Intelligence 37 (2009) 453–465

Contents lists available at ScienceDirect

Intelligence

The dependability of general-factor loadings: The effects of factor-extraction methods, test battery composition, test battery size, and their interactions

Randy G. Floyd [a,*], Elizabeth I. Shands [a], Fawziya A. Rafael [a], Renee Bergeron [b], Kevin S. McGrew [c]

[a] The University of Memphis, United States
[b] Hawaii Department of Education, United States
[c] Woodcock-Munoz Foundation, University of Minnesota, United States

ARTICLE INFO

Article history:
Received 7 June 2007
Received in revised form 19 May 2009
Accepted 21 May 2009
Available online 11 June 2009

Keywords:
g factor
g loadings
Factor analysis
Generalizability theory
Dependability

ABSTRACT

To understand the extent to which the general-factor loadings of tests are inherent in their characteristics or due to the sampling of tests, the number of tests in the correlation matrix, and the factor-extraction methods used to obtain them, test scores from a large sample of young adults were inserted into independent and overlapping batteries of varying sizes. Principal factors analysis, maximum-likelihood estimation, and principal components analysis yielded general-factor loadings for each test. Generalizability theory analyses revealed that the characteristics of the tests consistently contributed the largest percentage of variance. Variance attributable to the factor-extraction method and its interactions was sizeable when principal components analysis was included in the analysis but negligible when it was excluded. Psychometric sampling error produced sizeable variance components in some analyses, and its effects were magnified when test batteries diminished in size. When results from principal components analysis were excluded and when the effects of psychometric sampling error were reduced, general-factor loadings were highly dependable across varying conditions.

1. Introduction

Many modern psychometric theories of intelligence converge in agreement that the *general factor* (a.k.a., psychometric g) meaningfully represents the majority of the positive relations among specific measures of human cognitive abilities, such as scores from the tests of intelligence test batteries (Carroll, 1993; Jensen, 1998; Spearman, 1927; Sternberg & Grigorenko, 2002). Despite this general agreement, several prominent criticisms have been levied against the construct validity of the general factor. One criticism is that the general factor is dependent on the factor-analytic methods used to extract it from a matrix of correlations, and another criticism is that the general factor is dependent on the measures used to operationalize it. These criticisms seem to have become part of virulent memes that pervade the minds of many professionals and consumers of tests results.

1.1 Factor-extraction methods

Gould (1981, 1994) asserted that the type of factor-analytic methods used to extract the general factor from a matrix of correlations affected the identification of the general factor to the extent that it undermined its meaningfulness. In fact, Gould concluded that the general factor is a statistical artifact with no representation in reality. A key target of his criticism was the factor-extraction method principal components analysis, which analyzes all variance across scores (including error variance). Snook and Gorsuch (1989), B. Thompson (2004), and others have demonstrated that principal components analysis tends to produce inflated parameter estimates. In addition, principal components analysis can produce a first component (designed to represent the general factor) with uniformly positive loadings from constituent tests when correlations among at least some of those tests are weak and not significantly different than 0 (Jensen & Weng, 1994). Despite these criticisms of principal components analysis, when an appropriate correlation matrix

* Corresponding author.
E-mail address: rgfloyd@memphis.edu (R.G. Floyd).

0160-2896/$ – see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2009.05.003

is analyzed, values stemming from varying factor-extraction methods appear to be remarkably consistent—even when principal components analysis is used to extract the general factor. For example, Ree and Earles (1991) demonstrated both (a) correlations between 14 general component scores and factor scores from principal components analysis, principal factors analysis, and hierarchical factor analysis derived for each person and also (b) coefficients of congruence from general-factor loadings derived from varying factor-extraction methods that were very strong and near unity. Jensen and Weng (1994) also supported and extended these results through their comparisons of results from 6 to 10 different methods of factor analysis, including principal factors analysis, hierarchical exploratory factor analysis, and confirmatory factor analysis, using artificial correlation matrices and an archival correlation matrix. Their results indicated (a) very high correlations between factor scores between general factors, (b) very high congruence coefficients between general factor loadings yielded by different methods of analysis and the "true" general factor loadings used to derive the artificial correlation matrixes, and (c) consistency in the percentage of variance attributable to the general factor across methods. For instance, the Spearman correlations between general-factor loadings from 24 tests obtained using 10 different factor-extraction methods ranged from .79 to 1.0 (M = .91). Although general consistency across these factor-extraction methods is apparent, variation across methods is also evident.

1.2. Test battery composition and test battery size

It is sometimes argued that the general factor is dependent on the measures used to operationalize it, and it seems rational to argue, for example, that scores from a battery including a preponderance of tests of language-based abilities, when entered into factor analysis, would yield a very different general factor than when analyzing scores from a battery including a preponderance of tests of visualization abilities. Such criticism has often been levied by the most vocal opponent of the interpretation of the general factor in recent decades, John Horn (Horn, 1985, 1989; Horn & Blankson, 2005; Horn & McArdle, 2007). Consistent with this criticism, Horn has referred to the general factor and its related scores as conglomerates, mixtures measures, and hodgepodges of distinct abilities. There is some evidence of the effect of the test battery composition on the identification of the general factor. Based on Carroll's (1993) re-analysis of more than 460 data sets, he offered that "the G factor for a given data set is dependent on what lower-order factors or variables are loaded on it. One could say that a higher-order factor is 'colored' or 'flavored' by its ingredients" (p. 596). In a similar manner, Jensen and Weng (1994) conveyed, "Just as there is sampling error with respect to statistical parameters, there is *psychometric sampling error* [emphasis added] with respect to *g*, because the universe of all possible mental tests is not perfectly sampled by any limited set of tests" (p. 236).

Several studies have directly investigated this potential criticism that the general factor is dependent on the measures used to operationalize it. For example, Thorndike (1987) examined the identification of the general factor when it is formed from different samples of test scores. He employed data from 65 tests from the Army Air Forces Aviation Aircrew Classification Battery. From this battery, 48 tests were divided into six batteries of 8 tests each. The remaining 17 "probe" tests were inserted one at a time into each of the six batteries, and the general-factor loadings for each of the 17 probe tests were obtained. The median Pearson product–moment correlation coefficient between general-factor loadings across analyses using the six batteries was strong (.85). The range of correlations was .52 to .94, and two correlations were lower than .70. The average standard deviation of the general-factor loadings for the 17 tests was .07 (range = .04 to .14) across the six batteries. Based on Thorndike's results, it appears that the magnitude of the tests' loadings on the general factor is determined largely by the characteristics of the tests, rather than by characteristics of the test batteries in which they are inserted, but the influence of psychometric sampling error is apparent in the varying general-factor loadings across the test batteries.

More recently, others have examined the relations between second-order general factors extracted from varying test batteries using confirmatory factor-analytic methods and maximum-likelihood estimation, and they have demonstrated relations between these general factors that are consistently very near unity. Keith, Kranzler, and Flanagan (2001) produced a correlation of .98 between general factors derived from scores from two individually administered intelligence test batteries formed by 12 and 18 tests. Johnson, Bouchard, Krueger, McGue, and Gottesman (2004) produced correlations of .99, .99, and 1.0 between general factors formed from each of three test batteries formed by 11 to 17 tests. Most recently, Johnson, te Nijenhuis, and Bouchard (2008) produced correlations from .77 to 1.0 between general factors from each of five test batteries formed by 4 to 13 tests. Of these correlations, 7 of 10 were .95 or higher.

These results indicate similar identification of the general factor across independent test batteries. However, it is evident that disproportionate sampling of tests, biased toward specific abilities, may not allow specific variances to "average out" and for common variance, attributable to the general factor, to remain as the primary source of variance. It is logical that, as the number of test scores included in the factor analysis diminishes, the greater the effects that psychometric sampling error will have on the general factor and resulting scores. For example, in Johnson et al. (2008), the general factor formed from only 4 tests demonstrated notably lower correlations with the other general factors (M = .85) than did all of the other general factors with each other (M = .95). In addition, as indicated by the Wilks theorem (1938), as the number of test scores included in the factor analysis increases, the relations between the general factors derived from independent test scores will be strengthened (Jensen & Weng, 1994).

1.3. Purpose of the study

We sought to understand better the strength of and interactions between the effects of the factor-extraction method, the composition of the battery, and an understudied influence, the number of tests in the battery, on the identification of the general factor as well as to determine how these effects compare to differences in characteristics of