The most commonly employed rules or heuristics include the Kaiser criterion, Cattell's scree test, the chi-square statistic resulting from maximum likelihood factor analysis, and interpretability (for a detailed description of these techniques and procedures see Kim & Mueller, 1978; Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1986). For the purposes of the present paper a brief discussion of the advantages and disadvantages of these techniques is presented, followed by empirical work comparing these decision rules, and finally methodologists' recommendations regarding use of these techniques.

The Kaiser criterion is probably the most commonly used method for determining the number of components to retain. The justification and description of this method was detailed by Kaiser (1960) and simply involves retaining components obtained from PCA with eigenvalues greater than 1.0. While some view this criterion as simply establishing a lower bound for the number of true factors in a data set (Kim & Mueller, 1978), it has been most frequently used as a criterion for determining the number of factors. When used in the latter fashion, the Kaiser criterion is advantageous in that it can be easily and objectively applied. However, the rule has been found to perform poorly in simulation studies, with occasional underfactoring and more frequently overfactoring (for a discussion of this issue see Fabrigar et al., 1999).

Cattell's (1966) scree test is almost as frequently implemented as the Kaiser criterion. This test involves plotting all of the eigenvalues obtained from PCA and drawing a line through the smallest values. Those eigenvalues judged to be above this line are retained. Cattell's scree test has been criticized by some individuals as being too subjective. Supporting this position, one published study reported low interrater reliability (Crawford & Koopman, 1979), while other studies have found good performance when strong common factors are present (Cattell & Vogelmann, 1977; Hakstian, Rogers, & Cattell, 1982). Rater disagreement typically occurs in one of three situations: when the plot of all eigenvalues results in a gradual slope with no obvious breaking point in the line, there is more than one break point in the line, or more than one line can be drawn through the smallest eigenvalues (Zwick & Velicer, 1986).

The chi-square statistic resulting from maximum likelihood EFA is logically similar to Cattell's scree test in that the focus of the analysis is the equality of eigenvalues leftover after interpretable components are excluded. This technique comprises retaining successive factors until the chi-square statistic becomes non-significant. At this point the null hypothesis of equality of the remaining eigenvalues is no longer rejected. While this method is advantageous in that a definite stopping point can be easily identified, it has been criticized for leading to retention of too many factors. Overfactoring is particularly problematic when sample sizes are large (Fabrigar et al., 1999).

Interpretability generally refers to retaining only those components that make sense to the researcher (Fabrigar et al., 1999; Gorsuch, 1983). While this may mean use of several of the above criteria in combination with subjective judgement, it may also refer to subjective judgment alone. Interpretability has been frequently employed in factor analytic studies of cognitive ability tests, where authors sometimes refer to an examination of the "meaningfulness" of rotations of varying numbers of factors (for an example see Wechsler, 1991). Interpretability is fraught with difficulties when used in determining the number of cognitive ability factors. First, this method is extremely subjective and what is interpretable to some may not appear interpretable to others. Secondly, subjectivity makes it likely that confirmation bias will affect the judgment of researchers employing this method. Investigators may be willing to halt further analyses when their expectations have been met, even though the true structure of the data has not been approximated. Lastly, in determining the structure of cognitive ability tests, vast arrays of intelligence theories are available to interpret the results of exploratory analyses. This leads to an undesirable situation where almost any solution can be interpreted according to one or more of the existing theories. For these reasons, and the inherent difficulty in

operationalizing this criterion it was not evaluated in the present study.

Two other decision rules exist for determining the number of components or factors to retain: minimum average partial (MAP) analysis described by Velicer (1976) and Horn's (1965) Parallel analysis (HPA). These rules have been used infrequently in factor analytic literature. This is probably due to a general lack of knowledge regarding these techniques outside of the methodological community, the fact that they are not included in commonly used statistical programs, and the laborious computations needed to perform these analyses. MAP is an iterative procedure in which successive partial correlation matrices are examined. Initially, the average squared correlation of the observed correlation matrix is computed. Then, successive components resulting from PCA are partialed from the original matrix beginning with the first component. At each step, the average squared partial correlation is computed and the step at which the minimum average squared partial correlation is observed represents the number of components to retain. For example, if after partialing out the first and second components the minimum average squared partial correlation is reached, two factors are retained (see Velicer, 1976 for a description of the MAP procedure).

HPA was originally described as an adaptation of the Kaiser criterion (Horn, 1965). This procedure was developed to account for the fact that, for a given sample, principal components analysis of randomly generated, uncorrelated variables will result in initial eigenvalues exceeding 1.0 and later eigenvalues drifting below 1.0. Deviation of the eigenvalues from 1.0 is a result of sampling error of the observed correlations. In these situations implementations of the Kaiser criterion will lead to one or more factors being retained even though none of the variables are truly correlated. HPA corrects for this by comparing the eigenvalues obtained from PCA of the observed correlation matrix to eigenvalues obtained from PCA of a randomly generated correlation matrix. The randomly generated matrix is based upon the same sample size and number of variables as the observed correlation matrix. Components derived from the observed (real data) matrix are only retained if the eigenvalues for those components are larger than the eigenvalues derived from the randomly generated matrix.

Glorfeld (1995) noted that the above methodology was generally accurate but occasionally resulted in retention of more components than warranted. As a result of this observation, he further refined and improved HPA by generating several sets of eigenvalues from random correlation matrices, instead of only one set. From the multiple sets of random generated eigenvalues, 95% confidence intervals are constructed. Only components from the actual data that have eigenvalues that are larger than the upper bound of the 95% confidence interval of randomly generated eigenvalues are retained. This addition to HPA ensures that poorly defined components are not included.

3.2. Comparison of available decision rules

As a result of the multitude of decision rules available and an apparent lack of consensus over which techniques are most appropriate, factor analysis has been viewed by some researchers and statisticians as more subjective than other statistical techniques (Tabachnick & Fidell, 1996). However, Monte Carlo simulations involving data with a known structure have found MAP and HPA to more accurately recover the true number of existing factors (Zwick & Velicer, 1982, 1986). More conventional decision rules were found to inconsistently recover the true number of factors (Cattell's scree test) or led to overfactoring (the Kaiser criterion and the chi-square statistic). Results of these methodological studies led Velicer, Eaton, and Fava (2000) to recommend the use of HPA and MAP for determining the number of components. The other decision rules were not recommended as stand alone techniques. Unfortunately, results of studies examining these criteria have done little to influence current factor analytic practice. In fact, to the authors' knowledge neither MAP nor HPA has been used in the development of currently or previously available cognitive ability tests. Rather, more conventional criteria that may lead to over-extraction have been employed, supporting the notion that ability tests are being overfactored.

3.3. CFA decision rules

The advent of CFA and the apparent preference of this technique over EFA/PCA have led authors of recent commercial tests of cognitive ability to rely heavily on these methods (McGrew & Woodcock, 2001; Thorndike, Hagen, & Sattler, 1986; Wechsler, 1991, 1997b). CFA methods often utilize maximum likelihood estimation, thus unrestricted CFA models are mathematically identical to EFA using maximum likelihood estimation. CFA differs conceptually from EFA and PCA in that researchers using confirmatory methods specify the number of factors prior to the analysis. Ideally, a series of CFA models differing in factor complexity are specified and evaluated in order to determine the fit of the models, and consequently, the number of factors measured by the

data. Several indices are typically used to determine the fit of CFA models, including the chi-square statistic resulting from maximum likelihood estimation techniques and other statistics based upon chi-square. Two statistics which are frequently used in CFA were examined in the present study, the Comparative Fit Index (CFI; Bentler, 1988) and the Tucker–Lewis Index (TLI; Tucker & Lewis, 1973). Both indices are adjustments on chi-square for the degrees of freedom of the model and previous empirical work has suggested that values of CFI and TLI $\geq .95$ indicate good to excellent model fit (Kline, 1998).

3.4. EFA/PCA vs. CFA

We are unaware of any published empirical (Monte Carlo) investigations comparing CFA fit statistics in the recovery of the true number of factors in a data set. However, in previous studies comparing EFA and PCA decision rules, Bartlett's chi-square statistic frequently overfactored (Zwick & Velicer, 1986). This was particularly true when large sample sizes and conventional alpha levels ($p < .05$) were examined, as they frequently are in factor analytic studies of cognitive ability tests. Although Barlett's chi-square is not the same computationally as the chi-square resulting from maximum likelihood CFA, it is logically identical to these tests. This suggests that heavy emphasis on CFA and the chi-square statistic over recent years may have resulted in increased overfactoring of more current ability tests.

Other empirical work has provided evidence that CFA may be a less desirable technique for determining the number of factors measured by a data set. Many applications of CFA techniques involve specification searches, processes by which an initial model is modified in order to improve its fit. MacCallum and colleagues (MacCallum, 1986; MacCallum, Roznowski, & Necowitz, 1992) found that specification searches in covariance structure modeling often do not uncover the correct population model. This appears to be especially true in situations where the researcher's initial model is not close to the true model, when the search is terminated upon finding the first statistically plausible model, when small samples are used, and when the investigator cannot place valid restrictions on possible modifications (MacCallum, 1986). Based upon these findings, MacCallum and colleagues recommended using alternative a priori models. However, this approach may be difficult to implement when examining the structure of cognitive ability tests due to the large number of intelligence theories from which to generate alternative models.

4. The present study

For the previously discussed reasons, CFA methods and more conventional EFA/PCA based decision rules may be inappropriate for determining the number of factors measured by ability tests, leading to retention of too many factors. Yet, these methods have been used exclusively to determine the structure of ability tests, suggesting that recent ability tests may be overfactored. The purpose of the present study is to empirically evaluate the number of factors measured by current and historically relevant cognitive ability tests using recommended criteria (MAP and HPA), more conventional EFA/PCA decision rules, and CFA statistics with unknown accuracy. In particular, the present study will examine whether the observation of historical increases in the length and complexity of commercially available cognitive ability tests is statistically reliable, if recent increases in the overfactoring of ability tests have occurred, and (if overfactoring is observed) whether the use of liberal statistical criteria has influenced recent increases in overfactoring. The latter hypothesis will be evaluated in three ways. First, a qualitative examination of the statistical criteria used by recent ability tests will be performed. Secondly, the number of factors retained by HPA and MAP will be compared to the number of factors retained by more conventional criteria to determine whether these criteria overfactor. Finally, analyses will examine whether the number of factors retained using commonly implemented decision rules approximates the number of factors proposed by test authors.

5. Method

5.1. Commercial tests

Historically relevant and recent commercial tests of cognitive ability examined in this study were the: Wechsler Intelligence Scale for Children-Original Version, Revised, Third-Edition, and Fourth-Edition (WISC; Wechsler, 1949; WISC-R; Wechsler, 1974; WISC-III; Wechsler, 1991; WISC-IV; Wechsler, 2003), Differential Ability Scales (DAS; Elliott, 1990), Stanford–Binet Fourth-Edition (SB-IV, 15 subtests; Thorndike et al., 1986), Kaufman Assessment Battery for Children (K-ABC, 8 subtests, ages 7–12; Kaufman & Kaufman, 1983), Wechsler Primary and Preschool Scale of Intelligence-Original Version, Revised, and Third-Edition (WPPSI; Wechsler, 1967; WPSSI-R; Wechsler, 1989; WPPSI-III; Wechsler, 2002), Woodcock Johnson-Revised and Third Edition (20 subtests from the cognitive batteries, WJ-R; Woodcock & Johnson,

1989; WJ-III; Woodcock, McGrew, & Mather, 2001), and Wechsler Adult Intelligence Scale-Original Version, Revised, and Third-Edition (Wechsler, 1955, 1981, 1997b). Only 20 subtests from the WJ-R cognitive assessment battery could be analyzed due to missing correlations for one subtest in the subtest intercorrelation matrix of the test manual. Correlation matrices were based upon all individuals reported in the standardization samples for each of the batteries examined except for the SB-IV, where pairwise correlations reported by Boyle (1989) were used to fill in the matrix reported in Table 6 of the test manual[1] (Thorndike et al., 1986). In situations where the test manual did not report an average correlation matrix across groups (K-ABC, WJ-R and WJ-III), correlation matrices were averaged using Fisher's *z*-transformation to produce an average inter-subtest correlation matrix. This procedure used only age groups receiving all of the subtests. For the WJ-R and WJ-III, the minimum number of subjects administered any subtest was used to estimate the *N* for the averaged matrix (WJ-R, $N=499$; WJ-III, $N=2034$). The correlation matrices reported in the WISC and WAIS manuals were averaged in order to obtain one inter-subtest correlation matrix based upon all individuals reported in the component matrices. Previous versions of the Stanford–Binet Intelligence Scales and the Woodcock–Johnson cognitive assessment battery could not be analyzed because correlation matrices for these scales were not available.[2]

5.2. Statistical tests and decision rules

For each ability test, PCA, maximum likelihood EFA, and unrestricted one factor model CFA's were performed. Results of these analyses were used to determine and evaluate several decision rules. Kaiser's criterion, retaining any components with eigenvalues greater than 1.0, was determined using the results of PCA analyses performed in SPSS 11.0 (2002). Cattell's scree test was implemented by performing Principal Axis Factoring to extract eigenvalues resulting from squared multiple correlations in the input matrix using SPSS 11.0 (SPSS, 2002). Then, eigenvalues were plotted using Microsoft Excel. Four raters naïve to the purpose of the present study rated each graph. They were instructed to find the break in the line resulting from the lowest values and then count the number of data points falling above the line. Inter-rater reliability was computed using model two of the intraclass correlation coefficient which accounts for mean differences between raters (Shrout & Fleiss, 1979). Average inter-rater reliability was adequate (ICC (2,4)=.75; the second number indicates the number of raters averaged), but individual ratings were quite unreliable (ICC (2,1)=.43), indicating that an average rating was required for further analyses. Inspection of the ratings revealed particular difficulties with three measures, the Woodcock–Johnson-III, the Differential Ability Scales, and the Stanford–Binet IV. For these measures, ratings differed by as much as 3 (1 vs. 4 factors estimated). This appeared to be due to the odd shape of the plot of eigenvalues (multiple breaks in the line or gradual slope) for both of these measures and further reinforces the difficulty with implementing Cattell's scree test for measures without strong common factors.

HPA was computed using Watkins (2000) software. For each HPA analysis, 95% confidence intervals for the mean of 100 sets of randomly generated eigenvalues were created, based upon Glorfeld's (1995) extension on Horn's (1965) original analysis. Any components resulting from PCA that were larger than the upper bound of these confidence intervals were retained. The chi-square rule was derived by performing a series of maximum likelihood EFA's with successively larger numbers of factors being retained in each set of analyses. The point at which the chi-square value resulting from these analyses became non-significant was taken as the number of factors. MAP analyses were performed as described earlier using SPSS 11.0 macro language (O'Connor, 2000).

Computation of CFI and TLI occurred in stages. First one-factor unrestricted maximum likelihood CFA's were performed for each test using AMOS (Arbuckle, 1999). This was done to obtain the value of chi-square and the degrees of freedom for the independence model. Then, increasingly factorially complex, unrestricted models were performed using maximum likelihood EFA in SPSS 11.0 (2002). This was done to obtain the chi-square and degrees of freedom for these models. The chi-square values obtained for the independence model and the chi-squares values obtained from the various unrestricted models were used to compute CFI and TLI. The points at which CFI and TLI became greater than .95 were taken as the number of factors retained for these criteria, based upon previous research suggesting that

[1] Intersubtest correlations for the SB-IV were obtained from Table 6.1 of the manual. Missing correlations were derived from values obtained from other reported matrices. The median correlations analyzed were based upon different numbers of subjects so that some correlations were based upon less than 100 people while others were based upon 5000 people.

[2] Copies of the matrices examined in the present study can be obtained from the first author.

values in this range indicate good to excellent fit (Kline, 1998; Tabachnick & Fidell, 1996).

6. Results

Table 1 presents test publication date, test length, number of factors purportedly measured, and the number of factors indicated by each statistical criterion for all of the commercial tests of cognitive ability. As expected, a moderate, although non-significant, linear relationship between test publication date and test length, $R^2=.11$, $r(14)=.33$, $p=.236$, was observed (see Fig. 1, panel A). The linear relationship between publication date and the number of factors purportedly measured by these tests was large in magnitude and significant, linear $R^2=.31$, $r=.56$, $F(1,13)=5.91$, $p=.030$. A significant quadratic trend was also observed, quadratic $R^2=.51$, $\Delta R^2=.20$, $r=.45$, $F(1,12)=4.88$, $p=.047$ (see Fig. 1; panel B). The difference between the linear relationships between publication date and test length and publication date and the number of factors purported approached significance in spite of the small number of observations, $t(12)=1.81$, $p=.096$, suggesting that the number of factors purportedly assessed by cognitive tests has grown more quickly than the length of these tests (Cohen & Cohen, 1983). It should be noted that removal of the two measures with extremely high values for purported number of factors (WJ-R and WJ-III, both 7 factors proposed) decreases the magnitude of the quadratic effect described above, quadratic $R^2=.09$, $r=.30$, $F(1,10)=1.89$, $p=.20$, but the linear effect remains large and significant, $R^2=.40$, $r=.63$, $F(1,11)=8.98$, $p=.012$. Additionally, the ratio of test length to factors proposed (no. of subtests/no. of factors proposed) was significantly and strongly negatively correlated with test publication date, $R^2=.46$, $r(14)=-.68$, $p=.005$. After 1983, 3 of the 8 measures did not even include enough subtests to adequately identify each factor (3 subtests per factor; Fabrigar et al., 1999).

Based upon previous research indicating that HPA and MAP accurately recover the true number of factors in a data set (Zwick & Velicer, 1986), analyses comparing the number of factors determined by HPA and MAP and the number of factors purportedly measured by cognitive ability tests were performed. These analyses were used to evaluate the prediction that commercial cognitive ability tests have been overfactored. The number of factors retained using HPA and MAP was significantly less than the purported number of factors measured by the tests, $t(14)=4.08$, $p=.001$ and $t(14)=4.01$, $p=.001$ respectively. On average, the number of factors indicated by MAP and HPA did not differ, $t(14)=1.29$, $p>.20$; however these two methods showed poor agreement, ICC (2,2)=.21 (Shrout & Fleiss, 1979).

To examine whether there has been a historical increase in the overfactoring of ability tests, difference scores were computed by subtracting the number of factors indicated by HPA and MAP from the number of factors purportedly measured by the tests. Consistent with prediction, a marginally significant relationship

Table 1
Test publication year, test length, number of factors measured, and the number of factors retained using each statistical criterion for each commercial test of cognitive ability

Test	Publication year	Subtests	Factors proposed	HPA	MAP	Scree (avg.)	Kaiser	$\times^2$	CFI	TLI
WISC	1949	12	2	2	2	2	2	3	2	2
WAIS	1955	11	2	1	1	2	1	4	2	2
WPPSI	1967	11	2	2	1	2	2	3	2	2
WISC-R	1974	12	2	2	1	3	2	6	2	2
WAIS-R	1981	11	2	1	2	1.25	1	6	2	2
K-ABC	1983	8	2	2	1	2	2	3	2	2
SB-IV	1986	15	4	1	1	1.75	2	10 *	3	4
WJ-R	1989	20	7	3	1	2.75	5	11	5	9
WPPSI-R	1989	12	2	2	2	2	2	4	2	2
DAS	1990	6	3	1	1	1.25	1	3 *	2	3 *
WISC-III	1991	13	4	2	2	3.25	3	5	3	3
WAIS-III	1997	13	4	1	2	2.50	2	7	3	4
WJ-III	2001	20	7	3	2	2.50	3	10	4	5
WPPSI-III	2002	12	3	2	2	2.25	2	6	2	2
WISC-IV	2003	15	4	2	2	2	2	8	3	3
Mean(S.D.)	1983.8(16.4)	12.7(3.8)	3.3(1.7)	1.8(0.7)	1.5(0.5)	2.17(.56)	2.1(1.0)	5.9(2.8)	2.6(0.9)	3.1(1.9)

* For these values the degrees of freedom were used up ($df<1$) before the criteria was satisfied so the values displayed are for one greater than the number of factors retained prior to using up degrees of freedom. The scree test estimates are based on PAF extraction, and are the average of ratings from four independent judges naïve to the purported number of test factors.

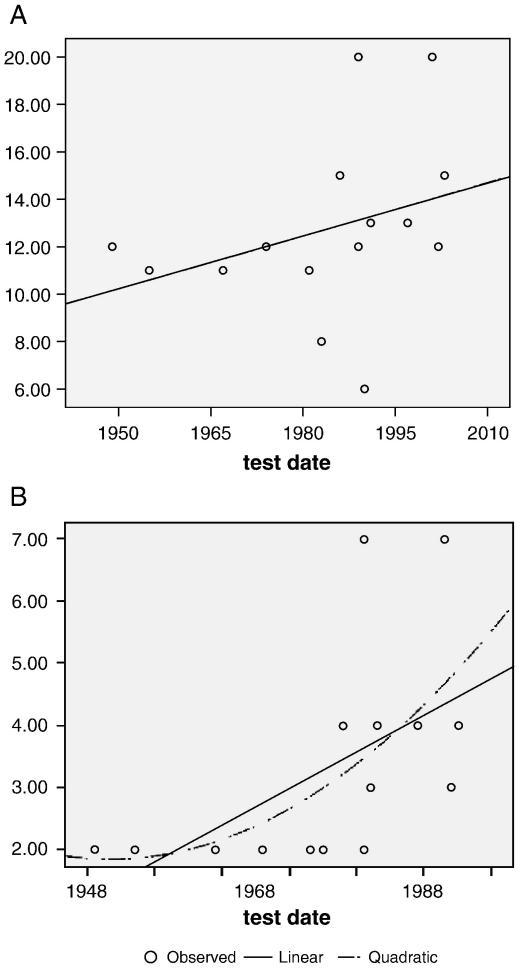

Fig. 1. Panel A presents the linear relationship between test publication date and test length. Panel B presents the linear and quadratic relationships between publication date and the number of factors purportedly measured by the test.

was found between test publication date and the difference between the number of factors indicated by test authors and the number of factors indicated by HPA, $r(14)=.51$, $p=.055$. The relationship between test publication date and the difference between the purported number of factors and the number of factors indicated by MAP was also large and showed a trend toward significance, $r(14)=.40$, $p=.146$.

The hypothesis that conventional criteria lead to overextraction was evaluated by comparing the number of factors retained by Cattell's scree test, Kaiser criterion, chi-square statistic, CFI, and TLI and the number of factors retained by HPA and MAP. The number of factors retained by Kaiser's criterion was marginally significantly larger than the number of factors retained by HPA and MAP, $t(14)=2.09$, $p=.055$ and $t(14)=2.07$, $p=.057$ respectively. Cattell's scree test, chi-square statistic, CFI, and TLI all retained a larger number of factors than indicated by HPA and MAP, smallest $t(14)=2.41$, $p=.030$. Even if the Woodcock–Johnson data are removed, the number of factors purported by HPA and MAP is substantially lower than other criteria (smallest $t(12)=1.76$, $p=.104$, Cohen's d=.50) and approaches significance in spite of poor power.

To further examine whether the use of liberal criteria, has led to recent increases in the overfactoring of ability tests, both qualitative and quantitative analyses were performed. Table 2 presents test publication date and the criteria used to determine the number of factors measured by each test. Beginning with the publication of the K-ABC in 1983, rules resulting from either exploratory or confirmatory approaches or both were used to determine the structure of all subsequent ability tests. Combining information from Tables 1 and 2, following publication of the K-ABC in 1983 (the point at which factor methods began to be used extensively in choosing the number of factors measured), the WPPSI-R is the only test whose purported number of factors matches the number of factors indicated by MAP and HPA. Interestingly, the WPPSI-R was also the only test that relied exclusively on Kaiser criterion for determining the number of factors measured, and the Kaiser criterion was shown earlier to perform the closest to MAP and HPA of any of the comparison criteria (see Table 1). Interestingly, while previous simulation work has indicated that Cattell's scree test is often fairly accurate, but difficult to score in some cases, in the present study it was found to perform similarly to the Kaiser criterion $t(14)=-0.17$, $p=.866$.

Quantitative evaluation was also performed to determine whether the more conventional criteria suggested retention of the same number of factors, on average, as the number indicated by test authors. Cattell's scree test, Kaiser criterion, and CFI all suggested retaining significantly less factors than the number proposed by the tests' developers, smallest $t(14)=2.86$, $p=.013$. The number of factors suggested by chi-square was significantly larger than the number of factors purportedly measured by the tests, $t(14)=6.15$, $p<.001$, while the number of factors indicated by TLI did not differ significantly from the number of factors suggested by test authors, $t(14)=0.90$, $p=.384$. Examination of the technical manuals for recent commercial ability tests indicated that, in most instances, a number of different criteria were used to determine the number of factors measured by these tests, including rules based upon EFA and CFA analyses. Therefore, the number of factors retained by each of the more conventional criteria were averaged and compared to the purported number of

Table 2
Publication year, administration time (min), and criteria used to determine the number of factors measured for each ability test

Test	Publication year	Administration time (min)	A priori theoretical considerations	EFA/PCADecision rules	CFA
WISC	1949	None reported	×		
WAIS	1955	None reported	×		
WPPSI	1967	62	×		
WISC-R	1974	75	×		
WAIS-R	1981	75	×		
K-ABC	1983	80	×	×	×
SB-IV	1986	75	×		×
WJ-R	1989	100	×	×	×
WPPSI-R	1989	62	×	×	
DAS	1990	53	×	×	×
WISC-III	1991	73	×	×	×
WAIS-III	1997	80	×	×	×
WJ-III	2001	100	×		×
WPPSI-III	2002	69	×	×	×
WISC-IV	2003	73	×	×	×

factors. The number of factors retained based upon this average did not differ from the number of factors chosen by test authors, $t(14) = 0.75$, $p = .466$.

7. Discussion

Several important findings emerged from the present study. As predicted, commercial ability tests have become increasingly complex. While the length of these tests has risen only moderately, the number of factors purportedly measured by these tests has risen substantially, possibly even exponentially. It should be noted, however, that the possibility of an exponential increase in the number of factors purportedly measured may be due to inclusion of two outliers, the WJ-R and WJ-III. Possibly even more convincingly, the ratio of test length to factors purported has decreased dramatically. These trends suggest that test authors may be positing additional factors without including a sufficient number of subtests to measure these factors. When more accurate, recommended, statistical criteria were examined commercial ability tests were found to be substantially overfactored. This held true even when less accurate (Kaiser criterion) and potentially subjective (Cattell's scree test) criteria were used.

Results of the present study also suggest that overfactoring of ability tests may be worsening, as the discrepancy between the purported number of factors and the number indicated by MAP and HPA has risen over time and the ratio of subtests to factors purported has decreased substantially as well. While commercial pressures and increasingly complex models of human cognitive abilities are likely contributing to these recent increases, these explanations were not investigated in the present study. Rather, evaluation centered on the hypothesis that test developers have been determining test structure using liberal, and often inaccurate, statistical criteria. This hypothesis was supported by three findings. First, tests developed in the early-to-mid 1980s up to the present used the more conventional, and often inaccurate or subjective, criteria (Kaiser criterion and Catell's scree test) along with CFA-based indices. In contrast, none of the more recent ability tests used HPA or MAP to determine the number of factors measured. Secondly, all of the more conventional statistical criteria evaluated in the present study (both CFA and EFA-based indices) suggested retention of more factors than proposed by HPA and MAP. Lastly, the number of factors retained using a combination of Cattell's scree test, Kaiser criterion, chi-square, CFI, and TLI approximated the number purported by test authors. The latter evidence was not conclusive, however, since specific differences were found between the number of factors proposed by test authors and the number indicated by Cattell's scree test and Kaiser criterion (fewer factors retained), and CFI and chi-square (more factors retained).

These specific differences leave two possibilities. It may be the case that either the number of factors purportedly measured by the tests was based only on the TLI or other CFA fit indices, or that combinations of fit indices and decision rules were used to determine the number of factors. The latter possibility appears more likely. Examination of the technical manuals for the more recent ability tests indicated that most tests used more than one criterion to determine structure, with each test using slightly different combinations of analyses and decision rules (see Table 2). The use of combinations of criteria makes it likely that the most liberal rule (chi-square statistic) was offset by consideration of the least liberal rule (Kaiser criterion).

The present study was the first to the authors' knowledge to compare the performance of HPA and MAP to CFA-based indices. That fact that chi-square retained significantly more factors than HPA and MAP was not surprising, since previous empirical work has found this rule to overfactor (Zwick & Velicer, 1986). However, it was interesting that both CFI and TLI suggested retention of more factors than HPA and MAP. This finding suggests that these indices lead to overextraction. It is possible that other CFA-based indices (such as AIC, BIC or RMSEA) may be more useful in determining the number of factors. While these statistics were not used to derive the published factor structures examined in the present study, these indices should be examined in future attempts to derive the structure of cognitive batteries.

Monte Carlo studies comparing CFA-based indices to HPA and MAP are needed to determine whether CFA indices do, in fact, lead to retention of too many factors. The present study also found poor agreement between MAP and HPA. This is in contrast with other unpublished work we have performed using these criteria and may be due to the restricted range observed in this study (range 1–2 for MAP and 1–3 for HPA with only one analysis indicating 3). Further, MAP and HPA suggested retention of approximately the same number of factors on average, and only disagreed by two factors on one occasion. Clearly more research is needed to determine the situations in which MAP and HPA are most effective. Used in conjunction they will, at worst, significantly reduce the number of plausible structures to be investigated.

7.1. Implications for future test development and clinical practice

Recent increases in overfactoring have important implications for future test development. Authors of future cognitive ability batteries will have two basic choices. They can continue to attempt to measure additional aspects of cognitive ability beyond *g* or they can focus on developing shorter measures that provide estimates of general ability, and possibly one or two additional factors. Those that wish to measure more aspects of cognitive ability will need to increase the length of these batteries beyond the length of present assessment batteries. Presently, the rate of increase in the number of factors purportedly measured by these tests is much greater than the rate of increase in test length. Reversing this trend will require use of decision rules that accurately determine the number of factors measured as well as inclusion of a number of subtests that adequately measure additional dimensions and do not load significantly on other dimensions.

Methodologists have stated that at least three indicators are needed to identify a factor and four are needed to determine if a factor is over-identified (Fabrigar et al., 1999; Velicer et al., 2000). Many current ability tests do not meet these requirements. Two examples are the WAIS-III and the DAS. Currently, the WAIS-III includes three 3-subtest factors and one 2-subtest factor, while the DAS includes three 2-subtest factors in the core cognitive battery. In each case, all of the factors fall short of the four indicators recommended by methodologists. In fact, the DAS does not even include enough subtests in the core battery to examine the proposed structure. Having a small number of indicators, measuring one or more of the factors in a test battery, results in several problems. The most clinically important of these are the unreliability of factor scores and resulting difficulties in establishing predictive validity.

7.1.1. Reliability

The minimum level of internal consistency reliability required to make decisions about individuals is generally thought to be .85–.90 (Nunnally & Bernstein, 1994; Rosenthal & Rosnow, 1991). In situations where only a few indicators are used to measure each factor, the reliability of factor score comparisons is likely to be insufficient for making decisions about individuals. This is due to the fact that such comparisons are typically based upon a difference score and difference scores often have lower reliability than their parent scores. At present, even the most reliable factor score comparisons from the WAIS-III and DAS fall short of this level. Using the formula from Streiner and Norman (1995) for the reliability of a difference score, the reliability of WAIS-III and DAS factor score comparisons range from .76 to .86 and .73 to .80, respectively. As a result, the comparisons between factor scores, that clinicians routinely use to make decisions, lack sufficient reliability.

Increasing the length of the WAIS-III by 5 subtests and the DAS by 6 subtests, so that each index has 4 indicators, enhances the reliability of factor score comparisons. Using the Spearman–Brown prophecy formula for estimating the reliability of a longer test, one can estimate the reliability of factor score comparisons based upon a larger number of indicators (Streiner & Norman, 1995). The reliability of lengthened WAIS-III and DAS factor score comparisons is estimated to range from .87 to .91 and .87 to .90. While these are not huge increases, these estimates should be viewed as conservative because the reliability of score differences could be greatly enhanced by adding subtests with greater specificity. Nonetheless, these estimates put comparisons for the WAIS-III and DAS close to or above the recommended levels of reliability. In both

cases, adding indicators to the existing factors results in significant improvements in the reliability of factor score interpretations, on average an increase in true score variance of 7% for the WAIS-III and 12% for the DAS.

7.1.2. Validity

Establishing predictive validity is essential for justifying the extra time and effort required in giving extensive test batteries. Yet, the less than desirable reliability of current factor-score comparisons makes it difficult to establish validity. Consistent with this, research has suggested that commonly used factor scores from commercial tests have had a disappointing track record, historically, in terms of demonstrating divergent, incremental, or predictive validity (Glutting, Youngstrom, Ward, Ward, & Hale, 1997; Moffitt & Silva, 1987). Yet, it is not clear whether these findings have resulted from poor measurement of additional factors (i.e. overfactoring) or if additional measurement does not provide information. Results of the present study suggest that previous failures in establishing incremental validity may be due to the fact that many of the factors measured in these studies are poorly defined and lacking in specificity. Future research in this area should attempt to improve the measurement of additional ability factors. This will necessarily involve inclusion of larger numbers of subtests, with sufficient reliability and specificity, tapping additional cognitive abilities. Future work should also employ more accurate statistical techniques for determining the number of factors measured by assessment batteries, such as HPA and MAP. This will help ensure that failure to find incremental validity is not a result of poorly defined factors.

7.1.3. Cost

Longer, more carefully designed, test batteries will undoubtedly provide reliable assessment of additional ability factors. Unfortunately, however, movement toward lengthier assessment tools will also result in substantial increases in cost to the consumer. Glutting and colleagues (2003) describe how longer test batteries yield increased administration, scoring, and report writing time. To demonstrate the increased cost of longer assessments, they computed that an hour increase in the current length of school-related assessments would result in a $55,000,000 increase in costs to the US public, K-12 educational system. Using the numbers provided by Glutting et al. (2003), which are based upon the average number of school psychological evaluations performed per year, and the mean administration time computed from Table 2, the cost of current intellectual assessment batteries to the educational system is approximately 1.25 h administration × $33.33 per hour × 72 assessments per year × 23,000 practitioners = $68,993,100 per year. Psychometrically sound measurement of additional ability factors will probably require increasing the length of existing batteries by at least 1/2 the current length and in some cases doubling the length of existing batteries. As a result, the cost to the education system will increase. If one assumes that increasing the length of existing batteries by 1/2 the current length will also increase administration time by 1/2 the current length, then the cost of these longer batteries to the education system will be $103,489,650 per year, an increase of $34,496,550 per year over current assessment instruments.

These numbers suggest that the benefits of longer assessments are probably outweighed by the substantial costs. Future research employing more reliable factor score comparisons may substantiate the validity of additional measurements and thereby shift this cost/benefit ratio. However, validity evidence will not eliminate the need for briefer, more frugal measures of general ability in some clinical settings. For example, lengthier tests are impractical in environments where quicker assessment of cognitive functioning is needed or longer assessments are too costly or simply not possible. At present, there are only two test batteries designed exclusively to provide quick assessments of general ability, the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999) and the Wide Range Intelligence Test (WRIT; Glutting, Adams, & Sheslow, 1999). These batteries allow for a quick assessment of general ability in a fraction of the time required to administer more traditional intelligence tests. Tests such as the WASI and WRIT have clear utility in managed care settings or other environments where longer assessments significantly reduce the number of individuals that can be tested and increase the cost for individual assessments. As clinicians become aware of the lack of validity evidence for additional ability factors and continue to be pressured toward shorter assessments by managed care companies, longer assessment batteries will likely give way to shorter instruments.

8. Summary

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. The results of this study do not suggest that CFA is not a useful approach to examining the structure of cognitive abilities. CFA methods involve attempts to determine the factor structure of a data set in the population, excluding measurement error. This fact

makes CFA methods particularly useful for developing theory regarding the structure of intellectual abilities, since poorly defined factors can be identified. However, minor factors may not possess sufficient reliability to make decisions on the individual level. Since, the common use of cognitive ability batteries is to assess and make decisions about individuals, more conservative decision rules, retaining only well defined and replicable factors may be much better suited for test development. Therefore, it is recommended that future test developers focus on conservative, but accurate, criteria based upon EFA methods, such as HPA and MAP, in order to ensure that factor scores derived from commercial ability tests are clinically useful and provide sufficient reliability and specificity for making score comparisons. Alternatively, CFA methods are likely to be most useful to researchers in developing theory about specific distinctions among intellectual abilities.

Authors of future cognitive ability tests may also decide to direct their focus on briefer assessments of general ability, or lengthy assessment batteries that provide psychometrically sound measurement of additional cognitive ability factors. The future of lengthy assessment batteries will depend upon research substantiating the incremental validity of additional measurement.

References

Arbuckle, J. L. (1999). *Amos (Version 4.0).* Chicago: Smallwaters.

Bentler, P. M. (1988). Comparative fit indexes in structural equation models. *Psychological Bulletin*, *107*, 238−246.

Boyle, G. J. (1989). Confirmation of the structural dimensionality of the Stanford–Binet Intelligence Scale (Fourth Edition). *Personality and Individual Differences*, *10*(7), 709−715.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York, NY: Cambridge University Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1−22.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*, 245−276.

Cattell, R. B., & Vogelmann, S. (1977). A comprehensive trial of the scree and KG criteria for determining the number of factors. *Multivariate Behavioral Research*, *12*, 289−325.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*, 2nd edition. Hillsdale, NJ: L. Erlabaum Associates.

Crawford, C. B., & Koopman, P. (1979). Note: Inter-rater reliability of scree test and mean square ratio test of number of factors. *Perceptual and Motor Skills*, *49*(1), 223−226.

Elliott, C. D. (1990). *Differential ability scales: Introductory and technical handbook.* San Antonio, TX: The Psychological Corporation.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272−299.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York, NY: Basic Books.

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.

Glutting, J. J., Adams, W., & Sheslow, D. (1999). *Wide range intelligence test manual.* Wilmington, DE: Wide Range, Inc.

Glutting, J. J., Watkins, M., & Youngstrom, E. A. (2003). Multifactored and cross-battery assessments: Are they worth the effort? In C. R. Reynolds & R. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children*, 2nd ed. New York: Guilford.

Glutting, J. J., Youngstrom, E. A., Ward, T., Ward, S., & Hale, R. (1997). Incremental efficacy of WISC-III factor scores in predicting achievement: What do they tell us? *Psychological Assessment*, *9*, 295−301.

Gorsuch, R. L. (1983). *Factor analysis*, 2nd ed. Hillsdale, NJ: Erlbaum.

Gridley, B. E., & McIntosh, D. E. (1991). Confirmatory factor analysis of the Stanford–Binet Fourth Edition for a normal sample. *Journal of School Psychology*, *29*(3), 237−248.

Guilford, J. P. (1967). *The nature of human intelligence.* New York, NY: McGraw-Hill.

Hakstian, A. R., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number-of-factors rules with simulated data. *Multivariate Behavioral Research*, *17*, 193−219.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179−185.

Horn, J. L. (1968). Organization of abilities and the development of intelligence. *Psychological Review*, *75*, 242−259.

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141−151.

Kaufman, A. S., & Kaufman, N. L. (1983). *K-ABC: Kaufman Assessment Battery for Children administration and scoring manual.* Circle Pines, MN: American Guidance Service.

Kim, J., & Mueller, C. W. (1978). *Factor analysis: Statistical methods and practical issues*, *Vol. 14.* Iowa City, IA: Sage Publications.

Kline, R. B. (1989). Is the fourth edition Stanford–Binet a four-factor test? Confirmatory factor analyses of alternative models for ages 2 through 23. *Journal of Psychoeducational Assessment*, *7*(1), 4−13.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: Guilford.

Lees-Haley, P. R., Smith, H. H., Williams, C. W., & Dunn, J. T. (1996). Forensic neuropsychological test usage: An empirical survey. *Archives of Clinical Neuropsychology*, *11*(1), 45−51.

MacCallum, R. (1986). Specification searches in covariance structure modeling. *Psychological Bulletin*, *100*(1), 107−120.

MacCallum, R., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: The problem of capitalization on chance. *Psychological Bulletin*, *111*(3), 490−504.

McGrew, K. S., & Woodcock, R. W. (2001). *Woodcock–Johnson III: Technical manual.* Itasca, IL: Riverside Publishing.

Moffitt, T. E., & Silva, P. A. (1987). WISC-R verbal and performance IQ discrepancy in an unselected cohort: Clinical significance and longitudinal stability. Special Issue: Eating disorders. *Journal of Consulting and Clinical Psychology*, *55*(5), 768−774.

Naglieri, J. A., & Kaufman, A. (1983). How many factors underlie the WAIS-R? *Journal of Psychoeducational Assessment*, *1*, 113−119.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*, 3rd ed. New York: McGraw-Hill, Inc.

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, and Computers*, *32*, 396−402.

O'Grady, K. E. (1983). A confirmatory maximum likelihood factor analysis of the WAIS-R. *Journal of Consulting and Clinical Psychology*, *51*, 826−831.

Parker, K. C. H. (1983). Factor analysis of the WAIS-R at nine age levels between 16 and 74 years. *Journal of Consulting and Clinical Psychology*, *51*, 302−308.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis*, Second Edition. New York: McGraw-Hill, Inc.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420−428.

Silverstein, A. B. (1982). Factor structure of the Wechlser Adult Intelligence Scale-Revised. *Journal of Consulting and Clinical Psychology*, *50*, 661−664.

Spearman, C. (1904). "General Intelligence," objectively determined and measured. *American Journal of Psychology*, *15*, 201−293.

SPSS (2002). *SPSS professional statistical package (Version 11.0).* Chicago, Illinois: SPSS Inc.

Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales: A practical guide to their development and use*, 2nd ed. New York: Oxford University Press.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics*, 3rd ed. New York, NY: HarperCollins.

Thorndike, R. M. (1990). Would the real factors of the Stanford–Binet Fourth Edition please come forward? *Journal of Psychoeducational Assessment*, *8*(3), 412−435.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Technical manual, Stanford–Binet intelligence scale: Fourth edition.* Chicago: Riverside.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs*, *1*.

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for maximum likelihood factor analysis. *Psychometrika*, *38*, 1−10.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*(3), 321−327.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: Honoring Douglas N. Jackson at seventy* (pp. 41−71).

Velicer, W. F., & Jackson, D. N. (1990). Component analysis versus common factor analysis: Some issues in selecting the appropriate procedure. *Multivariate Behavioral Research*, *25*, 1−28.

Vernon, P. E. (1950). *The structure of human abilities.* London: Methuen.

Ward, L. C., Ryan, J. J., & Axelrod, B. N. (2000). Confirmatory factor analyses of the WAIS-III standardization data. *Psychological Assessment*, *12*(3), 341−345.

Watkins, M. (2000). *Monte Carlo PCA for parallel analysis.*

Wechsler, D. (1949). *Manual for the Wechsler Intelligence Scale for Children.* New York, NY: The Psychological Corporation.

Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale.* New York, NY: The Psychological Corporation.

Wechsler, D. (1967). *Manual for the Wechsler Preschool and Primary Scale of Intelligence.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children-Revised.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1981). *Manual for the Wechsler Adult Intelligence Scale-Revised.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1989). *Manual for the Wechsler Preschool and Primary Scale of Intelligence-Revised.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children-Third Edition.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *Manual for the Wechsler Adult Intelligence Scale-Third Edition.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1997). *WAIS-III WMS-III technical manual.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2002). *WPPSI-III: Technical and interpretive manual.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003). *WISC-IV: Technical and interpretive manual.* San Antonio, TX: The Psychological Corporation.

Widaman, K. F. (1993). Common factor analysis versus principal component analysis: Differential bias in representing model parameters? *Multivariate Behavioral Research*, *28*, 263−311.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock–Johnson psycho-educational battery-revised.* Allen, TX: DLM Teaching Resources.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock–Johnson III tests of ability.* Itasca, IL: Riverside Publishing.

Zwick, W. R., & Velicer, W. F. (1982). Factors influencing four rules for determining the number of components to retain. *Multivariate Behavioral Research*, *17*, 253−269.

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.