Intelligence 37 (2009) 487–494

Contents lists available at ScienceDirect

Intelligence

Identifying *g*: A review of current factor analytic practices in the science of mental abilities

Charlie L. Reeve [a,*], Nikki Blacksmith [b]

[a] *Interdisciplinary Health Psychology Program, University of North Carolina Charlotte, United States*
[b] *Gallup, Inc, United States*

ARTICLE INFO

Article history:
Received 21 March 2009
Received in revised form 5 June 2009
Accepted 9 June 2009
Available online 1 July 2009

Keywords:
Factor analysis
g
Cognitive abilities

ABSTRACT

Factor analysis is arguably one of the most important tools in the science of mental abilities. While many studies have been conducted to make recommendations regarding "best practices" concerning its use, it is unknown the degree to which contemporary ability researchers abide by those standards. The current study sought to evaluate the typical practices of contemporary ability researchers. We analyzed articles reporting factor analyses of cognitive ability tests administered to adult samples over a 12 year period. Results suggest that, in aggregate, the science of mental abilities seems to be doing well with respect to the issues of sample size, number of indicators (relative to number of factors) and breadth of indicators. Further, our results suggest that the majority of ability researchers are using methods of factor analysis that allow for the identification of a *g* factor. However, 14.57% failed to use a method that allowed a common factor to emerge. These results provide insights regarding the methodological quality of the science of mental abilities, and will hopefully encourage further "introspective" research into the science of mental abilities.

The use of factor analysis in the study of cognitive abilities is in many ways the equivalent to the use of the telescope in the study of astrological bodies. Both are the primary tool for their respective science and it is hard to imagine either science having advanced in any meaningful way without these essential tools. Indeed, although Galton's (1883) theorizing was critical to the founding of the science of mental abilities, it was not until Spearman (1904) and his students (e.g., Garnett, 1919, and El Koussy, 1935) introduced factor analysis, that the psychometric structure of cognitive abilities could be explored in a more systematic fashion. However, the quality of information gained by factor analyses depends on a number of rather basic methodological considerations. For example, research shows that choices about sample size, number and breadth of indicators used, and the specific method of analyses used may affect the specific results obtained and the conclusions drawn from them (e.g., MacCallum, Widaman, Zhang, & Hong, 1999).

While many studies have been conducted to make recommendations regarding "best practices" (e.g., Jensen & Weng, 1994; MacCallum, et al., 1999; Marsh, Hau, Balla, & Grayson, 1998), there appears to be little research regarding the actual practices of contemporary ability researchers. In short, when it comes to conducting factor analyses of cognitive ability tests, it is relatively clear what researchers ought to do, but it is not clear what researchers actually do.

We believe it is important for researchers to occasionally take a step back and assess how their tools and methods are being used. Such self-reflective examinations of these basic methodological issues are not uncommon in other areas of psychology. For example, in organizational psychology, several researchers have investigated trends in research designs, (e.g., Sackett & Larson, 1990; Scandura & Williams, 2000), scale development practices (e.g., Hinkin, 1995), and sample sizes (Ghiselli, 1973; Monahan & Muchinsky, 1983; Salgado, 1998). Unfortunately, there are relatively few examples of such introspection among ability researchers. A rare exception is a recent paper by Frazier and Youngstrom (2007), who examine the historical increase in the number of factors extracted from common intelligence tests. Their

* Corresponding author. Department of Psychology, University of North Carolina Charlotte, 9201 University City Boulevard, Charlotte, NC 28223-0001, United States.
E-mail address: clreeve@uncc.edu (C.L. Reeve).

0160-2896/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2009.06.002

results showed that contemporary researchers are "over-factoring" the data from ability tests. In our opinion, this is a critical example of why such introspection is important, as it has important implications for both scientists and practitioners. Thus, the purpose of the current study is largely descriptive; we seek to provide information regarding current practices in conducting factor analyses of cognitive ability tests. In particular, we consider method of factor analysis, number and breadth of indicators, and sample size.

1. Method of factor analysis

In the abstract it is of course non-sensical to discuss which method of factor analysis is best, as different methods have different purposes. Similarly, the definition of "appropriate application" of factor analysis depends on the nature of theoretical model thought to underlie the data (e.g., number and nature of factors, factor covariances, etc). While the factor model and method of factor analysis are distinct concepts, they are inter-related in practice. That is, generally speaking, an appropriate method of factor analysis is the one that is best suited to estimate the model that represents the construct space thought to be the source of variance in the manifest indicators. Within the context of mental abilities, there seems to be little question among contemporary researchers regarding the appropriate model. Generally speaking, there seems to be consensus that abilities are best represented in something akin to the Cattell-Horn-Carroll (CHC) three-stratum model with a general '*g*' factor at the apex (McGrew, 1997, 2005). Indeed, it seems few would argue against the idea that a model of abilities must account for the pervasive positive manifold among the indicators.

From this perspective then, the primary question concerns the appropriate method for best representing the correlation matrix of mental ability tests. Jensen and Weng's (1994) treatment of this topic stands out as a seminal answer to this question. First, they note that methods or models specifying orthogonal factors are "absolutely inappropriate" because these methods mathematically preclude a *g*-factor. Such models are contradictory to the now well-accepted theoretical structure of mental abilities which includes a *g*-factor. Thus, exploratory factor analytic (EFA) methods using orthogonal rotations such as Kaiser's (1958) varimax method are considered inappropriate. Similarly, confirmatory factor analytic (CFA) methods (aka, LISREL models) which specify independent group factors are also inappropriate for the same reasons. In contrast, Jensen and Weng note that methods such as Principal Components Analysis (PCA) and EFA methods like Principal Factors Analysis (PFA) using oblique rotations, and CFA methods that posit a general factor, such as the bi-factor model or higher-order models, are generally appropriate methods. Moreover, at least with respect to estimates of the *g*-factor itself, they empirically demonstrated that these methods generally provide highly similar results.

1.1. Number and breadth of indicators

Like the decision regarding the appropriate method of extraction to use, the question of how many indicators should be included depends on nature and breadth of the construct space one is attempting to measure. Clearly, if one is trying to assess the broad spectrum of cognitive abilities, a greater number of varied indicators are needed than if one is only measuring a single specific ability. Thus, in the abstract, it is difficult to define any general guidelines. That being said, we believe it is instructive to frame this question around the issue of how best to represent *g*. We propose this as an important focal question for four reasons. First, we note the central role *g* appears to play in a wide range of educational, occupational, and social outcomes. Second, we note recent evidence (Reeve & Charles, 2008) showing expert consensus for the idea that, although there is certainly more to intelligence than just *g*, *g* is the single most important ability determinant of cognitively-loaded performances. Third, empirical evidence confirms that it is the *g*-saturation of ability indicators that is generally responsible for the broad criterion-related validity of ability tests (Jensen, 1986), whereas specific ability factors typically only predict narrow, content specific components of criterion variance (e.g., Reeve, 2004). Fourth, as will be noted below, an appropriate estimate of *g* requires the use of a broad and diverse set of ability indicators, where "broad and diverse" is defined with respect to the range of second stratum abilities (aka, broad abilities, group factors, specific abilities).

Again, Jensen and Weng (1994) provide clear, empirically driven recommendations on this issue. They note that the goodness of *g*-loadings drawn from a factor analysis is a function of the number and the diversity of mental abilities or tasks represented. Indeed, they point out that increasing number and breadth of reliable and valid indicators increases the quality of the *g* estimate because it decreases the amount of psychometric sampling error. Thus, from a purely psychometric perspective, more is usually better (assuming an appropriate degree of validity, diversity and reliability). Additionally, from a purely empirical perspective, Marsh et al. (1998) demonstrated that more indicators per factor are always better, regardless of sample size, when the number of factors was relatively restricted. However, there are obviously practical limitations in the number of indicators that can be reasonably obtained, and MacCallum et al. (1999) indicate that it is optimal to avoid situations where both the number of indicators and number of factors are large.

Thus, the question remains, how many indicators should be used in the factor analysis of cognitive abilities? Although there is no clear answer, if it is assumed that one is trying to generally model the basic domain of cognitive abilities and/or obtain a reasonable estimate of *g*, a few reasonable guidelines can be extracted from the literature. First, from a purely conceptual standpoint, it can be recommended that an ideal battery would contain at least one indicator per specific ability. Given that most models of cognitive abilities include 8 (e.g., Carroll, 1993) to 10 (e.g., McGrew, 1997) specific group factors, this would imply that 8 to 10 indicators are appropriate. That is, from a content validity perspective, if one wishes to broadly sample and cover the domain of abilities, there should be at least the same number of indicators as there are stratum two abilities. Consistent with this, Jensen and Weng (1994) demonstrate that 10 diverse tests appear to be a sufficient number to obtain a highly stable estimate of *g*, without worrying about the specific weights (i.e., *g*-loadings) of the tests included. Likewise, Ree and Earles (1991) empirically demonstrated that different methods of analysis converge on virtually identical *g*-factors when

10 diverse tests were included in the matrix. Thus, it would seem reasonable to conclude that researchers should be using batteries of indicators with at least eight diverse indicators (if one subscribes to the Carroll, 1993, model), though 10 or more can be recommended on the basis of McGrew's (2005) CHC model and previous empirical work.

1.2. Sample size

The question of sample size is not unique to factor analysis; the validity of almost all research methodologies and quantitative analyses hinge on the use of an appropriate number of observations. For example, from a theoretical perspective, the validity generalization hypothesis (Schmidt & Hunter, 1977) suggests that sampling error explains a large portion of the variability seen in validity coefficients when the average sample size is not large. Likewise, with respect specifically to factor analysis, researchers have demonstrated that sample size influences model fit and standard errors of the parameters in a factor model (e.g., Gagne & Hancock, 2006; Marsh et al., 1998). Indeed, it seems to be widely understood that, all things else being equal, larger sample sizes yield more reliable solutions and factor loadings that are more precise estimates of population parameters (e.g., Bryant & Yarnold, 2000; Gagne & Hancock, 2006; MacCallum, et al., 1999; Marsh, et al., 1998). However, in practice, the number of observations that can be reasonably obtained may not always be sufficient. The question of course is what constitutes a "reasonable sample size."

Currently, it seems many researchers rely on "rules of thumb" to determine the minimum sample size needed. For example, one rule states that there needs to be a subjects-to-variables ratio (aka, N/p ratio) of at least 5 to 1 (Bryant & Yarnold, 2000). Another rule suggests that there should never be less than 100 subjects, despite the subjects-to-variable ratio (Gorsuch, 1983). Other researchers (e.g., Anderson & Gerbing, 1984; Guilford, 1954) suggest that 200 subjects should be the minimum. Others have suggested that samples of 500 or more should be sought (e.g., Comrey & Lee, 1992).

However, the accuracy of these rules of thumb has been questioned. For example, Marsh et al. (1998) directly investigated the impact of sample size on model convergence across different indicator per factor ratios. When the indicator per factor ratio (i.e., p/K ratio) was 2.0, the proper convergence was not reached until the sample was as large as 1000. However, when the number of indicators per factor was 6.0, proper convergence could be reached with a sample as small as $N = 50$. Perhaps the most relevant explication of the impact of sample size can be found in MacCallum et al. (1999). These authors show that sample size has relatively little impact only when communalities are high (i.e., over .60) and a small number of factors that are well-determined (i.e., only a few, highly saturated indicators per factors); in such cases, which they caution rare, $N < 100$ can lead to a good recovery of parameters. However, as communalities decrease to .50 or lower, much larger samples are required. For example, in the more realistic situation where not all communalities are greater than .6 and the indicator to factor ratio (i.e., p/K ratio) less than 6, at least $N = 300$ is needed for good recovery. If communalities are low, and there are a larger number of factors (i.e., more than 3 or 4) with a small p/K ratio, a sample size of more than 500 is likely required.

2. Rationale for the current study

Understanding how factor analysis is being implemented in the science of mental abilities is important for several reasons. First, if inappropriate methods are being employed, the apparent factor structure or saturation of various indicators might be misrepresented. For example, as Jensen and Weng (1994) pointed out, the use of orthogonal rotation methods is "absolutely inappropriate" (p. 237) as it scatters the variance due to *g* across the other factors. Use of such a solution would likely lead to inaccurate inferences regarding the importance of the predictive power of narrow abilities. Second, factor loadings are often used to generate factor scores for individuals, which are then used for further analysis, selection, or diagnosis. To the degree that inappropriate methods, insufficient sample size, or insufficient number of indicators are used, the validity of scores generated from such a solution may suffer. Said differently, these first two issues stem in part from indeterminacy of solutions. For any given factor solution, there are a range of equally viable factors (often referred to as different rotations). Unless communalities are very high, these different factors may be quite different from one another, rendering our scoring of subjects on the theorized latent variables questionable. Accurate factor scores require selection of factor solutions that have multiple and sound links to theory and data (Loehlin, 2006, p. 199). Third, factor solutions can be used to help practitioners select from among a larger battery of tests to create focused or shorter batteries. Reliance on results from analyses based on inappropriate or insufficient methods could result in the selection of tests that fail to provide an optimal estimate of the targeted ability. Fourth, the degree to which manifest indicators are *g*-saturated (reflected by the standardized factor loading of a variable on the *g* factor) can be used to gain insight into the nature of performance determinants (Spearman & Jones, 1950). For example, vocabulary tests and matrix reasoning tests have similar *g*-loadings because the acquisition of vocabulary and solving matrix problems both draw upon deductive or inductive reasoning (Jensen, 1998). Fifth, prior research indicates the *g*-saturation of a manifest variable is related to criteria such as the manifest variable's predictive validity, and heritability (Jensen, 1986). Such research assumes that *g*-loadings have been well estimated.

Importantly, each of these endeavors implicitly assumes a certain degree of stability and accuracy of the estimation of factor loadings. However, although an indicator's theoretical level of ability-saturation is arguably stable (i.e., the degree to which an indicator taps a given ability or abilities), it is also true that the factors loadings extracted from a factor analysis are simply estimates which can vary across analyses. As such, we believe it is valuable to examine basic methodological issues concerning the factor analytic practices of contemporary researchers such that we can better assess the overall health of our science.

3. Method

3.1. Procedures

We sought articles reporting factor analyses of cognitive ability batteries published during the dozen years since Jensen and Weng's (1994) proscriptive paper appeared. To obtain articles, we conducted hand searches of each article published

in *Intelligence* and the *Journal of Applied Psychology* between 1994 and July 2006 (the month in which we began the literature search). Our decision to rely heavily on two journals was based on the belief, as noted in our introduction, that the factor analysis of ability tests are common and important for both basic and applied researchers. Thus, we selected a premier journal from each of these areas. In addition, we conducted a search on PsychInfo using the search terms "cognitive ability" and "factor analysis." Given that the factor analysis of cognitive ability batteries is often not the primary purpose of articles, it is difficult to estimate how frequently such analyses are reported across the entire psychological literature. Further, such analyses are not likely to always be listed as keywords defining articles in databases. Thus, we do not assume our search was comprehensive. However, we believe the hand search of two journals that publish research in which cognitive ability batteries are often used, along with the PsychINFO search, provides at least a reasonably representative sample of the contemporary literature.

Articles meeting the following inclusion criteria were kept: (a) the study was published between January 1994 and July 2006, (b) the study reports results of factor analysis on cognitive ability measures, (c) enough information is reported to code for type of analysis, number of indicators, and sample size, and (d) the data is based on non-clinical and non-child sample (child defined as 11 years and below). The decision to exclude research based exclusively on clinical samples (i.e., research conducted by clinical psychologists on samples of institutionalized individuals, or samples with documented disabilities/disorders) as well as child and adolescent samples was based on desire to better define the population of researchers and focal literature. While it is acknowledged that these criteria likely excludes most of the school and clinical psychology literatures, we felt it was better to have a well defined population and literature in order to maximize the interpretation of our results. For example, by restricting our sample as such, all of the articles we reviewed used tests/indicators designed for adults within the normal range of ability.

The literature search provided 98 unique articles with 124 unique samples (a full listing of the articles used for this study is available from the first author). Of the 98 articles, 78 were published in *Intelligence*, 9 were published in *Journal of Applied Psychology*, and the remaining 11 were published in eight different psychology journals. For the analysis of sample size, the literature search resulted in an $N = 124$. In cases where the authors performed multiple factor analyses on the same sample with the same set of ability scales as steps towards a final model (e.g., performed an EFA first to gauge number of factors, then fit a CFA of some sort as the final analysis), we only recorded the "final" analysis (i.e., most complete model of the ability test data or the last step in a series of analyses) to avoid dependency or redundancy across our units of observation. However, in cases where competing models or methods were applied to the same sample, we recorded both analyses as unique observations. This resulted in a sample of $N = 144$ analyses, which constituted our sample for the analysis of indicators used and method of factor analysis.

To code the articles on our focal variables (i.e., sample size, number of indicators, breadth of indicators, factor method used, and *g*-saturation), the one author read each article and recorded the relevant information. Next, the other author compared the first author's records against each article and noted any apparent discrepancies. We then discussed these disagreements until consensus was reached. To code for the breadth of indicators, we used a coding scheme based on McGrew's (1997) analysis of the major intelligence batteries according to the CHC framework. Specifically, we used McGrew's classification of specific tests and subscales from the major intelligence batteries to classify each manifest variable used in each study as an indicator of one or more of eight broad ability factors. In cases were a study used one or more of the batteries classified by McGrew, we simply used McGrew's classifications. In cases where scales came from batteries not analyzed by McGrew, we matched scales based on content to scales within the batteries analyzed by McGrew.

4. Results

Across the 124 unique samples in our collection of studies, the average sample size was $M = 5{,}869.78$ (S.D. = 30,245.75); however, the range of sample sizes was quite large with a minimum of 19 and a maximum of 323,723. Examination of the distribution indicated the minimum and maximum were likely outliers, thus we also computed the inter-quartile mean. The inter-quartile mean of 647.08 (S.D. = 556.08) is likely a more accurate reflection of the typical sample size used, and is more in line with the median sample size of 373. Two-thirds of the sample sizes in our database were 200 or greater. To better assess the appropriateness of the sample sizes, we computed the N/p ratio for each study. The average N/p ratio was 583.56 subjects per indicator (SD = 2,121.29; range is 2.71 to 14,714.68). However, again we computed the median and inter-quartile mean due to the outlier sample sizes. The median N/p ratio was 28.25, and the inter-quartile range is 12.74 to 207.14 (inter-quartile mean = 61.12).

Table 1 provides the frequencies of factor analysis methods used in the articles. The first three methods listed refer to cases where only a single, general factor was extracted (residual variance was not further defined). Exploratory methods (including PCA) were used 38.18% of the time. CFA methods were used 45.84% of the time. Interestingly, despite the widely known hierarchical nature of cognitive abilities, and Jensen and Weng's (1994) admonishment against the use of methods that disallow the emergence of a *g*-factor, 14.57% of the analyses failed to use a method that would allowed an identifiable *g*-factor to emerge.

Across the 144 unique analyses in our database, the average number of indicators was $M = 13.91$ (Mdn = 11.00; S.D. = 10.12), with a range from 3 to 60 indicators Despite the large overall range, examination of the distribution suggests there may actually be less variation across a majority of studies. The inter-quartile range (8.00 to 15.75) and inter-quartile S.D. (1.83; Inter-quartile $M = 11.77$) indicate much less variability. Overall, the mean, median, and inter-quartile mean all suggest the typical number of indicators used is around the recommended $p = 10$ suggested by Jensen and Weng (1994). In fact, about two-thirds (67.13%) of the analyses in our sample used 10 or more indicators.

Table 2[1] displays the results of our more detailed analysis of the nature of test batteries used. For each of our specific

[1] This set of analyses excludes the two studies that did not actually conduct factor analyses (i.e., the two that used the item-total correlation method to estimate *g*-loadings). We also excluded two studies that used the items from the Raven's scale as the indicators. The reason for this was that we felt these two studies would arbitrarily inflate the statistics computed (i.e., using 36 items is qualitatively different from using ability 36 scales).

Table 1
Frequency of use of different factor analytic methods.

Estimation methods	Freq	%
1. First unrotated principal component	27	18.75
Exploratory factor analytic methods		
2. First Unrotated Principal Factor (Principal Axis)	22	15.28
3. First Unrotated Factor with ML estimation	0	0.00
4. Higher Order using EFA methods		
-PC Extraction with Oblique Rotation		
- Oblimin Rotation	0	0.00
- Promax Rotation	1	0.69
-PF Extraction with Oblique Rotation		
- Oblimin Rotation	1	0.69
- Promax Rotation	3	2.08
-ML Extraction with Oblique Rotation		
- Oblimin Rotation	1	0.69
- Promax Rotation	0	0.00
Confirmatory factor analytic methods (i.e., SEM methods)		
5. ML estimation with Higher-order model specified	38	26.39
6. ML estimation with Nested Model (aka, Bi-factor)	15	10.42
7. ML estimation with single factor specified	13	9.03
Non-FA methods		
8. Item-total correlations	2	1.39
Non-g models		
9. Oblique Rotated PCA (with no higher-order analysis)	1	0.69
10. Oblique Rotated PFA (with no higher-order analysis)	4	2.78
11. Oblique Rotated ML (with no higher-order analysis)	1	0.69
12. CFA (ML estimation) with correlated group factors	14	9.72
13. Varimax Rotated PCA (with no higher-order analysis)	1	0.69
14. Varimax Rotated PFA (with no higher-order analysis)	0	0.00
15. Varimax Rotated ML (with no higher-order analysis)	0	0.00

analyses, we report results for the entire set of analyses, and then subsets for those analyses specifying multi-factor models and those specifying single-factor models. The results of the p/K ratio analysis are shown in the top portion of Table 2. As shown, the average p/K ratio overall 6.02 indicators per factor. As would be expected, the average p/K ratio is noticeably larger for analyses employing single factor models than for those employing multi-factor models.

To assess the diversity of indicators typically used, we evaluated the number of broad factors assessed by each study's collection of indicators using the CHC classification scheme as developed by McGrew (1997, 2005). For each study, we classified each indicator used as a key indicator of one or more of the eight broad factors used by McGrew (1997) to classify the major intelligence batteries[2]. Results of this analysis are shown in the lower portion to Table 2. First, we examined the number of broad factors (out of eight) that were tapped by at least one manifest indicator within each study. These results are shown in the middle of the table. The results show that, on average, researchers appear to be using batteries of tests that cover 4 to 5 of the broad CHC factors. This does not appear to vary much based on the nature of the model to be employed, though those studies employing one factor models appear to assess about 1 less factor, on average, than those employing multi-factor models. To better understand which factors are usually assessed, we evaluated the proportion of studies using at least one indicator assessing each broad factor as classified by McGrew (1997). These results, shown in the lower portion of the table, indicate that five factors in particular are assessed the most frequently. Overall, Gc (verbal:educational) and Gf (fluid intelligence/abstract reasoning) are covered the most frequently, followed closely by Gvs (visual-spatial), Gq (quantitative abilities), and Gs (cognitive speededness). Auditory abilities and long-term memory/associated retrieval appear to be assessed relatively infrequently. However, it should be noted that our selection of journals may have influenced these results. For example, neuropsychology-oriented journals would likely have more articles using measures tapping Glr (long-term storage and retrieval). Likewise, more speech/language-oriented journals would likely have revealed more measures of Ga (auditory abilities).

Finally, we were interested in examining the typical size of the communalities (h^2) seen in the factor analysis of cognitive ability tests. That is, given MacCallum et al.'s (1999) work showing that the range and average size of the communalities influence the sample size needed for good recovery of factors, we wanted to better assess the adequacy of the sample sizes used within the context of the science of mental abilities. For each study for which we could obtain or compute h^2 estimates, we calculated the average h^2 estimate for that study, the highest and the lowest h^2 estimate within that study, and proportion of h^2 estimates that were greater than

Table 2
Analysis of test batteries used in factor analyses.

	All models	M.F. models only	1 Factor models only
p/K ratio			
Mean	6.02	3.17	9.56
Median	4.00	2.8	8.0
S.D.	5.48	1.62	6.30
Max	46.00	15.00	46.00
Min	1.43	1.43	3.00
Number of Broad Factors tapped by indicators			
Mean	4.59	4.90	4.21
Median	5.00	5.00	4.00
S.D.	1.49	1.46	1.45
Max	8.00	8.00	8.00
Min	1.00	1.00	1.00
Proportion of analyses tapping each Broad Factor			
Gf	.82	.87	.77
Gq	.69	.69	.71
Gc	.92	.90	.97
Gsm	.51	.65	.34
Glr	.15	.27	.13
Gvs	.80	.84	.76
Gs	.67	.77	.56
Ga	.06	.08	.03

Note. Overall $N = 140$. Multifactor analyses $N = 79$. One-factor analyses $N = 61$. MF = Multifactor Models.

[2] Although there are 10 broad factors in the CHC framework, we were forced to collapse the Grw (reading/writing) and Gc (crystallized knowledge) categories into a broad verbal:educational category (denoted here as Gc), and the Gs (processing speed) and Gt (decision speed) categories into a broad cognitive speededness factor (denoted here as Gs). This was done because the highly similar nature of these categories and the general lack of detail reported regarding the specific nature of many of the scales resulted in an inability to clearly classify scales into one or the other. Likewise, McGrew's (1997) own classification often showed that verbal tests could be classified as assessing both Gc and Grw. Additionally, we believe this is consistent with other models of abilities such as Vernon (1961), Guttman and Levy (1991), and Johnson and Bouchard (2005). As such, we felt it would be better to collapse these categories for the purpose of our analysis as we could be certain of the scales classification with respect the broader categories but could not be certain with respect to the specific categories.

Table 3
Analysis of range and average size of communalities.

	Communalities obtained in each study			
	Avg. h^2	Highest h^2	Lowest h^2	*Prop.* of $h^2 \geq .60$
Mean	.47	.73	.24	.25
S.D.	.13	.14	.16	.25
Max	.81	.95	.71	1.00
Min	.25	.46	0.00	0.00

Note. $N = 49$.

or equal to .60. Table 3 shows the descriptive statistics from the analysis of the range and of the average size of communalities. Across the studies, the mean average h^2 estimate was .47, and the mean range was .73 to .24, with a mean proportion of estimates greater than .60 of .25.

5. Discussion

In many ways, we believe the results of this review speaks well of contemporary users of factor analyses in the science of mental abilities. Specifically, with respect to typical sample size employed and number of indicators, the "average" research report can be evaluated positively on these methodological considerations. First, although the results of this analysis suggest there is a significant variation in sample sizes, it would appear the typical sample size meets and exceeds Anderson and Gerbing's (1984) suggested minimum of $N = 200$. In fact, a full two-thirds of the samples in our database were greater than 200. Similarly, the median N/p ratio is well above the common recommendations such as the "5/1 rule". Of course, the determination of an adequate sample size needed for good recovery of parameters depends on a number of other issues. For example, MacCallum et al.'s (1999) simulation showed that a situation where communalities vary in range from .30 to .70 and the indicator to factor ratio less than 6, a sample size of at least 300 is needed for good recovery. Our results indicate that this scenario is a likely and in many ways typical in the factor analysis of ability tests. That is, the median p/K ratio was less than 6, and the typical range of communalities is .24 to .73. The median sample size of the studies in our review was over 300, suggesting that a majority of published studies likely use adequate sample sizes to achieve good recovery.

Second, the contemporary researchers can also be evaluated favorably with respect to the number of indicators typically used. Our review found that the typical number is 10 to 13, though it should be noted that we did uncover a number of studies with significantly smaller number of indicators (e.g., as small as 3). With respect to the estimation of the *g*-factor, this suggests contemporary research is quite good when viewed in the context of Jensen and Weng's (1994) recommendations. Of course, it is not simply the number of indicators that is important, but rather the use of a diverse (and obviously valid) set of indicators. Using 10 highly similar indicators of the same type or content is not the same thing as using 10 diverse ability scales that broadly sample the domain of abilities. That is, content validity is likely to be maximized when there are a sufficiently large number of indicators to adequately represent the entire construct domain. Though there is not a universally accepted model of abilities, most experts today accept some form of a hierarchal model of abilities with a single general cognitive ability factor, referred to as '*g*', at the apex (though see Horn's work for an important "no-*g*" perspective). Different models define the lower two strata in different ways, but most models either define three broad domains reflecting verbal: educational, quantitative, and visual-spatial capacities (e.g., Vernon, 1961; Guttman, & Levy, 1991; Johnson & Bouchard, 2005), or specify 8 to 10 narrow abilities (often referred to as "group factors") that also cover these domains. From this perspective, our results suggest that researchers are generally using an appropriately diverse set of indicators. The typical study appears to use a battery of manifest indicators that covers about five of the group ability factors as defined by the CHC model of abilities. From the perspective of models that define three broad domains at the second level (e.g., Johnson & Bouchard, 2005), over half of all studies (58%) included indicators that covered all three of the broad ability domains (i.e., verbal:educational, quantitative, and visual-spatial domains). This issue of content coverage is probably most important when one's interest is in obtaining a good *g*-factor. Of the studies specifying a single factor (which typically was done when the goal was to obtain *g*-factor estimates) 70.4% included indicators that covered all three broad domains.

With respect to the issue of the method of factor analysis, the results are a bit more mixed. Despite the widely known hierarchical nature of cognitive abilities which includes a *g* factor, and explicit admonishments against the use of methods which disallow a general factor to emerge (Jensen & Weng, 1994), 14.57% of the analyses failed to use a method that allowed a general factor to emerge. This is not to say that the only purpose of factor analyzing ability tests is to extract a *g*-factor, and in fact there are many cases where researchers may legitimately be more interested in the broad ability factors. However, we find the use of methods that exclude a *g*-factor curious for several reasons. First, arguably, no other psychometric question has been scrutinized and empirically tested more than the structure of cognitive abilities: There is a *g* factor that underlies human mental functioning (Carroll, 1993; Gottfredson, 1997; Jensen, 1998; Spearman, 1904; Thurstone, 1947). Second, a recent survey of experts (Reeve & Charles, 2008) revealed a strong consensus that ability tests measure *g*, and that *g* is the most important individual difference variable. Yet, despite this consensus, almost 15% of the research published in this literature fails to use a method that would be appropriate given a theoretical model that includes *g*.

Again, while the extraction of a *g*-factor is certainly not the only purpose in factor analysis of ability data, there are several potentially serious problems that can stem from the use of an inappropriate method (or more aptly, a method that disallows an appropriate model). The use of such methods scatters the variance due to *g* across the factors that are modeled, which can lead to inaccurate inferences regarding the importance of the predictive power of narrow abilities, or fail to provide optimal estimates of a targeted ability. For example, correlations between factor scores based on models that preclude a *g*-factor are ambiguous because those scores reflect variance due to both *g* and other narrow abilities. A clear understanding of how both *g* and narrow abilities relate to criteria or other constructs requires the variance due to *g* to be separated from the variance due to narrow cognitive abilities. Failure to appropriately distinguish sources of variance can both occlude true relations and give rise to false associations (See Reeve, Meyer & Bonaccio,

2006, for example regarding intelligence-personality associations). On the other hand, these same results suggest that 85% of published studies are using a method of factor analysis that allows for the modeling of variance due to a common factor.

5.1. Limitations and directions for further research

It should be recognized that there are important constraints on the extent to which the results of this survey can be generalized. First, as noted above, the selection criteria we used place restrictions on the degree to which these results can be extended across the psychological literature. Our study only included articles using adult, non-clinical samples. The decision to exclude research based exclusively on clinical samples as well as child and adolescent samples likely excludes most of the school and clinical psychology literatures. Similarly, our choice of journals to search may have skewed some of the findings with respect to the typical abilities assessed. For example, we found that auditory abilities and long-term memory/associated retrieval appear to be assessed relatively infrequently. However, a focus on neuropsychology-oriented or speech/language-oriented journals would likely have resulted in a different finding.

It should also be acknowledged that this review has limitations with respect to our general purpose, which was to assess the current practices in conducting factor analyses of cognitive ability tests. There are several additional methodological issues that should be of concern when considering the quality of our factor analytic research. Such issues include, but are not limited to, the reliability of indicators used, sampling issues such as the range of ability in the sample, the theoretical basis of the model specified, number of factors extracted (see Frazier & Youngstrom, 2007, on this issue), and the goodness of fit of the models imposed on the data (see Jackson, Gillaspy, & Purc-Stephenson, 2009; McDonald & Ho, 2002; MacCallum, Wegener, Uchino, & Fabrigar, 1993 for existing detailed discussions of this issue). Likewise, additional "introspective" research on the science of mental abilities should consider other important methodological issues that are not specific to factor analysis. For example, as far as we know, there is no existing research on how ability researchers deal with fundamental data management issues such as whether and how researchers screen for univariate or multivariate outliers, and test for normality of distributions or range restriction. Similarly, it is unclear how researchers in the science of mental abilities typically deal with missing data points and whether these decisions significantly influence substantive results.

6. Conclusions

The results of this review are useful for several reasons. First, they provide some evidence concerning the methodological quality of our literature. Perhaps no other area of psychology is criticized as often as the science of mental abilities. Thus, the ability to demonstrate the methodological rigor of our field can help dispel some myths and inappropriate criticisms. Second, these results can provide researchers, journal reviewers, and editors with some valuable information when considering new research. For example, researchers often struggle with balancing practical limitations against methodological ideals. These results can provide them with some sense of what is (or can be) considered "typical" in the literature. Similarly, reviewers may struggle with their evaluations of issues such as whether a sample is of sufficient size, and when it can legitimately be considered outside of the norm. Often when there is no objective criterion, norms within the scientific community are used as benchmarks against which to evaluate new research. Results such as these can provide some empirical basis for such benchmarks. Finally, we hope this review spurs others to evaluate the methodological quality of our literature. Indeed, any researcher in the science of mental abilities likely understands the multitude of methodological and design issues with which we must grapple. Additional reviews concerning these and other issues will likely provide valuable insights about the quality and rigor of quantitative methods used the science of mental abilities.

References

Anderson, J. C., & Gerbing, D. W. (1984). The effect of the sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, *49*, 155—173.

Bryant, F. B., & Yarnold, P. R. (2000). Principle-components analysis and exploratory and confirmatory factor analysis. In L. G. Grimm & P. R. Yarnold (Eds.), *Reading and Understanding Multivariate Statistics* (pp. 99—136). Washington, DC: American Psychological Association.

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

Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis*, 2nd ed. Hillsdale, NJ: Erlbaum.

El Koussy, A. A. H. (1935). Visual perception of space. *British Journal of Psychology Monograph Supplement*, *7*(20).

Frazier, T. W., & Youngstrom, E. A. (2007). Historical increase in the number of factors measured by commercial tests of cognitive ability: Are we overfactoring? *Intelligence*, *35*, 169—182.

Gagne, P., & Hancock, G. R. (2006). Measurement model quality, sample, size, and solution propriety in confirmatory factor models. *Multivariate Behavioral Research*, *41*, 65—83.

Galton, F. (1883). *Inquiries into Human Faculty and its Development.* London: MacMillan.

Garnett, J. C. M. (1919). General ability, cleverness and purpose. *British Journal of Psychology*, *9*, 345—366.

Ghiselli, E. E. (1973). The validity of aptitude tests in personnel selection. *Personnel Psychology*, *26*, 461—477.

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

Gottfredson, L. S. (1997). Why *g* matters: The complexity of everyday life. *Intelligence*, *24*, 79—132.

Guilford, J. P. (1954). *Psychometric methods*, 2nd ed New York, NY: McGraw-Hill.

Guttman, L., & Levy, S. (1991). Two structural laws for intelligence tests. *Intelligence*, *15*, 79—103.

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of Management*, *21*, 967—988.

Jackson, D. L., Gillaspy, J. A., & Purc-Stephenson, R. (2009). Reporting practices in confirmatory factor analysis: An overview and some recommendations. *Psychological Methods*, *14*, 6—23.

Jensen, A. R. (1986). *g*: Artifact or reality? *Journal of Vocational Behavior*, *29*, 301—331.

Jensen, A. R. (1998). *The g factor.* Westport CT: Praeger Publishers.

Jensen, A. R., & Weng, L. (1994). What is a good *g*? *Intelligence*, *18*, 231—258.

Johnson, W., & Bouchard, T. J. (2005). The structure of human intelligence: It is verbal, perceptual and image rotation, not fluid and crystallized. *Intelligence*, *33*, 393—416.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*, 187—200.

Loehlin, J. C. (2006). *Latent variable models*, 4th Ed Mahwah, NJ: Lawrence Erlbaum Associates.

MacCallum, R. C., Wegener, D. T., Uchino, B. N., & Fabrigar, L. R. (1993). The problem of equivalent models in applications of covariance structure analysis. *Psychological Bulletin*, *114*, 185—199.

MacCallum, R. C., Widam, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*, 84—89.

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*, 181—220.

McDonald, R. P., & Ho, M-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, *7*, 64—82.
McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. Harrison's (Eds.), *Contemporary Intelligence Assessment: Theories, Tests, and Issues* (pp. 151—179). New York: Guilford Press.
McGrew, K. S. (2005). The Cattell-Horn-Carroll Theory of Cognitive Abilities: Past, Present, and Future. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary Intellectual Assessment: Theories, Tests, and Issues* (pp. 136—181). New York: Guilford Press.
Monahan, C., & Muchinsky, P. M. (1983). Three decades of personnel selection research: A state-of-the-art analysis and evaluation. *Journal of Occupational Psychology*, *56*, 215—225.
Ree, M. J., & Earles, J. A. (1991). The stability of *g* across different methods of estimation. *Intelligence*, *15*, 271—278.
Reeve, C. L. (2004). Differential ability antecedents of general and specific dimensions of declarative knowledge: More than *g*. *Intelligence*, *32*, 621—652.
Reeve, C. L., & Charles, J. E. (2008). Survey of opinions on the primacy of *g* and social consequences of ability testing: A comparison of expert and non-expert views. *Intelligence*, *36*, 681—688.
Reeve, C. L., Meyer, R., & Bonaccio, S. (2006). Intelligence-personality associations reconsidered: The importance of distinguishing between general and narrow dimensions of intelligence. *Intelligence*, *34*, 387—402.
Sackett, P. R., & Larson, J. R. (1990). Research strategies and tactics in industrial and organizational psychology. In M. D. Dunnette & L. M. Hough (Eds.), *Handbook of industrial and organizational psychology* (pp. 419—489). Palo Alto, CA: Consulting Psychologists Press.
Salgado, J. F. (1998). Sample size in validity studies of personnel selection. *Journal of Occupational & Organizational Psychology*, *71*, 161—164.
Scandura, T. A., & Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management Journal*, *43*, 1248—1264.
Schmidt, F. L., & Hunter, J. E. (1977). Development of a general solution to the problem of validity generalization. *Journal of Applied Psychology*, *62*, 529—540.
Spearman, C. (1904). General intelligence, objectively determined and measured. *American Journal of Psychology*, *15*, 201—293.
Spearman, C., & Jones, L. L. W. (1950). *Human Ability.* London: Macmillan.
Thurstone, L. L. (1947). *Multiple factor analysis: A development and expansion of The Vectors of the Mind.* Chicago: University of Chicago Press.
Vernon, P. E. (1961). *The structure of human abilities*, 2nd ed. London: Methuen.