JSLHR

Research Article

Simulating the Effects of Common and Specific Abilities on Test Performance: An Evaluation of Factor Analysis

Dennis J. McFarland[a]

Purpose: Factor analysis is a useful technique to aid in organizing multivariate data characterizing speech, language, and auditory abilities. However, knowledge of the limitations of factor analysis is essential for proper interpretation of results. The present study used simulated test scores to illustrate some characteristics of factor analysis.
Method: Linear models were used to simulate test scores that were determined by multiple latent variables. These simulated test scores were evaluated with principal components analysis and, in certain cases, structural equation modeling. In addition, a subset of simulated individuals characterized by poor test performance was examined.
Results: The number of factors recovered and their identity do not necessarily correspond to the structure of the latent variables that generated the test scores. The first principal component may represent variance from multiple uncorrelated sources. Practices such as correction or control for general cognitive ability may produce misleading results.
Conclusions: Inferences from the results of factor analysis should be primarily about the structure of test batteries rather than the structure of human mental abilities. Researchers and clinicians should consider multiple sources of evidence to evaluate hypotheses about the processes generating test results.

Key Words: factor analysis, human abilities, general and specific processing disorders

Box and Draper (1987, p. 424) once observed, "Essentially, all models are wrong, but some are useful."

Researchers in audiology, speech, and language have frequently used factor analysis to organize data from multiple measures of individual abilities. For example, Watson and colleagues (2003) created factor scores to evaluate effects of sensory, cognitive, and linguistic factors on academic performance. Semel, Wiig, and Secord (2003) used factor analysis to provide support for the validity of the Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF–4), an instrument widely used to test language abilities. Domitz and Schow (2000) reported results of a principal components analysis that produced a four factor solution that they describe as closely linked to aspects of auditory processing disorders as defined by the American Speech-Language-Hearing Association (1996). Results of factor analysis were an important aspect of each of these investigations.

A variety of methods are used in factor analysis, such as principal components analysis, principal factors analysis, and structural equation modeling. Both principal components and structural equation modeling are used to summarize covariance matrices. Often these covariance matrices are normalized to be correlation matrices. Consider the correlation matrix produced by 12 tests, which will produce 66 unique correlations. Principal components and structural equation modeling provide a means of summarizing the complex series of relationships between these variables. These methods are often referred to as exploratory and confirmatory factor analysis. Actually these two methods differ not so much as in why they are used but rather because principal components analysis is a linear transformation, and structural equation modeling is a form of statistical modeling, as described below.

Principal components analysis is an approach to factor analysis that performs a linear transformation on a set of variables to produce a new set of orthogonal (i.e., uncorrelated) components. The first principal component accounts for the maximum amount of the covariance in the

[a]Wadsworth Center, New York State Department of Health, Albany
Correspondence to Dennis J. McFarland: mcfarlan@wadsworth.org
Editor: Craig Champlin
Associate Editor: Jeffrey Weihing
Received June 28, 2013
Revision received November 1, 2013
Accepted March 1, 2014
DOI: 10.1044/2014_JSLHR-H-13-0166

Disclosure: *The author has declared that no competing interests existed at the time of publication.*

covariance matrix that is possible by linear transformation. The second principal component accounts for the maximum amount of the covariance remaining in the residual matrix. Each additional component is extracted in a similar manner from the current residual until as many factors are extracted as there are variables that produced the covariance matrix. This process is reversible prior to the elimination of "minor" factors. When minor factors are eliminated by some criterion, the original covariance is now described in a more compact manner. The analyst may then interpret each component on the basis of its loadings on the original variables.

In structural equation modeling, the analyst specifies one or more models that might potentially account for the observed covariance matrix. As a rule, this specification includes identification of the latent factors that produce a given observed score, whereas the actual weighting of this influence is estimated by some optimization algorithm. Because this process requires that there are more degrees of freedom in the data than in the model, the model is overdetermined and can be evaluated by statistical tests of how well it accounts for the data. Structural equation modeling is useful for evaluating alternative models because these are specified by the analyst and the procedure generates fit indices that can be compared.

There are limitations to the interpretation of the results of factor analysis. The interpretation of any single factor solution is problematic given factor indeterminacy (MacCalum, Wegener, Uchino, & Fabrigar, 1993). Factor indeterminacy refers to the fact that many alternative models may account equally for the observed data. In addition, a recent simulation study by the author (McFarland, 2012) showed that, under certain conditions, factor analysis may not be able to separate orthogonal latent sources of individual differences that were the basis of test scores.

Simulations of test batteries can provide useful insights into the nature of analysis procedures. For example, simulations have been used to compare methods to determine the number of factors to extract (Ruscio & Roche, 2012; Zwick & Velicer, 1986) and are commonly used with structural equation modeling to evaluate fit indices (Bentler, 1990; Sterba & Pek, 2012). However, prior simulations of factor analysis have generally been restricted to models with simple structure (de Winter & Dodou, 2012; Velicer & Fava, 1998). Simple structure refers to the case where only a single latent factor loads on any given observed variable. This is true when each test measures only one underlying factor; it is a common way to interpret test performance. Thurstone (1940) proposed simple structure as a solution to the factor indeterminacy problem and considered it to be a reasonable assumption. However, we have previously argued that tests of cognitive abilities can potentially be determined by multiple factors (Cacace & McFarland, 2005; McFarland & Cacace, 1995).

In the present simulation study, I was concerned with how the results of principal components analysis relate to the underlying structure of simulated data. Simulated data can be created with simple linear models. A linear model of test performance was first proposed by Hart and Spearman (1912). In its simplest form, individual variation in test performance t can be expressed as

$$t = c + u, \tag{1}$$

where c represents a component common to the collection of tests in question and u represents the test-unique component. Hart and Spearman viewed this common component to be g, or general intelligence. More recent linear models of test performance distinguish between common components and specific components. Here, individual variation in test performance can be expressed as

$$t = c + s + u, \tag{2}$$

where the term s represents a broad source of variation associated with a subset of the collection of tests in question (i.e., specific abilities). A currently popular model of abilities (Carroll, 1993) holds that individual performance on tests of cognitive ability are a result of a single common factor and individual differences in broad factors such as auditory processing, visual processing, working memory, and speed of information processing. Although models of test performance often assume that the data can be accounted for by a single common factor and one specific factor as in equation 2, for any given test, more complex models are possible, as for example, where individual variation in test performance can be expressed as

$$t = c_1 + c_2 + s_1 + s_2 + u, \tag{3}$$

where c_1 and c_2 represent two uncorrelated sources of individual differences in test performance common to all tests under consideration. Likewise s_1 and s_2 represent two uncorrelated sources of individual differences common to subsets of the tests in question (note that s_1 and s_2 may not necessarily co-occur on the same subset of tests).

In the present simulation study, I will explore the extent to which principal components analysis recovers the uncorrelated sources of variation that were used to generate hypothetical test performance. Several models, including common and specific level factors, will be considered. The results show that principal components analysis does not always recover as many uncorrelated factors as were used to simulate the data and that uncorrelated specific factors tend to be associated as a "contrast" in a single factor. Indeed, factor membership is determined by the correlation between test loadings across the latent traits rather than the structure of individual differences. Like any study using simulations, these results are limited by the scope of the models examined. However, these results do show that factor analysis needs to be supplemented with theory and the results of research using other methods. The implications of these results for the interpretation of results in speech, language, and auditory research will be discussed.

Method

All simulations were done in SAS with a C++ program used to organize the data. The basic model for the *kth* score on the *ith* test was

$$t_{ik} = \sum (w_{ij} \,.\, a_{jk}) + e_{jk}, \quad (4)$$

where a_{jk} is the magnitude of the *jth* ability for the *kth* observation, and e_{jk} is a random test-specific term. The value of w_{ij} is the weight given a_{jk} on the *ith* test. The value of a_{jk} was unique to each individual within a test battery simulation and was drawn from the SAS normal distribution function. The value of a_{jk} represents the ability of an individual on some hypothetical trait (e.g., an individual's general intelligence or auditory processing ability), whereas the value of w_{ij} describes the role of these abilities in determining test performance (e.g., to what extent a test measures general intelligence or auditory processing).

Several simulations of single test batteries were conducted with values of w_{ij} set to 1 for common effects and 2 for specific effects in order to illustrate some basic effects in simple models. Next, simulations of models produced with different values of w_{ij} for each the of multiple test battery simulations were done in order to extend the generality of findings. Each value of w_{ij} was unique to a single test battery simulation and was drawn from the SAS uniform distribution function. Use of a uniform distribution ensures that all abilities function in a similar manner within a given test battery simulation (i.e., if a given ability has positive effects on one test, it would be expected to have positive effects on other tests). This is a boundary condition for all of the simulations conducted in the present study and was more extensively investigated by McFarland (2012).

Each simulation involved generating scores for 1,000 subjects on each of the hypothetical tests. These simulations were done with SAS (SAS, 2010). The resulting data were then analyzed with the SAS FACTOR procedure and in some cases also with the SAS CALIS procedure.

Results

The first simulated example is of two test scores generated by one common latent variable and test-specific error. This can be represented as

$$t_{1k} = w_{11} \,.\, a_{1k} + e_{1k} \quad (5)$$

and

$$t_{2k} = w_{12} \,.\, a_{2k} + e_{2k}, \quad (6)$$

where t_{1k} represents the score for the *kth* individual on test 1, and t_{2k} represents the score for this same individual on the second test. The values of each w in this case are 1, and the values of each a and e are drawn from a normal distribution with a M of 0 and a variance of 1. The results of a principal components analysis of a simulation of 1,000 cases using Equations 5 and 6 are presented in Table 1. The results show that this simple case produces a factor that is defined by the sum of the two tests with an eigenvalue of 1.50 and a factor that is defined by the difference between the two tests with an eigenvalue of 0.50. The correlation between these two simulated tests was 0.50. It can be shown analytically that two correlated variables always produce a principal components solution consisting of the sum and the difference between the two variables (Harris, 1985). The sum will have the larger eigenvalue for a positively correlated pair, and the difference will have the larger eigenvalue for a negatively correlated pair. However, cases with more variables do not have this deterministic form.

This two-variable case illustrates the fact that principal components analysis is a reversible linear transformation. This linear transformation can be considered a rotation of the axis onto which the data are projected, as illustrated in Figure 1. We could recover the original data by simply computing half the sum and difference of the new factor scores. However, principal components analysis is commonly used as a data-reduction technique. Although criteria for omitting components have been discussed extensively (Zwick & Velicer, 1986), a common rule of thumb is to eliminate factors with an eigenvalue less than 1. In the case of the results shown in Table 1, this would mean eliminating factor 2 (i.e., the difference). With the elimination of factor 2, the transformation is no longer reversible.

Consider the case in which the two variables in our first example represent the height and weight of individuals. These two variables are positively correlated, and their sum would be the major factor in a principal components analysis. The difference would represent the extent to which weight is not accounted for by height (and vice versa). This example illustrates another important aspect of principal components analysis. The transformation does not necessarily identify clinically significant relationships. For example, if one were interested in obesity, then the excess in weight not predicted by an individual's height would be of interest. This would best be captured by the second factor in our first example, a factor that was discarded by the eigenvalue less than 1 rule.

The next simulated example is of two test scores generated by two common latent variables and test-specific error. This can be represented as

$$t_{1k} = w_{11} \,.\, a_{1k} + w_{21} \,.\, a_{2k} + e_{1k} \quad (7)$$

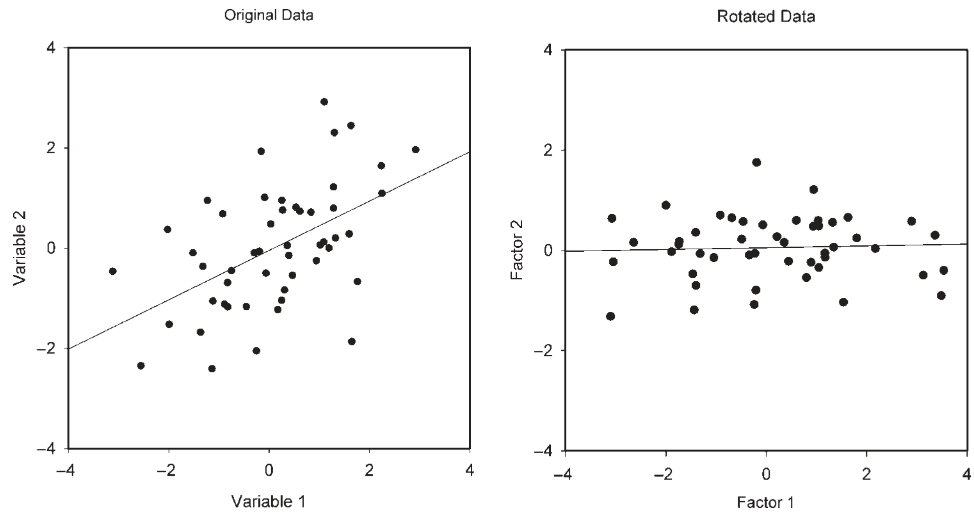
and

$$t_{2k} = w_{12} \,.\, a_{1k} + w_{22} \,.\, a_{2k} + e_{2k}. \quad (8)$$

Table 1. Simulation of two tests with a common factor.

Variable	Factor 1	Factor 2
Test 1	.8672	.4979
Test 2	.8672	–.4979
Eigenvalue	1.5041	0.4959

Figure 1. Illustration of axis rotation with the data from the two correlated variable example. The panel on the left shows the original simulated data (only the first 50 points). The panel on the right shows these same data transformed by the factor weights shown in Table 1. Note that the data are rotated so that they are no longer correlated on the new rotated axis. Also note that the variance along the axis of the first factor is greater than that along the axis of the second factor. This would not be the case for factor scores, which would be normalized by their respective eigenvalues.



This notation is similar to the first example with the exception that a second latent variable has been added. The results of a principal components analysis of a simulation of 1,000 scores using equations 6 and 7 are presented in Table 2. Note that the results are very similar to the first case and that only one factor would be retained by the eigenvalue greater than 1 rule. The results of this second example were obtained with data simulated with two common factors associated with individual differences in performance rather than just one as in the first case. The fact that the results are nearly identical illustrates the potential ambiguity in the results of factor analysis. If the results reflected the structure of abilities in individuals, then we would expect to have extracted two factors of roughly equal size representing a_1 and a_2 in equations 7 and 8. Instead, a single factor was produced that reflects the fact that a_1 and a_2 have identical loadings on the two tests. Thus, the results reflect the relationship between the tests (i.e., they measure the same thing) rather than the nature of the underlying abilities (i.e., they are each composed of two independent sources of variance).

The next example represents a simulation of 12 test scores generated from two common latent variables and two specific latent variables. Although these simulated latent variables are abstract and need not represent anything specific, one could think of them as representing common abilities, such as speed of information processing and freedom from distraction, and specific effects, such as auditory and visual processing abilities, or alternatively, receptive and expressive language abilities. For each of the 12 tests, the loadings on these latent variables as well as the loadings on the four factors extracted by principal components analysis are shown in Table 3. Note that the general effects have weights of 1 and the specific effects have weights of 2 so that their contributions are equivalent. The

Table 2. Two tests with two independent common factors.

Variable	Factor 1	Factor 2
Test 1	.8128	.5825
Test 2	.8128	–.5825
Eigenvalue	1.3214	0.6786

Table 3. Twelve tests simulated with two common factors and two independent specific factors.

Variable	C_1	C_2	S_1	S_2	Factor 1	Factor 2	Factor 3	Factor 4
Test 1	1	1	2	0	.6619	–.4676	–.2540	.1462
Test 2	1	1	2	0	.6643	–.4967	.1034	–.1840
Test 3	1	1	2	0	.6719	–.4549	.0416	.0929
Test 4	1	1	2	0	.6911	–.4450	–.1183	–.1199
Test 5	1	1	2	0	.6527	–.4946	.0159	.1977
Test 6	1	1	2	0	.6576	–.4777	.2102	–.1315
Test 7	1	1	0	2	.6555	.4834	–.0446	.1363
Test 8	1	1	0	2	.6375	.4880	–.3520	.0725
Test 9	1	1	0	2	.6466	.4752	–.2512	–.3806
Test 10	1	1	0	2	.6543	.4806	.1442	.3576
Test 11	1	1	0	2	.6446	.4908	.1442	–.1871
Test 12	1	1	0	2	.6461	.5006	.2843	.0010
Eigenvalue					5.1822	2.7632	0.4750	0.4661

Note. C_1 and C_2 are common latent variables with loadings on all tests (shown in the second and third columns). S_1 and S_2 are specific latent variables with loadings on subsets of tests (shown in columns 4 and 5). Loadings on the four factors and their eigenvalues obtained from principal components analysis are shown in columns 6–9.

last row of Table 3 shows the eigenvalues for these four factors. Applying the eigenvalue less than 1 rule would give a solution that represents these four latent variables in terms of two factors, the first being a single common factor and the second a "contrast" between the two specific factors. These same two factors would be retained if these eigenvalues were used to make a scree plot (Cattell, 1966; i.e., a plot used to select the point at which the rate of change in the eigenvalues becomes less steep). Table 4 shows the results of a varimax rotation. The varimax operation involves rotation of the principal components so as to maximize the variance of the weights of each factor (Harris, 1985), an operation that is thought to aid interpretation of the factors. As can be seen in Table 4, the first two varimax factors better represent the specific latent variables at the expense of the representation of the general latent variables. The last two factors do not correspond to any of the latent variables that generated the data. Allowing the factors to correlate using the obvarimax option in SAS produced factor loadings that were less coherent (data not shown).

The first two factors shown in Table 3 were next used to generate factor scores that were then correlated with the latent common and specific abilities used to generate the test scores. This provides a means of visualizing how the latent variables used to generate the simulated test scores relate to the extracted factors. The resulting correlations between the two extracted factors and the latent variables that generated the test scores are presented in Table 5. As can be seen in Table 5, all four of the generative latent variables correlated positively to an approximately equal extent with the first factor. The first factor thus represents a composite of all of the latent variables. This is to be expected since the criterion for determining the weights of the first principal component is to account for the maximum amount of variance in the pattern of test correlations. Only the specific latent variables correlated with the second factor, and this correlation was in opposite directions for the two. The second factor thus represents a "contrast" between the two specific latent variables. Once again, the results illustrate that the results of the principal components analysis reflect the relation between the tests rather than the structure of the abilities that the tests measure. In this case, all of the tests measure both of the common factors, so they all load on the first component. Tests that measure the first specific latent variable never measure the second specific latent variable so these two groups of tests load on the second latent variable with opposite signs.

Table 4. Twelve tests simulated with two common factors and two independent specific factors.

Variable	C_1	C_2	S_1	S_2	Factor 1	Factor 2	Factor 3	Factor 4
Test 1	1	1	2	0	.7971	.1030	.0757	.3016
Test 2	1	1	2	0	.8232	.0825	.1211	–.1826
Test 3	1	1	2	0	.8004	.1559	–.0508	.0339
Test 4	1	1	2	0	.8030	.1129	.2128	.0339
Test 5	1	1	2	0	.8144	.1297	–.1271	.1167
Test 6	1	1	2	0	.8075	.1175	.0201	–.2363
Test 7	1	1	0	2	.1351	.7968	.1116	.1356
Test 8	1	1	0	2	.1130	.7280	.3374	.3423
Test 9	1	1	0	2	.1271	.6611	.6313	–.0107
Test 10	1	1	0	2	.1411	.8629	–.1710	.1180
Test 11	1	1	0	2	.1258	.7896	.1719	–.3217
Test 12	1	1	0	2	.1202	.8207	.0635	–.1584
Eigenvalue					4.0113	3.7285	0.6725	0.4742

Note. C_1 and C_2 are common latent variables with loadings on all tests (shown in the second and third columns). S_1 and S_2 are specific latent variables with loadings on subsets of tests (shown in columns 4 and 5). Loadings on the four factors and their eigenvalues obtained from varimax rotation of the principal components analysis presented in Table 3 are shown in columns 6–9.

Table 5. Correlations between extracted factors and the generative variables.

Factor	C_1	C_2	S_1	S_2
1	.4817*	.4741*	.5060*	.4647*
2	–.0195	–.0147	–.6429*	.6755*

*$p < .0001$.

Table 5 shows that Factor 1 and Factor 2 both correlate with the specific latent variables, although Factor 2 correlates somewhat higher. However, in studies of human participants, the underlying latent variables are never actually known. Rather, they are estimated from test scores. To simulate the process of identifying individuals with poor performance on a specific factor, the average of the first six simulated tests was computed, as these all have loadings on the first specific latent variable. Then a group was formed that had scores at least two standard deviations below the mean (n = 63). The first 63 simulated scores not meeting this criterion were selected as a control group. To give this abstract experiment some context, these groups could be thought of as representing individuals with poor performance on auditory processing tasks and corresponding controls. Multiple regressions were then run with Factor 1 and Factor 2 used to predict group membership. Considered in isolation, Factor 1 had an r^2 value of .59, whereas Factor 2 had an r^2 value of .28. In a hierarchical model, Factor 2 had an r^2 value of .13 when it was corrected for Factor 1 scores (i.e., the variance unique to Factor 2). This process was also repeated using a cutoff of one standard deviation, producing groups of 179 simulated individuals. For this one standard deviation case, Factor 1 had an r^2 value of .47, whereas Factor 2 had an r^2 value of .27. In a hierarchical model, Factor 2 had an r^2 value of .20 when it was corrected for Factor 1 scores. Although all of these correlations are significant, the conclusion from this analysis would be that the individuals selected for on the basis of these tests differ mainly in terms of the general factor.

Table 6 shows fit statistics for several models evaluated with the simulated 12 test data using structural equation modeling (sometimes referred to as confirmatory factor analysis). The models considered are the null model, a model with two common and two specific factors (the

Table 6. Fit statistics for structural models of twelve tests simulated with two common factors and two specific factors.

Model	χ^2	*df*	*p*	AGFI	RMSEA	Akaike
NULL model	6405.6366	66	<.0001	.2105	.3101	6,273.6366
Two common and two specific	26.7236	30	<.6377	.9884	.0000	–33.2764
Two common	40.1923	42	<.5506	.9875	.0000	–43.8077
Two specific	143.8296	54	<.0001	.9672	.0408	35.8296
Two specific hierarchical	46.6796	50	<.6074	.9878	.0000	–53.3204
Two specific correlated	46.6796	53	<.7967	.9885	.0000	–59.3204

Note. AGFI = adjusted goodness-of-fit index; RMSEA = root-mean-square error of approximation; Akaike = Akaike information criterion (Akaike, 1974).

true generative model), a model with two common factors, a model with two specific factors, a model with two specific factors and a second order (i.e., hierarchical) common factor, and an oblique model with two specific factors. All specific factors corresponded to those used to generate the data. The hierarchical and oblique models are commonly used in modeling the correlations between tests of abilities. The null model includes only test-unique variance (sometimes referred to as error) and does not include any common or specific factors. The null model thus models only the variance of the tests and none of the covariance between tests. It serves as a basis for estimating goodness of fit.

Table 6 shows that the true generative model accounts for the most variance, as indicated by the lowest residual chi-square value. The two common factors model is a close second and actually is considered a better fit on the basis of Akaike's information criterion (Akaike, 1974). The Akaike information criterion takes into account goodness of fit via the log likelihood while penalizing for more complex models. Thus, because the two common factors model has fewer degrees of freedom (i.e., the residual *df* in Table 6 is larger) it is penalized less. The two common factors model has a pattern of weights very much like those produced by the principal components analysis shown in Table 3 with all tests having similar positive weights on the first factor and the second factor being a contrast between the two specific latent generative variables. Apart from the chi-square values, the two oblique common factor model has the best overall fit statistics. Only the null model and the two specific factors model have chi-square values significantly different from chance. Furthermore, all of the models (with the exception of the null) would be considered excellent fits by conventional criteria.

The results of the analysis of the simulated two common and two specific latent variable data thus show that principal components analysis does not necessarily uncover the true pattern of factor loadings nor does structural equation modeling necessarily provide an unambiguous basis for discriminating between several potential models. However, it could be that this example with equal weights for all latent variables is a special case. In order to establish the generality of these results, the next example considered 50 simulations where the weights determining the loading of abilities on tests were drawn from the SAS uniform distribution. This distribution produces values on the uniform distribution ranging from 0 to 1 (weights for specific factors were multiplied by 2). A unique random value was assigned to the weight of each latent variable that was constant throughout a given simulation but was unique for each of the 50 simulations.

For every one of the 50 simulations, the SAS principal components procedure recovered two factors by the eigenvalue greater than 1 criterion. Table 7 shows the average correlations of the latent factors with the factor scores. In some cases, the sign of the weights for the second component were reversed, so the common and specific weights were reordered in these cases so as to always present the positive value first. As can be seen in Table 6, these simulations produced a result very similar in nature to the example with equal weights for each latent variable. The results indicate that in this general case, principal components analysis does not recover the correct number of factors nor does it identify the proper factor loadings.

Discussion

As previously reported (McFarland, 2012), principal components analysis does not always recover the number of factors that correspond to the number of unique sources of individual differences. The pattern of results is related to the way the latent variables project onto the test scores rather than to the structure of individual differences. The results of principal components analysis should correspond to the structure of abilities if simple structure is a reasonable model of test performance (i.e., each test measures only a single aspect of an individual's abilities). However, this is not the case if a multivariate model of individual test performance is appropriate, as in the examples presented in the present study.

Concerns about the interpretation of the results of factor analysis are by no means new. For example, Eysenck

Table 7. Average (n = 50 simulations) correlations between factor scores and generative latent variables.

Factor	C_1	C_2	S_1	S_2
1	.4687	.4864	.4774	.4644
2	.0036	–.0151	.6355	–.6369

(1952) cautioned against viewing the results of factor analysis as absolute. Rather, he suggested that factor analysis might provide hypotheses that require further testing. Likewise, Overall (1964) provided several examples demonstrating problems with applying simple structure given that scores were generated by more complex models. Both Eysenck and Overall viewed factor analysis as a useful tool provided that researchers understood the limitations of this method.

It should be noted that the present results were obtained with models that had multiple latent variables determining simulated test performance. Furthermore, these latent variables had consistent effects across tests (i.e., larger latent variable values produced larger test scores). Prior simulation studies of factor recovery have produced more favorable results using simple models (e.g., Velicer & Fava, 1998; Ximénez, 2007). Velicer and Fava (1998) described these simple models as orthogonal patterns (i.e., only one factor with nonzero loadings per simulated test). They rationalized their restriction to simple structure by (a) suggesting that "most researchers try to achieve nearly orthogonal patterns whenever possible" (p. 235), (b) asserting that investigation of complex patterns would require an extremely large study, and (c) admitting that they could not generate a reasonable hypothesis about how complex patterns would affect results. The present study deals with only a limited number of examples. However, these examples do show how the results of factor analysis can be limited if the processes generating the data are complex.

In addition to examining more complex structures, one of the innovations of the present study was the computation of factor scores that could be correlated with the original latent variables that generated the test scores. In this way, the contribution of the generative latent variables to the extracted factors could be determined. The results showed that principal components may consist of multiple independent sources of individual variation. In addition, it allowed analysis of a subset of simulated subjects selected on the basis of poor test scores. The results showed that correcting for general abilities could produce misleading results.

It is important to remember that the criterion for optimizing the weights of the first principal component is to maximize the amount of covariance accounted for, rather than to detect the "true" structure of individual differences. The subsequent components are then extracted from the residual using the same criteria. With the simulations in the present study, we know the nature of the models generating the data. But with test data collected in actual human subjects this is not the case. There may not be any mathematical procedure that automatically extracts the true underlying dimensions of test performance. In fact, the assumption that there is in fact a "true" model may be a fiction (Chatfield, 1995). Rather, the concern should be on accuracy and the ability of our models to generalize to a wide variety of different circumstances (Foster, 2000).

The results show that principal components analysis produces factors that group tests together that measure similar traits or abilities. As shown in the present study and elsewhere (McFarland, 2012), this does not mean that these components necessarily reflect a single trait or ability. It is entirely possible that a given component reflects the influence of several independent abilities that determine performance on the tests that have large loadings on that component. Although it is a common practice to assume that each test measures only one trait or ability, this is by no means a universally held proposition (McFarland & Cacace, 1995).

On the basis of results obtained with factor analysis, Canivez and Watkins (2010) suggested that test batteries such as the Wechsler Adult Intelligence Scales—Fourth Edition (Wechsler, 2008) should be interpreted primarily at the level of general intelligence. This view holds that there is a single factor that accounts for most of the variation in human mental abilities. Given this point of view, it would seem reasonable to correct for or control for IQ when evaluating test battery performance as a predictor of target disorders. For example, the potential role of general cognitive abilities in diagnosis of disorders such as specific language impairment (SLI) is a common concern (e.g., Vugs, Cuperus, Hendriks, & Verhoeven, 2013). Nonverbal IQ is generally considered to be an exclusionary criterion for SLI (e.g., Karasinski & Weismer, 2010; Mainela-Arnold, Evans, & Coady, 2010). However, both the tests designed to identify specific deficits, such as SLI, and measures of general cognitive ability may be influenced by multiple independent factors. Individuals with specific deficits (e.g., SLI) may have difficulties associated with only one or a few of these. Isolating the specific factor(s) responsible for these difficulties may be very important for diagnosis and treatment. Yet this may be precluded by an approach that corrects and/or controls for general cognitive ability.

As shown in Tables 5 and 7 of the present study, principal components analysis can be a greedy algorithm, extracting variance from multiple uncorrelated sources. In these examples, the first principal component correlated almost equally with each of the four uncorrelated abilities that were used to generate the simulated test scores. If, for example, the specific abilities represent verbal and nonverbal skills, then controlling or correcting for the general component would eliminate a substantial amount of the influence of these specific abilities.

A fundamental issue here is whether test performance is best conceptualized as being determined by a single underlying latent variable or whether performance is determined by multiple influences. Experimental studies have shown that test performance can often be determined by multiple effects. For example, on the basis of dual-task interference, McFarland and Cacace (1997) showed that nonverbal auditory and visual patterns were influenced both by general factors and modality-specific factors. Milberg, Hebben, and Kaplan (2009) described a number of distinct processes that might limit performance on single subtests of the Wechsler Adult Intelligence Scales—Revised (Wechsler, 1981). This view is supported by examination of the impact of different patterns of brain pathology on test performance.

Conti-Ramsden, St. Clair, Pickles, and Durkin (2012) described different patterns of development of verbal and nonverbal skills in individuals with SLI. These examples illustrate that various experimental designs may allow for clarification of problems not easily resolved by factor analysis alone. Studies involving contrasts between well-defined characteristics of stimuli (e.g., auditory and visual) or responses (e.g., verbal and nonverbal) are more likely to meet this goal than studies concerned with abstract psychological constructs that are more difficult to characterize with specific test items (McFarland & Cacace, 2012).

We will next consider examples of how the present results might apply to understanding cases where factor analysis has been used in the speech, language, and hearing literature. As noted earlier, Watson et al. (2003) used factor analysis to organize data on the effects of sensory, cognitive, and linguistic factors on academic performance. They concluded that their speech processing factor accounted for less than 1% of the variance in reading achievement, a finding widely sited by others (e.g., Ferguson, Hall, Riley, & Moore, 2011; Kamhi, 2011). This 1% value was based on a multiple regression analysis predicting reading achievement scores from each of the four factors obtained by principal component analysis followed by varimax rotation. The resulting speech-processing factor had largest loadings on the subscales of the SCAN–3 for Children: Tests for Auditory Processing Disorders (SCAN–3:C; Keith, 2009). Table 4 from Watson et al. indicates that the correlation between the SCAN–3:C composite score and reading achievement was 0.34, which would account for 11% of the variance. Thus, the latent factor defined in part by the scales of the SCAN–3:C test was a much poorer predictor than the actual SCAN–3:C test scores. Whether the factor score or the raw SCAN–3:C scores are most appropriate is difficult to determine by factor analysis alone. For example, it is possible that the SCAN–3:C scales are contaminated by variance due to supramodal processes (McFarland & Cacace, 1995). Alternatively, as the results of the present study demonstrate, one cannot assume that the principal components analysis necessary corresponds to the true structure of human abilities. Unfortunately, Watson and colleagues did not provide the actual correlations between the test scores that they included in their principal components analysis. These correlations represent the actual data. The factor analysis presented by Watson et al. represents one of many possible interpretations.

The CELF–4 is one of the most frequently used test batteries for the diagnosis of SLI (Betz, Eickhoff, & Sullivan, 2013). The test manual describes the results of factor analysis as support for the validity of this test (Semel et al., 2003). The authors describe the results of previous exploratory analyses of the CELF–3 as providing evidence for an overall skill they call general language ability. They assert that receptive and expressive domains also exist that cannot be separated from each other. As the results of the present study show, this inseparability may be due to the method of analysis rather than the nature of human language abilities. The CELF–4 manual (Semel et al., 2003) describes hierarchical structural models associated with each of several age ranges. There are five factors used to model seven tests in separate models for each age group at 8 years old, 9 years old, and 13–21 years old. There are five factors used to model six tests for the 10- to 12-year-old group. The factors for each of these four models are arranged in a three-level hierarchy. However, no information is provided as to whether these separate models provide a better fit than simpler models. This is particularly striking given that there are so many factors used to model so few tests. Because correlation matrices for each of these age groups are not provided in the test manual, evaluation of alternative models is not so easily investigated. These issues are of importance because factor analysis provides part of the evidence for the validity of this commonly used test of language abilities (Semel et al., 2003).

In the previous two examples, a single model was evaluated by investigators. The results of Watson et al. (2003) have been widely cited as evidence for the role of auditory factors in school performance, and the results of Semel et al. (2003) provide evidence for the validity of a commonly used test of language skills. Our next example considers the factor analyses reported on a battery of tests of auditory processing skills. Domitz and Schow (2000) reported that results of a principal components analysis followed by an oblique rotation produced a four-factor solution closely linked to aspects of auditory processing defined by the American Speech-Language-Hearing Association (1996). Schow et al. (2000) applied structural equation modeling to these same data and concluded that the results reinforced the four-factor model with some revision. Because Domitz and Schow published their correlation matrices, McFarland and Cacace (2002) were able to evaluate alternative models of these data. McFarland and Cacace found essentially identical fit indices for the Schow et al. model and a model consisting of a single general factor and four test-specific factors. These test-specific effects were the result of including scores for both left and right ears on the same behavioral test. This example illustrates how there are multiple interpretations of the data that are equivalent in terms of fit indices. They also illustrate the importance of publishing the correlation matrices, which are the actual data, in addition to the results of a factor analysis, which is one of many possible interpretations.

The present study does not prove that factor analysis cannot identify useful models. It only points out certain possibilities. Studies using simulations are limited by the scope of the models examined. However, the present study does show that factor analysis needs to be supplemented with theory and the results of research using other methods. Methods such as principal components analysis are useful for identifying possible ways in which to organize multivariate data. The resulting organization is based on the relationship between tests rather than the structure of human abilities. Structural equation modeling is useful for comparing the fit of alternative models. As is also the case for fitting the psychometric function, several models may be indistinguishable on the basis of fit indices alone (Treutwein & Strasburger, 1999). These methods should not be applied in a "cookbook" fashion and given an absolute interpretation. Rather, they should be used creatively and considered as one part of the total evidence for the utility of the constructs they suggest. The most useful model of human abilities should be consistent with many sources of evidence with the ultimate criterion being clinical utility.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19,* 716–723.

American Speech-Language-Hearing Association. (1996). *Central auditory processing: Current status of research and implications for clinical practice* [Technical report]. Available from www.asha.org/policy

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107,* 238–246.

Betz, S. K., Eickhoff, J. R., & Sullivan, S. F. (2013). Factors influencing the selection of standardized tests for the diagnosis of specific language impairment. *Language, Speech, and Hearing Services in Schools, 44,* 133–146.

Box, G. E. P., & Draper, N. R. (1987). *Empirical model-building and response surfaces*. New York, NY: Wiley.

Cacace, A. T., & McFarland, D. J. (2005). The importance of modality specificity in diagnosing central auditory processing disorder. *American Journal of Audiology, 14,* 112–123.

Canivez, G. L., & Watkins, M. W. (2010). Investigation of the factor structure of the Wechsler Adult Intelligence Scale—Fourth edition (WAIS–IV): Exploratory and higher order factor analysis. *Psychological Assessment, 22,* 827–836.

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

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

Chatfield, C. (1995). Model uncertainty, data mining and statistical inference. *Journal of the Royal Statistical Society, Series A, 158,* 419–466.

Conti-Ramsden, G., St. Clair, M. C., Pickles, A., & Durkin, K. (2012). Developmental trajectories of verbal and nonverbal skills in individuals with a history of specific language impairment: From childhood to adolescence. *Journal of Speech, Language, and Hearing Research, 55,* 1716–1735.

de Winter, J. C. F., & Dodou, D. (2012). Factor recovery by principal axis factoring and maximum likelihood factor analysis as a function of factor pattern and sample size. *Journal of Applied Statistics, 39,* 695–710.

Domitz, D. M., & Schow, R. L. (2000). A new CAPD battery—Multiple Auditory Processing Assessment: Factor analysis and comparisons with SCAN. *American Journal of Audiology, 9,* 101–111.

Eysenck, H. J. (1952). Uses and abuses of factor analysis. *Journal of the Royal Statistical Society, Series C, 1,* 45–49.

Ferguson, M. A., Hall, R. L., Riley, A., & Moore, D. R. (2011). Communication, listening, cognitive and speech perception skills in children with auditory processing disorder (APD) or specific language impairment (SLI). *Journal of Speech, Language, and Hearing Research, 54,* 211–227.

Foster, M. R. (2000). Key concepts in model selection: Performance and generalizability. *Journal of Mathematical Psychology, 44,* 205–231.

Harris, R. J. (1985). *A primer of multivariate statistics* (2nd ed.). New York, NY: Academic Press.

Hart, B., & Spearman, C. (1912). General ability, its existence and nature. *British Journal of Psychology, 5,* 51–84.

Kamhi, A. G. (2011). What speech-language pathologists need to know about auditory processing disorder. *Language, Speech, and Hearing Services in Schools, 42,* 265–272.

Karasinski, C., & Weismer, S. E. (2010). Comprehension of inferences in discourse processing by adolescents with and without language impairment. *Journal of Speech, Language, and Hearing Research, 53,* 1268–1279.

Keith, R. W. (2009). *SCAN–3 for Children: Tests for Auditory Processing Disorders (SCAN–3:C)*. San Antonio, TX: Pearson.

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.

Mainela-Arnold, E., Evans, J. L., & Coady, J. A. (2010). Explaining lexical semantic deficits in specific language impairment: The role of phonological similarity, phonological working memory, and lexical competition. *Journal of Speech, Language, and Hearing Research, 53,* 1746–1756.

McFarland, D. J. (2012). A single g factor is not necessary to simulate positive correlations between cognitive tests. *Journal of Clinical and Experimental Neuropsychology, 34,* 378–384.

McFarland, D. J., & Cacace, A. T. (1995). Modality specificity as a criterion for diagnosing central auditory processing disorders. *American Journal of Audiology, 4,* 36–48.

McFarland, D. J., & Cacace, A. T. (1997). Modality specificity of auditory and visual pattern recognition: Implications for the assessment of central auditory processing disorders. *Audiology, 36,* 249–260.

McFarland, D. J., & Cacace, A. T. (2002). Factor analysis in CAPD and the "unimodal" test battery: Do we have a model that will satisfy? *American Journal of Audiology, 11,* 9–12.

McFarland, D. J., & Cacace, A. T. (2012). Establishing the construct validity of auditory processing disorder (APD). In R. Goldfarb (Ed.), *Translational speech-language pathology and audiology* (pp. 185–191). San Diego, CA: Plural.

Milberg, W. P., Hebben, N. A. Q., & Kaplan, E. (2009). The Boston process approach to neuropsychological assessment. In I. Grant & K. M. Adams (Eds.), *Neuropsychological assessment and neuropsychiatric disorders* (3rd ed., pp. 42–65). New York, NY: Oxford University Press.

Overall, J. E. (1964). Note on the scientific status of factors. *Psychological Bulletin, 61,* 270–276.

Ruscio, J., & Roche, B. (2012). Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factor structure. *Psychological Assessment, 24,* 282–292.

SAS. (2010). *SAS/STAT 9.22 user's guide*. Cary, NC: SAS Institute.

Schow, R. L., Seikel, J. A., Chermak, G. D., & Berent, M. (2000). Central auditory processes and test measures: ASHA 1996 revisited. *American Journal of Audiology, 9,* 63–68.

Semel, E., Wiig, E. H., & Secord, W. A. (2003). *Clinical Evaluation of Language Fundamentals, Fourth Edition (CELF–4)*. San Antonio, TX: The Psychological Corporation.

Sterba, S. K., & Pek, J. (2012). Individual influence on model selection. *Psychological Methods, 17,* 582–599.

Thurstone, L. L. (1940). Current issues in factor analysis. *Psychological Bulletin, 37,* 189–236.

Treutwein, B., & Strasburger, H. (1999). Fitting the psychometric function. *Perception & Psychophysics, 61,* 87–106.

Velicer, W. F., & Fava, J. L. (1998). Affects of variable and subject sampling on factor pattern recovery. *Psychological Methods, 3,* 231–251.

Vugs, B., Cuperus, J., Hendriks, M., & Verhoeven, L. (2013). Visuospatial working memory in specific language impairment: A meta-analysis. *Research in Developmental Disabilities, 34,* 2586–2597.

Watson, C. S., Kidd, G. R., Horner, D. G., Connell, P. J., Lowther, A., Eddins, D. A., . . . Watson, B. U. (2003). Sensory, cognitive, and linguistic factors in the early academic performance of elementary school children: The Benton-IU project. *Journal of Learning Disabilities, 36,* 165–197.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale—Revised* (WAIS–R). San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2008). *Wechsler Adult Intelligence Scale—Fourth Edition* (WAIS–IV). San Antonio, TX: Pearson.

Ximénez, C. (2007). Effect of variable and subject sampling on recovery of weak factors in CFA. *Methodology, 3,* 67–80.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432–442.