Multivariate Behavioral Research, 27 (2), 239-247

The Relevance of Factor Analysis for the Study of Group Differences

Jan-Eric Gustafsson
University of Göteborg

The Guttman article should primarily be seen as a critique of factor analysis as a tool for research on individual and group differences. I agree with most of this criticism, but I disagree with Guttman's negative conclusions about the usefulness of factor analysis. Because it is quite impossible to argue this position in abstract and general terms in the present context, the major share of the space allocated to this commentary is used to present a concrete example intended to illustrate the power of factor analysis for the study of group differences. Let me first, however, briefly emphasize some of the points of agreement.

In Defense of Spearman

One of the most valuable contributions of the Guttman critique is that it rectifies several widespread misunderstandings about the theoretical positions taken by Spearman (1923, 1927). Spearman's explicit rejection of reaction time as a physiological basis of *g* (1927) is an example of this, as is Guttman's observation that what Jensen labeled the "Spearman hypothesis" (1985) in Spearman's own writing was nothing but minor comments.

The most important point, however, is that Spearman (1927) did not view *g* as the first principal factor or principal component of the intercorrelations among a set of intellectual performances. As Guttman pointed out, Spearman's *g* is the common factor in a one-factor model. This definition of *g* demands that the one-factor model fits the data, but the hypothesis that a one-factor model is sufficient to reproduce the intercorrelations among intellectual performances is easily rejected even without the formal hypothesis testing capabilities of modern confirmatory factor analysis. However, in my opinion, it is not necessary to reject the idea of a *g*-factor even if the simplest form of the hypothesis must be rejected. The empirical example to be presented provides an illustration how the general factor may be preserved within the context of a multidimensional model.

Guttman's defense of Spearman (1927) thus is valuable, but I would like to add a few more words in his defense. Guttman's statement that Spearman's "... emphasis was more on algebra than on content" (p. 180) is not quite correct. Spearman (1923) wrote a 350-page book, which treatise should be recognized as the first major contribution to cognitive psychology. In this work, which certainly is not only of historical interest, he proposes "principles of cognition" which seem to capture much of Guttman's own distinctions between "rule inference" and "rule application", but which have not been discussed by Guttman.

An Alternative Factor Analytic Model

Guttman thus rejected the *g*-factor and he rejected multiple factor analysis as a useful technique. However, even though much of his criticism of present factor analytic practices and assumptions is correct, factor analysis is, in my opinion, much too useful a tool to be thrown away. In future research it is necessary, however, to rely upon other types of factor analytic models than oblique multiple factor models. It does seem necessary to liberate factor analysis from the "simple structure" principle which is so correctly criticized by Guttman. It also seems necessary to adopt a hierarchical approach, which allows simultaneous identification of general and specific dimensions of ability. Such models are currently being investigated, so there is reason briefly to review some of the recent developments in the factor analytic research on abilities.

Since about the mid 70's there has been a resurgence of interest in broad factors of ability, primarily because of disenchantment with the multitude of narrow ability factors produced by multiple factor analysis (see Lohman, 1989). The currently most popular hierarchical model is the *Gf-Gc* (Fluid Intelligence-Crystallized Intelligence) model developed by Cattell (e.g., 1963, 1987) and Horn (e.g., 1986, 1989) (Cattell-Horn model), but there also are alternative models (e.g., Vernon, 1950). Recently confirmatory higher-order factor analytic techniques have been used to compare different hierarchical models (Gustafsson, 1984, 1988; Undheim & Gustafsson, 1987). This research has resulted in a model similar to the hierarchical model proposed by Vernon, but it also shares many characteristics with the Cattell-Horn model (see also Undheim, 1981).

The model distinguishes factors at three levels. The lowest level recognizes the narrow abilities identified by Thurstone (1938), Guilford (1967), and other researchers working within the multiple factor tradition. The intermediate level includes factors which closely correspond to broad abilities within the Cattell-Horn model. Among the handful of factors identified at this level, three

seem to be of particular importance: Fluid intelligence (*Gf*), which subsumes abilities such as Induction (*I*) and General Reasoning (*R*); Crystallized intelligence (*Gc*), which is most strongly shown in the factor Verbal Comprehension (*V*); and General Visualization (*Gv*), which is involved in abilities such as Visualization (*Vz*), Spatial Orientation (*SR*) and Flexibility of Closure (*Cf*). At the highest level the model includes a factor of general intelligence (*g*), on which all the broad abilities have loadings. Interestingly enough, however, the loading of *Gf* on *g* consistently has been found to be unity, which implies that the *g*-factor is equivalent with *Gf*. This result thus should make it possible to define the *g*-factor in the same invariant manner as *Gf* is identified.

Gustafsson (1985) argued that the factor identified as *g* by Jensen (1985) is biased in the *Gc* direction because the complex nonverbal reasoning tests which best measure *Gf* are less frequently represented in the studies than those measuring *Gc*. It was also suggested that a more appropriate technique for investigating the nature of black-white difference in performance would be afforded by Sörbom's (1974) technique for analyzing differences in factor means. An attempt will therefore be made to turn these suggestions into practical application.

Kaplan (1989) has suggested a multistage method for studying mean structures in multiple group higher-order confirmatory factor analysis. In this procedure the means on the lower-order factors are first studied, and if differences are found a second step tests whether these differences can be explained by higher-order factor mean differences. However, this procedure confounds the higher-order factors and residuals in lower-order factors when the means on the lower-order factors are studied. Thus, the influence from the higher-order factors should be partialled out from the lower-order factors in an orthogonalized model (see Schmid & Leiman, 1957), leaving residual factors at lower levels in the hierarchy.

This can rather easily be done within the framework of the type of hierarchical model labeled nested-factor (*NF*) model by Gustafsson and Balke (in press). In such models a set of latent variables, which usually are orthogonal, are specified to have direct relations with the observed variables. Typically one latent variable is hypothesized to influence all observed variables, others are hypothesized to influence large sets of observed variables, and others still are hypothesized to influence only a few observed variables. Within such an *NF*-model, mean differences on the latent variables between groups of persons may be investigated with standard techniques. If the "Spearman hypothesis" (Jensen, 1985) is correct, we would expect a large difference in favor of whites on the general factor, and no differences on other factors.

An Empirical Illustration

One of the data sets included in the Jensen (1985) analysis has been selected for reanalysis with such techniques. The study selected is the Jensen and Reynolds (1982) investigation of the U.S. standardization sample (age groups 6.5 to 16.5) of whites (N = 1868) and blacks (N = 305) on the WISC-R (Wechsler, 1974). This study was selected because the descriptive data needed for the reanalysis is included in the article.

A model for the WISC-R (Wechsler, 1974) was hypothesized to include a general factor (G), a broad verbal factor (Gc'; the prime is added to indicate that the factor is a residual factor), a broad spatial-figural factor (Gv'), and a narrow memory span (Ms') factor. With some modifications such a model could be fitted, using LISREL VII (Jöreskog & Sörbom, 1988), to the covariance matrices for whites and blacks with all parameters constrained to be equal for the groups.

The goodness-of-fit test for the model is statistically significant (χ^2 = 384.16, df = 150, $p < .00$), but the fit of the model may nevertheless be regarded as acceptable. Thus, the chi-square value is about 2.5 times as large as the degrees of freedom, which indicates a relatively good fit with samples as large as those analyzed here (Loehlin, 1987). The GFI-values computed by LISREL VII (Jöreskog & Sörbom, 1988) also indicate that the fit is quite good. (whites: .93; blacks: .926). This model may, therefore, be accepted as a first approximation. The overall test of invariance of the measurement model over the two ethnic groups is weakly significant (χ^2 = 67.56, df = 41, $p < .006$), but the differences seem small enough to be disregarded in the present study. Thus, in the analyses of mean differences all parameters in the measurement models have been constrained to be equal over the two groups of persons. The factor loadings are presented in Table 1.

The first factor is a general factor (G) in the sense that every subtest loads on it. To obtain a G-factor which may be interpreted as an invariant Gf-factor we would ideally need three or four Gf-tests, and we would require the G-factor to account for all the variance common to these tests (i.e., there should not be any Gf-factor). Unfortunately, the WISC-R (Wechsler, 1974) battery does not include subtests which clearly belong to the inductive category. The Arithmetic test is not influenced by any other factor than G, and at least for the higher age-groups the items do involve a fair amount of problem solving. The Digit Span test is also known to be a rather clean, but weak, indicator of Gf (Gustafsson, 1984). The basis for interpreting the general factor as Gf thus is weak but there is nothing in the pattern of loadings obtained on the G-factor which contradicts the hypothesis that this factor is close to Gf.

The second factor in the model (Gc') is most strongly related to the verbal tests. It may be noted, however, that three performance tests (Picture

Table 1
Standardized Factor Loadings in the *NF*-model for the WISC-R Estimated for Whites and Blacks in U.S. Standardization Sample

	Factor Loadings			
	G	*Gc'*	*Gv'*	*Ms'*
Information	.69	.35		
Similarities	.62	.42		
Arithmetic	.73			
Vocabulary	.68	.50		
Comprehension	.55	.48		
Digit Span	.55			.39
Tapping Span	.40		.09	.41
Picture Comprehension	.40	.24	.39	
Picture Arrangement	.39	.25	.32	
Block Design	.60		.55	
Object Assembly	.36	.18	.59	
Coding	.42		.11	
Mazes	.35		.34	

Completion, Picture Arrangement, and Object Assembly) load on the *Gc'*-factor as well. These relations were not included in the originally hypothesized model, but were allowed to improve the fit of the model to data. The relations do seem quite reasonable, however, The Picture Completion subtest asks for identification of missing parts of familiar objects and a good vocabulary should make it easier to produce a verbal response. In the items of the Picture Arrangement subtest, the task is to arrange a set of pictures in the right order so they tell a story that makes sense. Examinees with experiences of stories from books and other media are likely to perform better than examinees without such experiences. The Object Assembly test, which has the weakest *Gc'* relation, may possibly have this relation because the objects to be assembled are real objects present in the culture (e.g., apple, horse, and car), and not abstract patterns as in the Block Design subtest.

The performance tests have loadings on the *Gv'*-factor. Most loadings are comparatively small, however, and only in the Block Design and Object Assembly subtests a more substantial amount of variance is due to this factor. The fourth factor (*Ms'*) is a very narrow factor which only accounts for variance in the two memory span tests.

In the next step, differences in factor means for the two groups of subjects have been analyzed. The overall test of the vector of four differences in factor means is very highly statistically significant ($\chi^2 = 423.49$, $df = 4$, $p < .00$). Table 2 presents the estimates along with t-tests of individual parameters.

For three of the factors (*G*, *Gc'*, and *Gv'*) there are very highly significant differences in favor of whites, and for the fourth factor (*Ms'*) there is a borderline significance in favor of blacks. Thus, the pattern of results obtained here does not support the "Spearman hypothesis" (Jensen, 1985).

Discussion and Conclusions

The results presented here are somewhat preliminary and before any major substantive conclusions about differences in performance on WISC-R between whites and blacks are drawn, some further analyses should be conducted. In order to investigate better the nature of the *G*-factor it is necessary to include at least two or three inductive reasoning tests in a model along with the WISC-R subtests. To get further information on the hypothesized *Gc'*- and *Gv'*-factors, it also seems a good strategy to bring in further variables with known properties into the model. The standardization sample also should be broken down into sub-groups more homogenous with respect to age, in order to investigate possible changes in the structure of the model as a function of age. Possible interactions between age and ethnicity with respect to the pattern of mean differences also need to be investigated. These analyses require access to the raw data, however.

Even before these further analyses have been conducted, it may be interesting to speculate about possible interpretations and implications of the present findings. As has already been pointed out they do not support the "Spearman hypothesis" in the sense that the differences between whites and

Table 2
Estimates of Means on Latent Variables for Whites and Blacks

	G	*Gc'*	*Gv'*	*Ms'*
Estimates:				
Whites	0	0	0	0
Blacks	-0.80	-0.77	-0.86	0.26
t-value	-9.56	-7.08	-9.21	2.02

Note. The estimates shown have been divided by the standard deviations of the latent variables to allow comparisons between variables.

blacks are at least as large with respect to *Gc'* and *Gv'* as they are with respect to *G*.

It is true that the present analysis does show a large and significant difference with respect to *G*. It must be emphasized, however, that even though the *G*-factor may be interpreted in terms of *Gf*, this does not imply that differences between individuals or groups are due to genetic factors (see, e.g., Horn, 1986). The *G*-factor does reflect variance from factors such as test-taking skills, persistence, attitude and familiarity with the testing situation, and such factors may account for a substantial amount of the difference in level of performance between whites and blacks. Just to take one example, Fuchs and Fuchs (1986) showed in a meta-analysis that low SES examinees performed worse when tested by unfamiliar rather than familiar examiners, while for high SES subjects no effect was found. Examinee ethnicity was not included as a variable in the study, but the results are at least of indirect relevance for the present discussion because of the observed correlation between SES and ethnicity (Wechsler, 1974). It is, of course, impossible to tell how large a portion of the *G*-factor differences may be accounted for by this and other contextual variables of the testing situation, but the Fuchs and Fuchs results indicate a considerable power of such factors as determiners of group differences in test performance.

The present study does not provide any information about the causes of the differences with respect to *Gc'* and *Gv'*. It is interesting to note, however, that the technique employed here does show that there are large differences in means on these factors. Jensen and Reynolds (1982) failed to find large differences for any factor but the *G*-factor when they computed factor scores from a four-factor exploratory solution. However, even though the factor scores were computed from an orthogonal model there were substantial correlations among the estimated factor scores, which may be one reason for the different patterns of results.

Let me, finally, make a few comments about differences and similarities between the *dimensional* factor analytic methodology applied here, and the *regional* analysis advocated by Guttman. There is in fact a rather close correspondence between the radex model and the hierarchical factor model. As has been demonstrated by Snow, Kyllonen and Marshalek (1984; see also Marshalek, Lohman & Snow, 1983) the *G*-factor loading of a test is closely related to the *centrality* of the test in the two-dimensional radex plot. The *G*-factor thus is closely related to the *operations* facet (facet C) with the highest *G* loadings for *rule-inference* tests, intermediate *G* loadings for *rule-application* tests, and lowest *G* loadings for *learning* tests. The *content* facet is also easily recognized in the hierarchical factor model. The *verbal* element corresponds to the *Gv*-factor. The *numerical* element is not so easily mapped onto a broad

quantitative factor (*Gq*) because the *G*-factor tends to predominate in complex numerical tests, but in less complex numerical tasks a numerical factor is easily identified.

Thus, despite the seemingly very important differences between a regional analysis as advocated by Guttman, and the hierarchical factor model, a deeper analysis reveals striking similarities. Given these similarities, it does not seem wise to reject one of these approaches as flawed and only pursue the other, since they are likely to have both advantages and disadvantages.

The technique of multidimensional scaling seems particularly useful when the aim is to uncover the major structural aspects of the data. This approach also avoids concepts such as *latent variables* or *traits*, which concepts involve risks of reification and hypostazation of human characteristics (see Snow & Lohman, 1989, for an extended discussion).

But the regional approach also has disadvantages when compared with the dimensional one. Thus, in spite of the strong claims made by Guttman about the ease with which group differences may be studied over different parts of the region, the technique does not seem to provide any support for actually doing such an analysis. Any point in the radex or cylindrex is a combination of elements from two or more facets and a simple ocular inspection does not suffice to separate the different sources of influence. As has been shown above this separation can, however, easily be done with the dimensional approach.

It is also my impression that the dimensional approach allows a more fine-grained and powerful analysis than does the regional one. Not only are the modern factor analytic techniques based on a solid mathematical and statistical foundation, with powerful and robust methods for estimation and testing, but they are very versatile as well. Thus, in the model presented in Table 1, some tests are influenced both by *Gc'* and *Gv'* in addition to *G*. A regional analysis could not easily represent such a radical deviation from the notion of simple structure, because it would require simultaneous classification of a task into two elements of the same facet. In conclusion, I cannot think of any other technique more relevant and useful for the study of group differences than factor analysis.

References

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

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action*. Amsterdam: North Holland.

Fuchs, D., & Fuchs, L. S. (1986). Test procedure bias: A meta-analysis of examiner familiarity effects. *American Educational Research Journal, 56*, 243-262.

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

Gustafsson, J.- E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence, 8*, 179-203.

Gustafsson, J.- E. (1985). Measuring and interpreting *g*. *The Behavioral and Brain Sciences, 8*, 231-232.

Gustafsson, J.- E. (1988). Hierarchical models of individual differences in cognitive abilities. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence Vol. 4* (pp. 35-71). Hillsdale, NJ: Erlbaum.

Gustafsson, J.- E., & Balke, G. (in press). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research.*

Horn, J. L. (1986). Intellectual ability concepts. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence Vol. 3* . Hillsdale, NJ: Erlbaum.

Horn, J. L. (1989). Models of intelligence. In R. L. Linn (Ed.), *Intelligence. Measurement, theory, and public policy* (pp. 29-23). Urbana: University of Illinois Press.

Jensen, A. R. (1985). The nature of the black-white difference on various psychometric tests: Spearman's hypothesis. *The Behavioral and Brain Sciences, 8*, 193-263.

Jensen, A. R., & Reynolds, C. R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences, 3*, 423-438.

Jöreskog, K. G., & Sörbom, D. (1988). *LISREL 7. A guide to the program and applications.* Chicago: SPSS.

Kaplan, D. (1989). *A multistage method for studying mean structures in multiple group higher order confirmatory factor analysis.* Unpublished manuscript.

Loehlin, J. C. (1987). Latent variable models. *An introduction to factor, path, and structural analysis.* Hillsdale, NJ: Erlbaum.

Lohman, D. F. (1989). Human intelligence: An introduction to advances in theory and research. *Review of Educational Research, 59*, 333-373.

Marshalek, B., Lohman, D., & Snow, R. (1983). The complexity continuum in the radex and hierarchical models of intelligence. *Intelligence, 7*, 107-127.

Schmid, J., & Leiman, J. M. (1957). The development of hierarchical factor solutions. *Psychometrika, 22*, 53-61.

Snow, R. E., & Lohman, D. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement.* New York: Macmillan.

Snow, R. E., Kyllonen, P. C., & Marshalek, B. (1984). The topography of ability and learning correlations. In R. J. Sternberg (Ed.), *Advances in the psychology of human intelligence Vol. 2.* Hilsdale, NJ: Erlbaum.

Spearman, C. (1923). *The nature of 'intelligence' and the principles of cognition.* London: Macmillan.

Spearman, C. (1927). *The abilities of man.* London: Macmillan.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structures between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229-239.

Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs, 7.*

Undheim, J. O. (1981). On intelligence IV: Toward a restoration of general intelligence. *Scandinavian Journal of Psychology, 22*, 251-265.

Undheim, J. O., & Gustafsson, J.- E. (1987). The hierarchical organization of cognitive abilities: Restoring general intelligence through the use of linear structural relations (LISREL). *Multivariate Behavioral Research, 22*, 149-171.

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

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