INTELLIGENCE 18, 165–187 (1994)

Hierarchical Models of Ability at Individual and Class Levels

KJELL HÄRNQVIST
JAN-ERIC GUSTAFSSON
University of Göteborg, Sweden

BENGT O. MUTHÉN
GINGER NELSON
University of California, Los Angeles

Scores in ability tests administered to students in Grades 4–6 (n = 1,274) and Grades 7–9 (n = 1,310) were simultaneously factor analyzed at class and individual levels. Muthén's (1990) multilevel factor analysis was used to test hierarchical models of intelligence. At the individual level the general factor was most highly loaded on "fluid" abilities. Also, there were residual factors for speed, verbal comprehension, spatial visualization, and numerical facility as well as test-specific factors derived from parallel versions of the tests. At the class level three common factors were established: one general factor more highly loaded on "crystallized" than on fluid abilities, and residual common factors for speed and, in Grades 7–9, verbal comprehension. The between factors accounted for relatively more variance in Grades 7–9 than in Grades 4–6 due to an intervening reorganization of classes based on choice of more or less academic courses. Demographic differences between neighborhoods, self-selection between academic and general programs, and variations in instruction and test administration were used to explain the between-class factors.

During the last couple of decades there has been a resurgence of interest in the concept of intelligence and other broad dimensions of individual differences in cognitive abilities (Lohman, 1989). Possible reasons for this are that narrowly specialized abilities of the kind identified by Thurstone (1938) and Guilford (1967) only marginally improve prediction of educational outcomes (cf. Gustafsson & Balke, 1993), and they have been found to be unproductive in research on interactions between aptitudes and treatments (Cronbach & Snow, 1977). The growing popularity of hierarchical models of ability (see, e.g., Gustafsson, 1988) may be another reason for the current interest in broad concepts of ability. Such models allow for both broad and narrow abilities, but the broader abilities have been emphasized.

Constructive criticism by the reviewers has greatly helped to improve the article.

Correspondence and requests for reprints should be sent to Kjell Härnqvist, Department of Education and Educational Research, University of Göteborg, P.O. Box 1010, S-431 26 Mölndal, Sweden.

#53

This is true for the currently most popular hierarchical model, the model of fluid and crystallized abilities developed by Cattell (1943, 1963, 1987) and Horn (1976, 1986, 1989). This model emphasizes a set of broad dimensions of ability, the two most prominent of which are fluid ability (Gf) and crystallized ability (Gc). Both these abilities involve abstraction, concept formation and perception, and eduction of relations. Gf, however, is involved in tasks that are new to the examinee, whereas Gc is shown in familiar tasks, which typically have verbal-conceptual content. Gf is thought to represent influences of biological factors and incidental learning on intellectual development, whereas Gc is interpreted as reflecting education and experience.

The concept of general ability has not been given any place within the Gc–Gf model, and Horn (1989) in particular has argued against the notion of general intelligence. However, Undheim (1981) argued on both theoretical and empirical grounds that Gf is equivalent to the Spearman *g* factor. Gustafsson (1984, 1988; see also Undheim & Gustafsson, 1987) employed techniques of higher order factor analysis to formulate and test alternative hierarchical models of ability and found a third-order *g* factor to be perfectly related to the second-order Gf factor. One interesting theoretical implication of this result is that it allows a unification of the hierarchical models of the British tradition (Vernon, 1961) with the Gc–Gf model.

However, there are also studies that indicate a closer affinity of the general factor to Gc than to Gf. Humphreys, Parsons, and Park (1979) studied variances and intercorrelations of school means to investigate the nature of the general factor. They argued that processes of unplanned social selection, such as choice of neighborhood and high school, might be informative about the relative importance of the general factor versus group factors on the one hand, and of the relative importance of socioeconomic factors versus cognitive factors on the other hand. Humphreys et al. used the Project TALENT Data Bank with data from 1960 for approximately 400,000 students to compute school means separately for boys and girls for a large set of measures of cognitive performances, socioeconomic status (SES), and school characteristics. The unweighted raw school means were used to compute variances and intercorrelations of the measures, which were then analyzed with a variety of techniques for exploratory factor analysis.

The preferred solution included four factors: one labeled *g* and three group factors. The *g* factor accounted for the major share of the variance. The highest loadings on this factor were obtained for Vocabulary (.96 for boys, .95 for girls), Reading Comprehension (.95, .96), and Social Studies (.92, .90). Tests that typically load highly on Gf had comparatively lower loadings (e.g., Abstract Reasoning: .85, .88; Visualization in Three Dimensions: .77, .78). The Gf tests, in fact, had lower loadings than variables such as Outdoor Activities (.89, .86) and Art (.87, .86). This pattern of results thus strongly suggests that the factor labeled *g* by Humphreys et al. (1979) is closer to Gc than to Gf. The authors did observe a possible verbal bias in the general factor and ascribed it to the relatively large number of verbal information tests in the battery.

There is, however, another possible explanation for the result that *g* appears to coincide with Gc at the school level, namely, that the mechanisms causing between-school differences primarily involve Gc rather than Gf. If it is indeed the case that *g* comes close to Gf at the individual level but comes close to Gc at higher levels of aggregation, this would be a result of theoretical significance, and of practical importance for the design and interpretation of large-scale studies.

Humphreys et al. (1979) only analyzed data at a high level of aggregation and used rather crude exploratory factor-analytic techniques. Recently, however, Muthén (1989, 1990, 1991) has developed a technique for modeling at two levels simultaneously. This technique of multilevel factor analysis (MFA) is applied here in a reanalysis of a set of ability test scores in order to study hierarchical models of ability at class and individual levels.

SUBJECTS, VARIABLES, AND SINGLE-LEVEL FACTORS

The analyses to be reported here are based on a battery of tests administered to intact classes in Grades 4 through 9 in Swedish comprehensive schools. The schools were chosen so that they covered practically the entire school-age population in their geographical districts. On the average, there were 478 students in each grade, about equally divided between boys and girls and distributed among 20 to 29 classrooms. The class sizes varied from 5 to 34 pupils. About 90% of the students were of normal age for their grade, that is, 11 years old in Grade 4 up to 16 years old in Grade 9.

Within school districts the students had been assigned to classes in different ways for Grades 4–6 and 7–9. In the elementary grades the assignment was normally done before Grade 1 according to neighborhood principles, and demographic differences between, for instance, rural and more densely populated areas of the districts may have had an influence. In the upper grades the students could choose between more or less academic programs and were assigned to classes according to their choice. Classes in Grades 4–6 were taught by classroom teachers, and by subject-matter teachers in Grades 7–9.

The original data collection (Härnqvist, 1960) was done at a time when Thurstone's (1938) Primary Mental Abilities (PMA) still was the dominating model of the factorial structure of intelligence. The tests were chosen or constructed in accordance with this PMA model—two well known tasks for each hypothesized factor:

Verbal Comprehension (V)	Synonyms and Opposites
Inductive Reasoning (I)	Letter Grouping and Figure Series
Spatial Visualization (Vz)	Metal Folding and Block Counting
Number Facility (N)	Additions and Multiplications
Perceptual Speed (P)	Identical Numbers and Highest Number

Test scores were recorded for two parallel forms, that is, odd and even halves of the V, I, and Vz tests and for separately administered forms of the speeded N and P tests.

In various early exploratory factor analyses the PMA model was, in general, found to be valid, but with higher order factors bringing V, I, and Vz together in one second-order factor and N and P together in another. In a manual from the early 1960s (Härnqvist, 1962) for the practical use of the tests in schools, it was recommended to base a measure of general intelligence on V, I, and Vz tests, a measure of speed on N and P tests, and, in addition, to use the contrast between verbal and spatial tests for differential prediction.

A reanalysis (Härnqvist, 1978) was inspired by Cronbach's (1976) report on research on classrooms and schools. There, Cronbach recommended a decomposition of the individual scores into components for different levels of aggregation resulting in separate estimates for different levels, for instance, between and within classrooms. The reanalysis was done on decomposed scores at class and individual levels. The class-level variation includes a minor component of variation between school districts, but these were too few to be kept at a separate level in the analysis. Exploratory factor analyses were performed for each grade separately. The results differed somewhat between grade groups but, tentatively, the empirical findings could be generalized in a model (Härnqvist, 1978, p. 713) where the test scores at the individual level form a PMA structure, and above that are found second-order factors for Power and Speed and a third-order *g* factor. At the class level, the basic PMA structure is not found, but Power versus Speed and *g*.

Gustafsson (1989) reanalyzed part of the original sample in order to demonstrate a new factor-analytic approach to hierarchical models. Instead of second-order factors accounting for correlations among PMA factors, he developed a so-called nested factor model with one general cognitive factor (G), and residual verbal (V′), spatial (Vz′), and number (N′) factors, orthogonal to the general factor, as well as test-specific factors based on the parallel halves or forms of the tests. The term "residual" means that the influence of higher order factors has been partialed out, in this case G from verbal, spatial, and numerical factors, and G, V′, Vz′, and N′ from the test-specific factors. This model was tested by means of LISREL (Jöreskog & Sörbom, 1986) on the total covariance matrix for Grade 6, $\chi^2(92, N = 207) = 94$, and a goodness-of-fit index (GFI) of .95.

Gustafsson's (1989) hierarchical modeling has also been applied in this reanalysis but with some modifications. The most important change was guided by the results of the 1978 analysis. The N and P tests were used as a basis for a general speed factor (Gs′) orthogonal to the G factor, and the numerical factor N′ was made orthogonal to both G and Gs′.

The analysis was done on covariance matrices including 20 variables, that is, parallel halves or forms of each of the 10 tests. As a rule, the resulting factor estimates differed only marginally between the parallel subtests. In order to save space, these estimates have been averaged in the following tables.

The factorial model resulting from the analysis of individual data is presented in Figure 1. Both parallel versions of each test are illustrated by one single box. In the starting model, the factor loadings were fixed according to the modified PMA structure as described previously in relation to Gustafsson's (1989) analysis, but with the following exceptions. In the first and general factor, all tests except Synonyms and Opposites were set free. In the residual general speed factor, Letter Grouping and Metal Folding, as well as the even halves of the numerical and perceptual speed tests, were set free as a result of the modification procedure. Thus, the factor analysis was of a confirmatory kind, but through the modifications, an element of exploration was, as in most cases, introduced: Grades 4–6, $\chi^2(152, N = 1{,}274) = 362$, Grades 7–9, $\chi^2(152, N = 1{,}310) = 444$; the GFI was .97 for both groups. Moderate ceiling effects were observed for some tests in Grades 7–9. These are not likely to have influenced chi-square analysis very much (cf. Muthén & Kaplan, 1985). The measures of fit must be regarded as satisfactory considering that the sample sizes with complete information were as large as 1,274 and 1,310.

Table 1 presents estimates of the relative contributions of different factors. Here, as well as later in the article, the contributions have been estimated from the unstandardized factor loadings and the factor variances. More specifically,

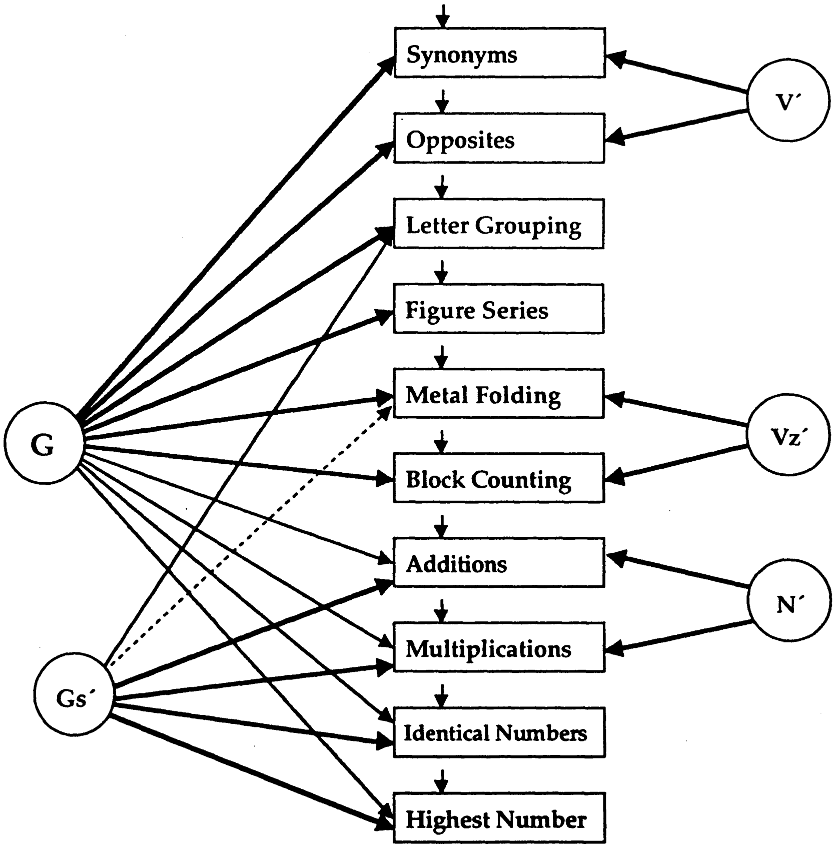

Figure 1. Latent variable model for individual test scores.

TABLE 1

Contributions of Different Factors to Estimated Test Variance (Average for Parallel Forms)

	G	Gs′	V′	Vz′	N′	Spec′
Grades 4–6						
Synonyms	.47		.37			.16
Opposites	.56		.44			
Letter Grouping	.51	.04				.45
Figure Series	.61					.39
Metal Folding	.59	.02*		.14		.25
Block Counting	.42			.15		.44
Additions	.23	.47			.20	.11
Multiplications	.23	.43			.19	.15
Identical Numbers	.11	.64				.24
Highest Number	.21	.51				.28
Grades 7–9						
Synonyms	.47		.40			.12
Opposites	.54		.46			
Letter Grouping	.58	.03				.39
Figure Series	.66					.34
Metal Folding	.45	.01*		.25		.29
Block Counting	.40			.22		.38
Additions	.11	.59			.26	.05
Multiplications	.10	.52			.22	.16
Identical Numbers	.04	.64				.31
Highest Number	.06	.56				.38

*Negative factor loading.

the loadings have been squared and multiplied by the factor variance, and these products were then divided by the sum of estimated (error-free) variances in the tests. The general cognitive factor (G) is highly loaded on the tests coming from the PMA factors I, V, and Vz. The numerical and perceptual speed tests have small but significant loadings.

The residual general speed factor (Gs′) has high saturations in the perceptual speed tests and also in the numerical tests as expected. The modifications called for in the analysis, however, consistently resulted in two additional small loadings, which may have to do with response patterns rather than cognitive content. In Letter Grouping (with positive loading) an alphabet was supplied as an aid. Quickly referring to that may have helped in the problem solving. In Metal Folding (with negative loading), on the other hand, too rapid responses may have been detrimental. However, their contributions to the variance can almost be ignored.

The V′ factor measures the part of verbal comprehension that is not explained by general cognitive ability but by verbal knowledge. This factor is a narrow representative of Gc in the present battery.

The Vz′ factor measures spatial performance to the extent that it is not based on general reasoning but rather on visualization. Among the two Vz′ tests, Metal Folding has more of G.

The numerical tests have their own residual N′, but it is less strong than their loadings on the general speed factor.

Two of the original PMA factors are missing in this set of residual factors. The reasoning factor, I, is fully merged with G, which indicates that G is close to inductive reasoning (i.e., Gf) in this analysis. This is in agreement with consistent findings by Gustafsson (1988, 1989) in his development of the hierarchical model used here. The perceptual speed factor P is merged with Gs′.

All tests except Opposites have significant contributions from test-specific factors (Spec′) based on the parallel halves or forms of the tests. In the reasoning and spatial tests, the test-specific contributions to the true variance are quite strong: from 25% to 45%. In contrast, the verbal and numerical tests have little or even no specificity at all. This difference evidently has to do with the varying degree of similarity of the tasks chosen to represent the same PMA factor.

THE MULTILEVEL FACTOR ANALYSES (MFA)

The single-level factor analysis presented previously serves mainly as a background to the multilevel analysis reported now. MFA, according to Muthén (1990), separates the total covariance of the tests into two parts—one between groups (e.g., classrooms) and one between individuals within groups. The factor analysis of these covariance matrices is done in one step, as in simultaneous analysis in several groups with LISREL or LISCOMP (Muthén, 1988; Nelson & Muthén, 1991).

Because the group means, on which the between covariances are calculated, contain sampling errors inversely related to the subgroup sizes, the observed between covariances have to be reduced for this sampling variation. Muthén (1990) has developed two different methods (MUML and FIML) for this adjustment. The MUML (Muthén's maximum likelihood based estimator) is an approximate method where the influence of the subgroup sizes on the stability of class means is taken care of by means of an ad hoc estimator related to the average class size. FIML uses full information maximum likelihood for each subgroup, which makes the computations much more time consuming. Comparisons have shown that the two methods give very similar MFA results (Härnqvist, Gustafsson, Muthén, & Nelson, 1991). In this article, only MUML is used.

The MFA analysis at class and individual levels is based on two sample covariance matrices: one for the subgroup means weighted by class size ($\mathbf{S_B}$) and one for the variation between individuals within classes pooled for the entire set of classes ($\mathbf{SP_w}$; for statistical documentation, see Muthén, 1990, Section 4).

In order to get stable results, it is necessary to base the between-covariance matrix on many observations, that is, many class means. Muthén (1990) recommended at least 50 to 100 classes. Thus, it is not meaningful to analyze each grade separately as was done in Härnqvist (1978). On the other hand, the number of classes is large enough for two groups of grades. The different principles of assignment make it natural to treat Grades 4–6 with 83 classes and 1,274 students as one group and Grades 7–9 with 68 classes and 1,310 students as another. Most test scores show an average increment from grade to grade. The pooling of grades thus implies some increase in the variation at between level. However, preliminary analyses not reported here have demonstrated that the factor model is valid in all grades.

In the single-level analysis (cf. Table 1) a 14-factor solution was found comprising a general cognitive factor and, orthogonal to that, factors for Speed, Verbal Comprehension, Spatial Visualization, and Number Facility, as well as test-specific factors for all tests except Opposites. In the MFA this model was tried at both class and individual level. The free loadings of Letter Grouping and Metal Folding on Gs′ were kept only at the within level. All other loadings were fixed: for 14 + 14 factor model, Grades 4–6, $\chi^2(316, N = 1{,}274) = 694, p < .01$; Grades 7–9, $\chi^2(316, N = 1{,}310) = 625, p < .01$. Among the factors at the *between* level, only the following were significant.

- Grades 4–6: General, General Speed, and the specifics of Letter Grouping, Figure Series, and Multiplications.
- Grades 7–9: General, General Speed, Verbal, and the specifics of Letter Grouping, Metal Folding, Block Counting, and Highest Number.

When the nonsignificant factors were eliminated, the changes had comparatively small effects on the fit of the model, and the loadings on the remaining factors changed only slightly. Therefore, and in order to facilitate comparisons between grade groups as well as levels, the loadings from the 14 + 14 model (*df* = 316) will be kept.

Table 2 shows the completely standardized loadings at the within-class level. The inductive and spatial visualization tests are most strongly loaded on G_w. The verbal tests also have high loadings on Gw but even more so on V'_w. The numerical tests are mainly found in Gs'_w and N'_w, and the perceptual ones in Gs'_w. The pattern is rather similar to that shown in Table 1, although it must be observed that Table 1 presents variance contributions, that is, squared loadings, and Table 2 presents standardized loadings.

The picture becomes radically different when one turns to the standardized loadings at the between level in Table 3. Loadings on nonsignificant factors are put in parentheses and should not be included if estimating the total test variance at between level, explained by the factors, from the squared loadings. In both grade groups the general factor between (G_b) dominates strongly. With the ex-

TABLE 2
MFA Factor Estimates at Within-Class Level: Standardized Factor Loadings for Grades 4–6 and 7–9 (Average for Parallel Forms)

	G_w	Gs'_w	V'_w	Vz'_w	N'_w	$Spec'_w$
Grades 4–6						
Synonyms	.52		.60			.40
Opposites	.57		.64			
Letter Grouping	.57	.20				.63
Figure Series	.68					.57
Metal Folding	.64	−.15		.36		.47
Block Counting	.52			.34		.60
Additions	.33	.64			.44	.32
Multiplications	.29	.65			.44	.39
Identical Numbers	.14	.72				.49
Highest Number	.23	.64				.54
Grades 7–9						
Synonyms	.48		.63			.38
Opposites	.51		.66			
Letter Grouping	.60	.17				.62
Figure Series	.71					.55
Metal Folding	.54	−.10		.46		.49
Block Counting	.52			.42		.55
Additions	.08	.70			.50	.25
Multiplications	.02	.66			.48	.41
Identical Numbers	.02	.65				.50
Highest Number	.09	.58				.61

ception of Identical Numbers and Highest Number in Grades 7–9, all loadings are .80 or above and nine of them are .90 or above. Moreover, they are also higher in G_b than in Gs'_b for the numerical and perceptual speed tests.

A full picture of the contributions of different factors to estimated ("true") test variance is found in Table 4, where it is easy to compare the within and between parts of variance pairwise. Contributions of nonsignificant factors are put in parentheses. The contributions from G_w to the estimated variance in different tests varies greatly: from .00 in Identical Numbers in the upper grades to .50 in Figure Series in the lower grades. The contributions from G_b vary much less with .10 in Block Counting (Grades 4–6) and Identical Numbers (Grades 7–9) and .34 in Opposites (Grades 7–9) as extreme values.

In the numerical and perceptual speed tests, the between contributions of G exceed the within contributions. On average, the between parts cover 80% (numerical speed) and 90% (perceptual speed) of the total contributions of G in these tests. The reverse relation holds for the inductive and spatial tests where the between parts, on average, cover only 32% (inductive) and 27% (spatial) of the

TABLE 3
MFA Factor Estimates at Between-Class Level: Standardized Factor Loadings for Grades 4–6 and 7–9 (Average for Parallel Forms)

	G_b	Gs'_b	V'_b	Vz'_b	N'_b	$Spec'_b$
Grades 4–6						
Synonyms	.94		(.29)			(.16)
Opposites	.95		(.28)			
Letter Grouping	.92					.36
Figure Series	.89					.42
Metal Folding	.94			(.25)		(.22)
Block Counting	.90			(.29)		(.25)
Additions	.82	.38			(.24)	(.26)
Multiplications	.80	.37			(.23)	.28
Identical Numbers	.81	.42				(.16)
Highest Number	.81	.28				(.28)
Grades 7–9						
Synonyms	.94		.33			(.12)
Opposites	.93		.33			
Letter Grouping	.93					.38
Figure Series	.94			(.30)		(.30)
Metal Folding	.88			(.32)		.38
Block Counting	.85					.41
Additions	.81	.54			(.36)	(.09)
Multiplications	.84	.48			(.32)	(.13)
Identical Numbers	.56	.56				(.18)
Highest Number	.51	.41				.48

Note. Parentheses indicate loading in nonsignificant factor.

total contributions of G. The verbal tests are found near the middle with a mean of 52% at the between level.

Averaging the contributions of G to each pair of tests gives the following results:

Within	
Inductive	.41
Spatial	.34
Verbal	.25
Numerical	.05
Perceptual	.02
Between	
Verbal	.28
Numerical	.19
Inductive	.19
Perceptual	.17
Spatial	.12

TABLE 4
Contributions of Different Factors to Estimated Test Variance (Average for Parallel Forms)

	G	Gs′	V′	Vz′	N′	Spec′
Grades 4–6						
Synonyms						
Within	.27		.34			.15
Between	.22		(.02)			(.01)
Opposites						
Within	.32		.40			
Between	.26		(.02)			
Letter Grouping						
Within	.34	.04				.41
Between	.18					.03
Figure Series						
Within	.50					.35
Between	.12					.03
Metal Folding						
Within	.45	.02*		.14		.24
Between	.14			(.01)		(.01)
Block Counting						
Within	.32			.14		.42
Between	.10			(.01)		(.01)
Additions						
Within	.10	.38			.18	.09
Between	.17	.04			(.02)	(.02)
Multiplications						
Within	.08	.37			.18	.13
Between	.18	.04			(.01)	.02
Identical Numbers						
Within	.02	.51				.23
Between	.19	.05				(.00)
Highest Number						
Within	.05	.35				.25
Between	.29	.03				(.03)
Grades 7–9						
Synonyms						
Within	.20		.34			.12
Between	.30		.04			(.00)
Opposites						
Within	.23		.39			
Between	.34		.04			
Letter Grouping						
Within	.31	.03				.34
Between	.27					.05
Figure Series						
Within	.49					.30
Between	.19					(.02)

(continued)

TABLE 4 (Continued)

	G	Gs′	V′	Vz′	N′	Spec′
Metal Folding						
Within	.30	.01*		.23		.26
Between	.15			(.02)		.03
Block Counting						
Within	.31			.20		.34
Between	.11			(.02)		.02
Additions						
Within	.01	.43			.23	.06
Between	.19	.09			(.00)	(.00)
Multiplications						
Within	.00	.36			.19	.14
Between	.23	.07			(.00)	(.00)
Identical Numbers						
Within	.00	.51				.27
Between	.10	.10				(.01)
Highest Number						
Within	.01	.34				.38
Between	.11	.07				.09

Note. Parentheses indicate loading in nonsignificant factor.
*Negative factor loading.

These comparisons clearly indicate that G at within level comes close to a factor of fluid intelligence (Gf), whereas G at between level is more related to crystallized intelligence (Gc).

Similar comparisons of average contributions of the general speed factor (Gs′) give the following results:

Within	
Perceptual	.43
Numerical	.40
Between	
Perceptual	.06
Numerical	.06

It seems that Gs'_w is a measure of individual differences in perceptual speed, whereas Gs'_b may reflect variations between classes in one or both of two respects: differences in emphasis on speeded performance, that is, a kind of treatment effect, and irregularities in the administration of the speeded tests between different classrooms.

The verbal factor is significantly represented at the between level only in Grades 7–9. The average contributions of this factor, nonsignificant loadings included, amount to the following values:

Within	
Verbal	.37
Between	
Verbal	.03

Evidently, the verbal tests come closer to the crystallized V'_w factor (.37) than to G_w (.25). This strengthens the interpretation of G_w as a factor of fluid intelligence.

ERROR-FREE VARIANCES

The parallel analysis of elementary and upper grades has demonstrated that similar models account for the pattern of relations in both groups. This can be regarded as a cross-validation of the model at large. So far, however, the differences in factor loadings or variance contributions between the two grade groups have not been studied closely enough. Such differences may emerge in the multilevel analysis as an effect of the different principles of student assignment in the elementary and upper grades.

As mentioned before, students were assigned to elementary classes according to neighborhood principles. In Grades 7–9, they were assigned to classes according to their own choice between more or less academic programs. Such programs are likely to attract students with different characteristics. In addition, more competent teaching in the academic programs and interaction between program characteristics and intellectual development may improve not only crystallized but also fluid abilities. Such mechanisms make it reasonable to expect a relative increase of the variation between classes from elementary to upper grades.

In his factor analysis of pre- and posttests of mathematics achievement, Muthén (1991) introduced methods to separate true from error variance in the tests, based on the results of the multilevel factor analysis. From these estimates, several indices can be computed, for example, measures of reliability at between and within class level, true intraclass correlations, and true changes in variances between and within classes from pre- to posttest or between different subgroups.

The variances that are needed for such indices are shown in the Appendix A. The information includes estimated true variances between (BF) and within classes (WF) as well as error variances between (BE) and within (WE), all of them for both grade groups (see also Appendices B–F).

The reliability estimates derived from these components can be summarized as follows. The within-class reliability of half-tests in Grades 4–6 varies from .74 to .86 with an average of .78, and in Grades 7–9 from .70 to .83 (average = .77). The between-class reliability is much higher (average = .97) for all tests except Identical Numbers and Highest Number, which have .82 in Grades 4–6 and as low as .63 and .65 in Grades 7–9.

Of greater interest in this context are the true intraclass correlations in the two grade groups that indicate the proportion of true variance that comes from variations between classes. These intraclass correlations are found in Table 5. The intraclass correlations vary from .11 in Block Counting (Grades 4–6) to .38 in Opposites (Grades 7–9). In all tests except the two perceptual speed tests mentioned before, the correlation is considerably higher in the upper than in the elementary grades. The differences vary between −.09 and .10 (average = .045).

Another piece of information comes from the differences between the two grade groups in true variances between and within classes. In Table 6, these differences are expressed as proportions of the variance in Grades 4–6. At the within level, most variances decrease, which is to be expected after self-selection has taken place to more or less academic programs; around an average decrease of −.14, the proportions vary from −.42 for Opposites to .14 for Block Counting. The extra large decrease for the verbal tests can be interpreted in two different ways. It can reflect an emphasis on verbal abilities in the choice of courses, and, also, the fairly large decrease in numerical tests supports an interpretation that self-selection, to a large extent, depends on crystallized abilities. The decrease, however, can also partly be due to a ceiling effect in the verbal tests, which is reflected in the observed score distributions.

At the between level, on the other hand, half of the variances increase, and some of them to a considerable degree. Around an average increase of .07, the proportions vary from .65 in Letter Grouping to −.31 in Highest Number. The inductive and spatial tests show increases, and the verbal and perceptual tests decrease. This indicates, also, that fluid abilities are strongly related to the assignment to classes—through self-selection or treatment—whereas ceiling effects may hinder the influence of verbal abilities to fully manifest itself.

In several respects, the perceptual speed tests show a different pattern than the rest of the test battery. They have lower between-class reliability than the other

TABLE 5
True Intraclass Correlations in Grades 4–6 and 7–9

	Grades 4–6	Grades 7–9	Differences
Synonyms	.24	.34	.10
Opposites	.28	.38	.10
Letter Grouping	.21	.32	.11
Figure Series	.15	.21	.06
Metal Folding	.16	.20	.04
Block Counting	.11	.15	.04
Additions	.24	.28	.04
Multiplications	.24	.31	.07
Identical Numbers	.24	.22	−.02
Highest Number	.36	.27	−.09

TABLE 6
Error-Free Increase in Between and Within Variance From Grades 4–6 (L) to Grades 7–9 (H)

	$(WF_H - WF_L)/WF_L$	$(BF_H - BF_L)/BF_L$
Synonyms	−.43	−.11
Opposites	−.42	−.08
Letter Grouping	−.08	.65
Figure Series	−.09	.38
Metal Folding	−.04	.29
Block Counting	.14	.52
Additions	−.24	−.08
Multiplications	−.15	.18
Identical Numbers	−.08	−.17
Highest Number	.02	−.31

tests, especially in Grades 7–9. Their intraclass correlations do not increase between grade groups, and they also behave differently from the numerical tests in Table 6. It seems that the larger error terms at the between level are behind all these deviations.

The error terms were based on the variation between two parallel forms with 5 min as the time limit in Identical Numbers and 3 min in Highest Number. The first forms were administered in Lessons 2 and 3 in the morning of the first day, the second forms in Lessons 2 and 3 of the following morning. One possible interpretation of this combination of circumstances may be that a decrease in testing discipline and test-taking motivation for such repeated elementary tasks occurred, and more likely so in the upper than in the elementary grades. Even without a ceiling effect on the observed scores, this might result in a reduced variation in true scores and a relative increase in error terms.

DISCUSSION AND CONCLUSIONS

The results at the between-class level of the multilevel analysis show striking similarities with the results of the school-level analysis reported by Humphreys et al. (1979). Their *g*, with its emphasis on vocabulary and reading comprehension, is very similar to G_B found in the MFA. In their Group Factor II, three highly speeded clerical tests have by far the highest loadings. This factor seems to correspond to our speed factor (Gs'_B).

In describing our findings at the between level, we have hypothesized the following sources of variation in the general factor (G_B).

- Demographic differences, including social and educational characteristics of homes and areas from which the students come, are shown in differences between classes, and for the elementary grades this is likely to be the primary source of variation captured by the general between factor.

- Self-selection to more or less academically oriented programs in the upper grades is an additional source of G_B variation at that stage. It registers in the differences between intraclass correlations for the two stages.
- Treatment effects through more competent teaching and intellectually more stimulating programs may help to improve the development of both crystallized and fluid abilities. These too can contribute to increased variation between classes at the upper stage.

Such sources of variation between schools in the general factor were also recognized by Humphreys et al. (1979). Moreover, a measure of average SES, which loaded .76 and .77 on *g*, gave direct evidence of this. On the other hand, only Rate of College Going, among 19 school characteristics, had moderate loadings on *g*. Instead, many of them were found in Group Factor I, which distinguished rural and small town schools from urban schools.

For the general speed factor between classes (Gs'_B), the following sources of variation have been discussed:

- Treatment effects that may affect the Gs'_B factor through classroom differences in emphasis on speeded performance and drill of numerical facility.
- Irregularities in test administration when tests are given with very short time limits, and variations in test-taking discipline and motivation. These, however, are more likely to have resulted in decreased reliability of the most speeded tests than having affected the common factor Gs'_B.

Humphreys et al. (1979) interpreted their speed factor in quite a different way, and the sources discussed previously are also more likely to affect individual classrooms than whole schools. They hypothesize that the speed factor represents largely black high schools. This is based on the finding that black students tend to perform comparatively better in speeded tests provided that wrong answers are not penalized in scoring. Thus, a kind of demographic interpretation is also proposed for the speed factor, which is hardly relevant for our more homogeneous population.

In the introduction, we mentioned the possibility that the selective mechanisms behind a general factor at group level may primarily operate on Gc rather than on Gf. A comparison of the two levels of analysis in this study supports such a hypothesis. The G_B is strong in verbal and numerical tests, typical for Gc, and comparatively weak in perceptual tests. The G_w has its strength in the inductive tests, and it seems to coincide with the Gf factor. The two levels of analysis, thus, are clearly not interchangeable: a conclusion already drawn by Cronbach (1976) and empirically supported in Härnqvist (1978).

Almost all earlier generalizations about the factorial structure of abilities have been based on single-level analyses of individual data, such as the factor analysis reported in Table 1. This structure has great similarities with the within structure in MFA, although with some reduction in the within loadings due to the reduced variability. But if they differ, which one is likely to give the "best" picture?

This question seems fairly easy to answer when it comes to abilities. The overall structure is likely to be of primary interest because there is no good reason why the structure of intelligence should be determined with the exclusion of important demographic and self-selection sources of variation that are part of "natural" variation in abilities. On the other hand, the MFA permits a more distinctive interpretation of results, and it gives the correct statistical tests when the sample studied has been selected by means of cluster sampling. An estimate of the total factor solution can be put together by combining the levels in a multilevel analysis, which gives richer information than an overall factor analysis only.

The case is different for achievement measures and other variables for which treatment effects are likely to appear. Then, the structure within each treatment might be of more interest than the overall structure.

This choice is part of a broader set of decisions regarding the design and analysis of large-scale studies: (a) In what order of priority should different sources of variation be considered? (b) When is it time to stop the breakdown process? (c) What level of analysis gives the most relevant information? Questions like these have to be answered in each study and should be answered preferably on the basis of theory rather than convention.

REFERENCES

Cattell, R.B. (1943). The measurement of adult intelligence. *Psychological Bulletin, 40,* 153–193.

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.

Cronbach, L.J. (1976). *Research on classrooms and schools: Formulation of questions, design, and analysis.* Stanford, CA: Stanford University, Evaluation Consortium.

Cronbach, L.J., & Snow, R.E. (1977). *Aptitudes and instructional methods.* New York: Irvington.

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

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

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). Hillsdale, NJ: Erlbaum.

Gustafsson, J.-E. (1989). Broad and narrow abilities in research on learning and instruction. In R. Kanfer, P.L. Ackerman, & R. Cudek (Eds.), *Abilities, motivation and methodology: The Minnesota Symposium on Learning and Individual Differences.* Hillsdale, NJ: Erlbaum.

Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28,* 407–434.

Härnqvist, K. (1960). *Individuella differenser och skoldifferentiering* [Individual differences and school differentiation]. Stockholm: Statens offentliga utredningar.

Härnqvist, K. (1962). *Manual till DBA (Differentiell begåvningsanalys)* [Manual of DBA (Differential analysis of ability)]. Stockholm: Skandinaviska Testförlaget.

Härnqvist, K. (1978). Primary mental abilities at collective and individual levels. *Journal of Educational Psychology, 70,* 706–716.

Härnqvist, K., Gustafsson, J.-E., Muthén, B.O., & Nelson, G. (1991, April). *Hierarchical models of ability at class and individual levels.* Paper presented at the meeting of the American Educational Research Association, Chicago, IL.

Horn, J.L. (1976). Human abilities: A review of research and theory in the early 1970s. *Annual Review of Psychology, 27,* 437–485.

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.* Urbana: University of Illinois Press.

Humphreys, L.G., Parsons, C.K., & Park, R.K. (1979). Dimensions involved in differences among school means of cognitive measures. *Journal of Educational Measurement, 16,* 63–76.

Jöreskog, K.G., & Sörbom, D. (1986). *LISREL VI: User's guide.* Mooresville, IN: Scientific Software.

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

Muthén, B.O. (1988). *LISCOMP: User's guide.* Mooresville, IN: Scientific Software.

Muthén, B.O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54,* 557–585.

Muthén, B.O. (1990). *Multilevel factor analysis of class and student achievement components* (UCLA Statistics Series No. 76). Los Angeles: University of California.

Muthén, B.O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28,* 338–354.

Muthén, B.O., & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology, 38,* 171–189.

Nelson, G., & Muthén, B. (1991). *Analysis preparation steps for multilevel factor analysis using SOURCE.BW and LISCOMP.* Unpublished manuscript, University of California, Los Angeles, Graduate School of Education.

Thurstone, L.L. (1938). *Primary mental abilities* (Psychometric Monographs No. 1).

Undheim, J.O. (1981). On intelligence: II. A neo-Spearman model to replace Cattell's theory of fluid and crystallized intelligence. *Scandinavian Journal of Psychology, 22,* 181–187.

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

Vernon, P.E. (1961). *The structure of human abilities* (2nd ed.). London: Methuen.

APPENDIX A
True and Error Variances Within and Between Levels in Grades 4–6 (L) and 7–9 (H)

	Grades 4–6				Grades 7–9			
	WF_L	WE_L	BF_L	BE_L	WF_H	WE_H	BF_H	BE_H
Synonyms	10.27	2.79	3.31	0.00	5.86	1.77	2.95	0.05
Opposites	8.23	2.94	3.22	0.05	4.78	2.04	2.95	0.08
Letter Grouping	6.22	1.98	1.64	0.02	5.74	1.58	2.70	0.00
Figure Series	8.73	2.37	1.52	0.04	7.91	1.90	2.10	0.04
Metal Folding	7.89	2.18	1.45	0.02	7.56	2.50	1.87	0.00
Block Counting	8.10	2.76	1.04	0.06	9.23	3.04	1.58	0.00
Additions	6.00	1.46	1.92	0.09	4.56	1.11	1.77	0.13
Multiplications	6.14	1.04	1.97	0.18	5.24	1.15	2.32	0.32
Identical Numbers	4.52	1.38	1.41	0.30	4.14	1.42	1.17	0.69
Highest Number	5.64	1.73	3.15	0.70	5.76	2.20	2.17	1.06

Note.	Reliability within:	$WF/(WF + WE)$
	Reliability between:	$BF/(BF + BE)$
	True intraclass correlation:	$BF/(BF + WF)$
	True change within:	$(WF_H - WF_L)/WF_L$
	True change between:	$(BF_H - BF_L)/BF_L$

APPENDIX B
Standard Deviations of Scores From Parallel Forms Within and Between Classes

	Between Classes		Within Classes	
	Grades 4–6	Grades 7–9	Grades 4–6	Grades 7–9
SYN1	3.63	2.79	8.80	8.05
SYN2	3.50	2.62	8.69	8.12
OPP1	3.35	2.62	7.62	7.69
OPP2	3.37	2.68	7.21	8.32
LETT1	2.85	2.68	5.84	7.55
LETT2	2.88	2.73	5.75	7.97
FIG1	3.34	3.10	6.18	7.30
FIG2	3.30	3.17	6.04	7.13
METF1	3.11	3.19	5.67	6.46
METF2	3.21	3.17	6.06	7.00
BLOCK1	3.53	3.69	5.55	6.82
BLOCK2	3.03	3.30	5.15	6.27
ADD1	2.60	2.32	5.62	6.77
ADD2	2.90	2.33	6.42	6.39
MU1	2.44	2.48	5.88	6.91
MU2	2.64	2.39	6.66	7.71
INDENT1	2.20	2.35	5.21	5.50
INDENT2	2.56	2.23	6.20	6.22
HIGH1	2.80	2.99	7.89	9.16
HIGH2	2.99	3.10	8.65	8.68

Note. In order to facilitate the computations, the individual raw scores in ADD, MULT, and INDENT were divided by 3 and the raw scores in HIGH by 4. Ad hoc estimators: Grades 4–6 = 3.912; Grades 7–9 = 4.385.

APPENDIX C
Correlations Within Classes for Grades 4–6

SYN1																			
SYN2	78																		
OPP1	69	70																	
OPP2	65	66	74																
LETT1	27	31	34	34															
LETT2	32	33	37	39	76														
FIG1	32	34	37	38	38	38													
FIG2	33	35	38	39	40	39	79												
METF1	32	31	32	34	31	30	44	45											
METF2	33	31	35	36	35	35	46	47	78										
BLOCK1	28	28	30	30	28	29	37	37	46	48									
BLOCK2	28	27	28	30	26	27	36	36	43	47	75								
ADD1	18	19	22	25	29	31	19	18	09	13	16	15							
ADD2	19	21	22	26	28	30	17	16	09	14	16	13	81						
MU1	22	24	22	28	26	29	17	18	06	09	09	08	67	66					
MU2	22	26	26	30	28	30	18	20	07	10	11	09	72	73	84				
INDENT1	07	09	08	09	25	26	09	10	−01	−01	06	06	46	46	40	42			
INDENT2	08	10	08	10	23	24	07	06	−02	00	05	05	46	47	42	44	76		
HIGH1	10	11	12	14	20	23	10	10	02	03	09	10	56	54	47	49	52	54	
HIGH2	12	14	13	15	24	28	14	13	04	06	09	07	59	59	52	56	56	60	79

Note. Decimal points omitted.

APPENDIX D
Correlations Within Classes for Grades 7–9

SYN1																			
SYN2	76																		
OPP1	68	67																	
OPP2	64	64	72																
LETT1	27	26	33	34															
LETT2	22	24	30	30	78														
FIG1	36	33	38	38	44	41													
FIG2	33	31	36	36	45	43	81												
METF1	23	24	29	26	28	25	34	37											
METF2	30	29	33	33	33	30	39	41	75										
BLOCK1	22	22	27	27	34	34	37	37	44	50									
BLOCK2	24	23	28	26	30	30	36	36	46	50	75								
ADD1	05	07	05	06	16	19	04	03	−06	−02	05	00							
ADD2	04	06	05	05	18	20	06	05	−05	−01	07	01	79						
MU1	05	09	06	07	12	15	02	00	−11	−08	00	−07	68	62					
MU2	06	08	07	06	10	14	−01	−01	−09	−07	−01	−06	70	69	81				
INDENT1	00	05	04	03	13	14	01	02	−04	−02	01	−02	40	40	39	37			
INDENT2	−01	02	00	−01	09	11	−02	00	−06	−06	−01	−05	40	42	38	40	73		
HIGH1	01	05	02	04	14	18	03	03	−03	−01	12	05	50	49	41	41	53	49	
HIGH2	01	04	04	03	13	17	06	05	−04	00	10	04	51	52	44	46	51	51	76

Note. Decimal points omitted.

APPENDIX E

Correlations Between Classes in Grades 4–6

SYN1																			
SYN2	96																		
OPP1	95	95																	
OPP2	93	93	93																
LETT1	74	75	78	75															
LETT2	74	76	79	74	93														
FIG1	68	68	67	67	69	68													
FIG2	74	76	74	74	76	75	92												
METF1	77	75	79	76	69	66	76	76											
METF2	79	75	79	77	69	66	75	76	93										
BLOCK1	59	59	61	57	62	64	75	70	75	74									
BLOCK2	64	63	66	62	73	71	70	71	76	77	87								
ADD1	64	66	70	64	65	63	46	52	56	54	51	56							
ADD2	70	74	76	71	69	66	55	58	63	60	55	60	92						
MU1	71	74	73	69	59	61	51	57	58	57	50	54	81	81					
MU2	68	72	70	66	66	65	51	57	55	53	53	57	82	86	90				
INDENT1	67	69	71	62	63	63	47	52	52	51	50	49	73	76	74	78			
INDENT2	61	63	64	59	60	61	43	49	45	41	48	51	73	78	75	76	83		
HIGH1	61	65	64	58	59	63	54	56	59	55	56	59	72	76	68	68	76	68	
HIGH2	67	67	68	63	68	67	53	57	61	58	58	65	76	79	67	73	81	75	81

Note. Decimal points omitted.

APPENDIX F
Correlations Between Classes in Grades 7–9

SYN1																			
SYN2	96																		
OPP1	95	96																	
OPP2	93	95	95																
LETT1	74	77	76	78															
LETT2	75	79	76	76	97														
FIG1	78	79	79	80	85	85													
FIG2	77	77	76	77	84	83	95												
METF1	79	79	79	78	65	64	68	69											
METF2	83	81	79	79	68	68	73	76	94										
BLOCK1	65	64	60	63	72	74	70	70	68	73									
BLOCK2	71	71	67	67	72	73	71	73	74	79	93								
ADD1	67	71	63	73	68	69	61	59	54	53	50	51							
ADD2	72	75	67	76	66	66	58	57	54	53	43	46	92						
MU1	70	71	64	71	66	66	57	57	56	55	48	51	91	89					
MU2	78	80	75	80	73	74	68	67	63	61	53	55	87	87	85				
INDENT1	58	59	55	63	51	50	50	48	47	46	37	38	82	76	81	74			
INDENT2	52	56	52	53	47	49	41	41	40	40	30	34	53	53	54	67	58		
HIGH1	30	33	25	31	26	26	28	25	11	18	11	17	60	58	54	47	57	43	
HIGH2	56	60	51	56	49	49	49	48	41	41	36	39	80	80	71	72	67	44	69

Note. Decimal points omitted.