DOCUMENT RESUME

ED 219 437 TM 820 455

AUTHOR	Gustafsson, Jan-Eric; And Others
TITLE	A General Model for the Organization oi Cognitive Abilities. [Report] 1981:06.
INSTITUTION	Goteborg Univ., Molndal (Sweden). Dept. of Education.
SPONS AGENCY	National Swedish Board of Education, Stockholm.; Swedish Council for Research in the Humanities and Social Sciences, Stockholm.
PUB DATE	81
NOTE	138p.
EDRS PRICE	MF01/PC06 Plus Postage.
DESCRIPTORS	Ability Identification; *Cognitive Ability; Data Analysis; *Factor Structure; Grade 6; *Intelligence; Intermediate Grades; *Models; Sex Differences; Visualization
IDENTIFIERS	*Cattell Horn Fluid and Crystallized Ability Theory; Crystallized Intelligence; Exploratory Factor Analysis; Fluid Intelligence; LISREL Computer Program

ABSTRACT

The report addresses the question of the structure of cognitive abilities, i.e., how many different abilities should be identified and the relations between abilities. Models suggested by Spearman, Thurstone, Guilford, Vernon, Cattell-Horn and others are reviewed. By combining features of the Vernon and Cattell-Horn models it is possible to construct a more general model, of which most other models are special cases. An empirical study is presented in which a test battery of 16 tests was administered to some 1200 pupils in the sixth grade. The test battery was designed to reflect the Cattell-Horn factors fluid intelligence (Gf), crystallized intelligence (Gc), and general visualization (Gv). A sequence of LISREL-models is presented in which the suggested model for the organization of abilities is tested. Good support is obtained for the model. In relation to interpretations of G (or Gf), it is proposed that this factor represents the ability to create and execute new assemblies of processes. Gc and Gv ara interpreted to reflect the ability to process verbal and figural information, respectively, both as a function of the specific processing requirements posed by these types of information, and as a consequence of previously acquired knowledge. (Author/PN)

1981:06

A general model for the organization of cognitive abilities

Jan-Eric Gustafsson

Berner Lindström

Eva Björck-Åkesson

(The research presented in this paper has been supported financially by the National Board of Education in Sweden and by the Swedish Council for Research in the Humanities and Social Sciences)

Contents

SUMMARY

The report addresses the question of the structure of cognitive abilities, i.e. how many different abilities should be identified and the relations between abilities. Models suggested by Spearman, Thurstone, Guilford, Vernon, Cattell-Horn and others are reviewed. It is noted that while some models include a general factor (G), others do not. Another difference is that in some models all abilities are placed at the same level, while in other models some abilities are subsumed under others in a hierarchical pattern. It is pointed out, however, that by combining features of the Vernon and Cattell-Horn models it is possible to construct a more general model, of which most other models are special cases.

Problems associated with exploratory factor analysis are discussed and it is concluded that the LISREL technique of Jöreskog and Sörbom may be particularly well suited to test the model. Two reanalyses of published data are presented, which both provide good support for the suggested integration of models.

An empirical study is presented in which a test battery of 16 tests was administered to some 1200 pupils in the 6th grade. The test battery was designed to reflect above all the Cattell-Horn factors fluid intelligence (Gf), crystallized intelligence (Gc), and general visualization (Gv). The report describes the test battery in rather great detail and presents analyses of each of the tests. Above all, however, a sequence of LISREL-models is presented in which the suggested model for the organization of abilities is tested. Good support is obtained for the model.

Sex differences are also analyzed with the LISREL technique. The sex differences in level of performance match a commonly established pattern with a higher performance for boys on numeric factors, and a higher performance for girls on verbal factors. In addition there are some sex differences in the variance of factors, factor loadings, and error variances of tests.

In relation to interpretations of G (or Gf) suggested by Simon and Snow it is proposed that this factor represents the ability to create and execute new assemblies of processes. Gc and Gv are interpreted to reflect above all the ability to process verbal and figural information, respectively, both as a function of the specific processing requirements posed by these types of information, and as a consequence of previously acquired knowledge.

1 INTRODUCTION

Research within the field of differential psychology has traditionally been concerned with questions such as the organization of human abilities, group differences in ability, and with the etiology of individual differences. Lately, however, research on individual differences has sought partly new directions, in attempts to seek a deeper understanding of ability in terms of process constructs derived from cognitive psychology.

One of the reasons for this development is a dissatisfaction with the highly empirical, atheoretical nature of much of the differential psychological research (e.g. Resnick, 1976). Another reason is the insight that the road to further practical applications is blocked urless a better understanding of individual differences is achieved. Thus, one of the conclusions drawn from research on aptitude x treatment interactions (ATI; Cronbach & Snow, 1977) is that adaptive instruction requires knowledge of individual differences which goes much beyond a description of the individual's standing on different traits. For such purposes a formulation of individual differences in process terms would seem highly desirable (e.g. Glaser, 1972; Gustafsson, 1971, 1981; Snow, 1977, 1980).

The present research is conducted within a project with the primary aim of analyzing and describing individual differences in learning strategies. The purpose thus is to identify "new aptitudes" (Glaser, 1972) useful in diagnosis of learning problems, and selection of instructional strategies. We share, however, Snow's (1980) conviction:

> that the new and the old (aptitudes) will be found to differ more in form than in kind, and that an improved conception of human learning and cognition will need to be built on their combination.... Since it is the old aptitudes that consistently predict learning from instruction, that is the place I think it best to start. The object is to convert existing aptitude constructs into more detailed models of individual differences in cognitive processing, and to trace the operation of these through the activities involved in instructional learning (Snow, 1980)

In designing the research project we have, therefore, selected for further research what seem the most promising variables developed in traditional differential psychological research. These variables serve as reference variables against which new constructs may be evaluated. But they also represent constructs to be accounted for by other, theoretically more potent, constructs.

The question of the structure of human abilities, and above all the question whether intelligence is unitary or multifacetted, has been of central concern in the research on "old" aptitudes, and so it should be in the research on "new aptitudes" as well. Quite a few different models of the organization of abilities have been suggested during the decades of differential psychological research, which each carries serious implications for how ability should be measured, for theory and for applications. Some are now of historic interest only, but several have survived into our days. However, while the rather intense battle fought between the adherents of different models appears to have ceased, it seems that the question which model is superior never got an answer. In the present report we make yet another attempt to answer this question.

In addition to this purpose, the report documents a reference material of tests and subjects to be used in further studies of individual differences in learning.

2 MODELS OF THE STRUCTURE OF ABILITIES

We will start by presenting a brief review of the history of differential psychology, along with a description of the most influential models of the organization of abilities.

2.1 The first steps of differential psychology

Differential psychology has, like so many other lines of psychological research, its roots in late nineteenth century. One of the first contributors was Sir Francis Galton. Inspired by Darwin's theses about the evolution of species, he conducted family pedigree studies of the inheritance of talents in various fields of work (Galton, 1869). He found that eminence tends to run in families, and concluded that genius is inherited. (Thereby overlooking the fact that family members have not only genes, but also environment in common).

But Galton also was a prolific contributor of tests and measurement techniques. These early instruments assessed in particular the acuity of sensory processes, such as reaction time, sensory discrimination and motor capacity. Sensory processes were not concentrated upon because of their intrinsic interest but because it was thought that they would provide an avenue to estimate an individual's intellectual level. Galton (1883) wrote:

> The only information that reaches us concerning outward events appears to pass through the avenue of our senses: and the more perceptive the senses are of difference, the larger is the field upon which our judgement and intelligence can act (p. 27).

Thus, Galton saw the sensory and intellectual processes as being quite distinct from one another, but for purposes of indexing the latter he concentrated upon the former.

Galton not only furnished differential psychology with tests and measures, but he also contributed to the statistical theory of correlation. As we shall see later on an interest in differential psychology often goes hand in hand with an interest in statistical techniques for describing and analyzing concomitant variation.

In late nineteenth century experimental psychology got established. While the experimentalists, like Wundt, did not pay much attention to individual differences, the procedures they worked with set standards also for research on individual differences. For example, the American J. McKeen Cattell (see Cronbach, 1960, p. 158) used a blend of procedures from Wundt's and Galton's laboratories to measure sensory acuity, strength of grip, memory for dictated consonants and so on, in an effort to identify talented individuals. It was soon to be discovered, however, that these measures did not forecast scholastic success (Wissler, 1901), which did put a rather abrupt end to the attempts to base an applied psychology on psychophysical measures of individual differences.

2.2 Binet and tests of general ability

The major early breakthrough in differential psychology instead came in quite another line of attack. This approach, the leader of which was the French physician Alfred Binet, was characterized by an orientation towards measurement of complex psychological processes, and it was strongly application oriented.

A paper by Binet and Henri (1895) provided a programmatic statement for research on individual differences. They argued that the research so far had been too concerned with simple psychophysical measurements, and suggested that performance on complex mental tasks should be assessed instead. While this could be assumed to be considerably more difficult to do with precision, Binet and Henri expected the range of individual differences in such aspects to be greater, which would simplify the task.

They suggested two main methods for studying individual differences in complex psychological processes. In the first of these, one studies the degree of association of a larger number of processes; and in the other attempts are made to change one process, to see the effects on other processes. Binet, and other differential psychologists to follow him, came to adopt in particular the first line of approach, while the second, with its combination of manipulation and observation, was not systematically conducted until ATI-research was instigated (Cronbach, 1957; Cronbach & Snow, 1977).

In 1905 Binet and Simon came up with the first intelligence test. This scale consisted of a rather motley collection of tasks, such as naming objects in a picture, discriminating two lines for length, memory span, defining simple words, folding and cutting paper, and completing sentences, just to mention a few examples. Attempts, however, were made to represent in the test categories such as judgement, common sense, initiative, and ability to adapt, and each task was included on the basis of its ability to differentiate between different age groups and between "bright" and "dull" groups of subjects.

The test met with almost immediate practical success and it was soon to be followed by revised editions, and by translations into other languages. One of the translators was Lewis Terman at Stanford University, and in 1916 he published the Stanford-Binet revision of the Binet-Simon test. The Stanford-Binet was widely accepted and it set a standard for other intelligence tests soon to be developed. One reason for this may be that it introduced the easily understood IQ concept (see Cronbach, 1960, p. 161).

Ever since, tests of general mental ability have been heavily used -- too heavily according to several critics of testing. The basis for the frequent application of this type of tests is, of course, that they have been shown empirically to work; in particular the validity coefficients for predicting school achievement are quite high. Otherwise, development in the area of tests of general mental ability has been slow:

> Ability tests have remained about the same since 1920 ... The practical tests of today differ from the tests of 1920 as todays's automobiles differ from those of the same period: more efficient and more elegant, but operating on the same principles as before (Cronbach, 1960, p. 159).

The principle of the intelligence tests, it might be reiterated, was thus simply to assemble items which discriminate between age groups, and between successful/unsuccessful groups of performers. No theory was developed, however, to account for why the items have such predictive capacity. Binet himself did not conceive of intelligence as a separate capacity, but rather as comprised of several more or less clearly identifiable capacities, such as judgement, common sense, initiative and ability to adapt. It can be argued, of course, that such a theoretical position is incompatible with the notion of rank-ordering individuals along a single dimension. Spearman (1927) levelled such criticism against Binet:

> although in actual testing he took account of his "general level" alone, still in all his theoretical psychology continued to rely altogether upon his old formal faculties, not withstanding that these and the "general level" appear to involve doctrines quite incompatible with each other (Spearman, 1927, p. 60).

Binet thus seemed to conceive of the score obtained on an intelligence test as some kind of average of several different capacities. However, from both a theoretical and a practical point of view this position is fraught with problems: The model of the structure of abilities is implicit rather than explicit, and the sampling of tasks from different domains is more or less arbitrary.

2.3 Spearman and the theory of Two Factors

The first explicit, empirically based, model of the structure of human abilities was contributed by Spearman (1904b, 1927), who also brought statistical sophistication to the emerging line of differential psychological research.

The Spearman model has its roots in the attempts to measure individual differences in intelligence by psychophysical assessments. It will be remembered that this line of research had proven quite unproductive. Spearman (1904a) showed, however, that errors of measurement tend to cause underestimation of the true correlation between variables. He developed techniques to correct for the underestimation, and applied these in yet another study of the relation between performance on laboratory tasks (e.g. light-, weight-, and pitch-discrimination) and independent ratings of intelligence. Spearman found a no less than perfect correlation between an underlying variable common to the ratings of intelligence and an underlying variable common to the sensory discrimination tasks. The conclusion was thus drawn "that the common and essential element in the Intelligence wholly coincides with the common and essential element in the Sensory Functions" (Spearman, 1904b, p. 269).

The method applied by Spearman in the analysis of data was the first factor analytic model -- in essence he showed that one common factor is sufficient to account for the intercorrelation among variables. By present standards his technique must be considered crude and it may, of course, be suspected that imperfections in the analysis caused Spearman to draw the conclusion quoted above. However, a reanalysis of parts of the Spearman data (Series I) with confirmatory factor analysis (see Appendix 1) supports the Spearman conclusion: The three sensory discrimination tasks and the three independent ratings of intelligence do fit a one-factor model (chi-square=8.1, df=9, p <.53). The only aspect of Spearman's conclusion which is challenged by the results obtained in the reanalysis is the statement that the common underlying variable is the essential one: the correlations between the sensory discrimination variables and the latent variable range between .37 and .54, which implies that only some 14 to 30 per cent of the variance in the sensory discrimination variables is due to the common factor.

On the basis ot these re-ults, and similar results in other studies, Spearman proposed the Two Factor theory, which states that performance on an intellectual task is affected by two factors only, one general (g) and one specific (s). The g factor enters more or less prominently into any intellectual activity, but the individual's standing on this factor is the same irrespective of task. The s factor is of great importance when the g factor is of little importance, and vice versa. However, for each type of intellectual activity a different s factor is assumed, so the individual's standing on this factor varies with the task.

Spearman (1927) also worked out a mathematical-statistical theory, based on the criterion of "tetrad differences", with which the relative importance of the g and s factors could be estimated for any task, and with which the Two Factor theory could be subjected to empirical tests. Results from a large number of empirical studies indicated that the Two Factor model showed a very good fit to data. In some studies it was found, however, that the s factors were not orthogonal but correlated, and thus gave rise to group factors. It also was found that when tests that were too similar were included in a test battery, this caused a correlation among the s factors.

Spearman (1927) contrasted the Two Factor theory with three rival doctrines in the understanding of individual differences: the "monarchic", the "oligarchic" and the "anarchic". The monarchic doctrine "assumes mental ability to lie under the sovereign rule of one great power named 'intelligence'" (Spearman, 1927, p. 4), and with the Binet-type of intelligence tests this doctrine got practical impact. Spearman argued, however, that the doctrine lacks theoretical foundation. the several attempts made to define the concept "intelligence" provide at best more or less loose "statements about" intelligence, and lack usefulness in both practical and theoretic work. Furthermore, several of the "definitions" of intelligence are essentially lists of more of less independent faculties, which would logically imply separate measurements of the faculties. According to Spearman the monarchic doctrine of intelligence thus is so fraught with problems that it should be abandoned.

The oligarchic doctrine assumes that there are several different powers (abilities), each of which constitutes a separate function, which can and should be separately measured. Spearman argued, however, that this position too is fraught with problems. For if it can be argued that intelligence must be split into several components, then it can also be argued that these components should be split further, and so on. Spearman (1927, p. 35) stated:

> Take, for instance, judgement. This, too seems to break up into several different kinds. Judgement for politics would appear to be

> one thing; that for sports, another; that for telepathy, yet a third; and so on. Is not then, here also, a separate measurement needed for each kind? Our answer must again be that all different kinds certainly require separate measurements unless they can be shown to be perfectly interdependent, so that the person who excels in any one kind does so to just the same extent in all others. (Spearman, 1927, pp. 35-36).

This would lead to an almost infinite regress, and as we will see later on this is almost exactly what has happened in certain lines of research on the structure of abilities.

In the anarchic doctrine it is held that individual differences in mental ability are caused by individual differences in very many more or less independent abilities. Within the anarchic position, Spearman made a distinction between two different positions. In one of these it is held that ability is subdivided into innumerable independent parts; in the other it is held that while ability is subdivided into very many aptitudes, these are correlated and stand in complex relationships with one another. Spearman claimed, however, that the first of these must be rejected on the basis of empirical findings, and the second he regarded as "true but sterile" (Spearman, 1927, p. 70).

A corollary to the anarchic position is that even though it may not be possible to measure each of the abilities, a sample can be made in order to determine a "general level" as an average. According to Spearman this was the approach taken by Binet, following a recommendation by him in the 1904b paper. There are problems, however, in the conception of general mental ability as an average of several abilities: the domain to be sampled from must be determined; the sampling must be representative; the problem of compatibility of scale units must be solved, and so on. According to Spearman "No genuine averaging, or sampling, of anybody's abilities is made, can be made, or even *has really been attempted".* (Spearman, 1927, p. 71, emphasis in original).

Spearman viewed the Two Factor theory as an eclectic construction, taking the best from each of these three incompatible doctrines:

> a certain amount of truth is to be found in *each* of the three great rival doctrines... Thus, the "monarchic" view is justified by g if we admit this ruler to be constitutional, not despotic: it forms a mighty factor in the state, but not the sole one. And a further truth -- qualifying and restricting the other -- is contained in the "anarchic" view. For besides the factor g which rules throughout all mental processes, there is also the factor s which is in every process independent; under the monarchic reign there is

> still some freedom for the individual citizens. And as much may be said, finally, for the third or "oligarchic" view, seeing that something of the nature of faculties or types -- quite distinct from the universal factor and fairly distinct from the ordinary narrow factors -- has revealed itself in what we have been calling the broad factors (Spearman, 1927, p. 84).

Spearman avoided a formal definition of g; instead the empirical attitude was adopted that g is whatever is contained in the factor. "All else about it -- including the question as to whether it has the least right to be regarded as a genuine measure of 'intelligence' -- lies still before us to ascertain" (Spearman, 1927, p. 161).

As one step towards the elucidation of g, he proposed a set of three laws to account for cognitive phenomena. The first law -- apprehension of one's own experience -- says that persons have more or less power to observe what goes on in their minds. The second law -- eduction of relations -- states that persons have more or less power to find relations between ideas. The third law -- the eduction of correlates -- "enounces that when a person has in mind any idea together with a relation, he has more or less power to bring up onto mind the correlative idea" (Spearman, 1927, p. 166).

Spearman studied empirically the involvement of g in a large number of specific classes of relations and found that this involvement is very large indeed. He concluded, therefore, that the best measure of g is provided by tests which involve eduction of correlates and relations, and especially so when the content is abstract; abstraction was by Spearman regarded as the "climax of eduction".

At a theoretical level, however, Spearman argued against the idea to reduce g to a power to grasp relations. For one thing this would involve only the second law, and leave beyond the scope of g the other two laws. For another thing such an explanation of g would be framed in terms of mental operations "whereas our g, as we have seen, measures only a *factor* in any operation, not the whole of it" (p. 80). Spearman favored, therefore, another interpretation of g and argued that it is "... an underlying something which -- by analogy with physics -- has been called mental energy". (Spearman, 1927, p. 89).

The hypothesis of mental energy was inspired by observations concerning "universal mental competition", i.e. one cognitive act interferes with another cognitive act, perception interferes with thought, thought interferes with perception, emotion interferes with perception and thought and so on. Thus "... the maximal output for *each kind of activity* is *not* constant, but becomes changed and lowered by any simultaneous occurrence of other *activities* " (Spearman, 1927, p. 111, emphasis in original). This was interpreted to mean "... that all the mental activity, just like the physical, consists in ever varying manifestations of one and the same underlying thing, to which may be given the name of energy" (Spearman, 1927, p. 133). Spearman also suggested that this energy fuels "engines" which correspond to specific neural structures, and the efficiency of which is reflected in the s factors.

These interpretations of the g- and s-factors are, unfortunately, quite vague, and they did not much influence further research on ability.

One of Spearman's greatest contributions was the mathematical/statistical theory upon which the doctrine of Two Factors is based. This was the first factor-analytical model, and it may even be regarded the first confirmatory factor-analytic model since it affords a test of the fit of data to the model. This possibility of actually testing whether the model fits the data was the great strength of the model, and upon which Spearman heavily relied in making his quite far-reaching conclusions about the generality of the theory of Two Factors.

As has already been pointed out Spearman did find a very good fit between observational data and the model. But there were also deviations. For one thing it was, as has already been pointed out, in some cases found that the s factors were correlated, thus giving rise to group-factors. For another thing it was found that the model broke down when tests that were "too similar" were included in the test battery, again because of a correlation between the s factors. These facts were readily admitted by Spearman, but they came to cause great problems for his theory when other researchers confronted it with data.

The problem is, of course, how it should be decided when tests are too similar to be included in the same test battery, and how it can be decided if a group-factor is so important as to disturb the theory. Spearman's advice in these matters was quite clear: "... this is a point not to settle intuitively, but to ascertain by experiment. Performances should be regarded as quite different .. so long as the tetrad equation is satisfied and no longer" (Spearman, 1927, p. 80).

This piece of advice was most unfortunate, however, since adoption of it implies the assumption that the model is correct; the model can never be rejected if deviations may be blamed on imperfections in the observational basis. It was no wonder, therefore, that the theory of Two Factors was soon to be thoroughly refuted by other researchers (e.g. Kelley, 1928).

By present standards the factor-analytic model with which Spearman operated is of course much too simplified, and more elaborate factor analytic procedures were soon to be developed, which were capable of representing systematic variance of immensely greater complexity. But as we shall see this development lost two great advantages of the Spearman approach: its concentration upon one factor of momentous importance, and the model-testing capabilities of the statistical technique.

2.4 Models based on Multiple Factor analysis

The Spearman factor analytic model assumes that the correlations among a set of observed variables are accounted for by relations between the observed variables and one common factor. Primarily through the efforts of Thurstone (1938, 1947) factor analysis was extended, however, to encompass multiple common factors. Along with the fact that in the American research there was a development towards larger and larger batteries of tests, this paved the way for a much more multi-facetted conception of the structure of abilities.

2.4.1 The Thurstone model

Thurstone (1938) applied his recently developed Multiple Factor analysis to a test-battery of 38 tests and found about a dozen factors. By locating the reference axes according to the principle of simple structure, which essentially states that any test should be affected by one factor only, it was found that each of the factors accounted for performance on only a subset of the tests in the battery. There was no sign of a general factor.

Most factors identified by Thurstone (1938) were replicated several times by Thurstone and his colleagues (e.g. Thurstone, 1940; Thurstone & Thurstone, 1941), and it was possible to set up a list of six or seven easily replicable Primary Mental Abilities (PMA´s).

The most important PMA´s were: Verbal Comprehension (V), involved in understanding of language and frequently found in tests such as reading, verbal analogies, and vocabulary; Word fluency (W), involved in the fluent production of language, and measurable by tests such as rhyming or naming words in a given category; Induction (I), involved in tests requiring the subjects to find a rule in complex material; Space (S), involved in manipulation of geometric or spatial relations; Perceptual Speed (P), involved in quick and accurate grasping of visual details; and Number (N), involved in quick and accurate arithmetic computations.

In the first set of factor analyses the PMA´s were kept orthogonal. But when test batteries were assembled to measure the PMA´s, it was found that the tests were intercorrelated, which did suggest that a general factor might be present after all. This led Thurstone to adopt instead an oblique factor analytic model, in which correlations were allowed among the PMA´s. These correlations could then be analyzed in another, so-called second-order, factor analysis. Thurstone and Thurstone (1941) conducted such analyses and they did, indeed, find a general factor in the second-order analysis, which factor was most highly loaded by the I-factor.

Studies conducted during the 1940´s and 1950´s by Thurstone and others paid, however, very limited attention to the general factor. Instead the list of factorially identified primary abilities was extended considerably, partly by showing that several of the original PMA´s were differentiable into more narrow factors, and partly by investigation of new domains.

There was, thus, a rapid proliferation of mental abilities, presenting the "consumer" of results from differential psychological research with a rather bewildering picture. From the 1950´s and onwards several attempts were, therefore, made to bring order to the multitude of factors.

French (1951; cf French, Ekstrom & Price, 1963) presented a survey of the research, trying to determine which factors were distinct and cross-identified in several studies. The list came to encompass some 60 factors measurable with aptitude and achievement tests. While some of the factors were broad and comprehensive, others were very narrow and circumscribed, and had to be considered the result of a subdivision of a broader factor. However, all factors were primary in the sense that they represented results from application of Multiple Factor analytic techniques to matrices of intercorrelations between tests. Therefore, all factors had to be placed more or less at the same level, and it was not possible to indicate a hierarchical pattern among the factors, such that certain groups of factors could be subsumed under other factors.

2.4.2 The Guilford model

Guilford (1967) took another approach to organize the factor-analytic findings, and to develop guidelines for further research. He introduced a model with three facets: operation, content and product, in terms of which tests and factors could be described. The operation facet includes 5 levels (cognition (C), memory (M), divergent production (D), convergent production (N), and evaluation (E)), the content facet 4 levels (figural (F), symbolic (S), semantic (M) and behavioral (B) content) and the product facet 6 levels (Units (U), classes (C), relations (R), systems (S), transformations (T) and implications (I)).

In this system each test can be uniquely identified as a combination of levels on the three facets, for example cognition of semantic units (CMU). It was also hypothesized that each combination of levels on the three facets defines a unique factor. This "Structure-of-intellect" (SI) model therefore predicts no less than 120 (i.e. 5 x 4 x 6) identifiable factors.

Guilford (1972) argued that each of the PMA's could be mapped into the SI-model (for example, the V-factor would correspond to the CMU-factor), and that in addition the model provides guidelines for constructing tests so that also the other cells in the model can be factorially identified. In the latest version of the model Guilford and Hoepfner (1971) claim identification of at least one factor (and sometimes more) in each of 98 of the cells in the SI model.

In the factor analyses conducted to test the SI-model, Guilford has favored orthogonal rotations. This implies an assumption that factors having levels in common on one or two facets are no more related than are factors which have no levels in common. Therefore, the SI-model does not afford a parsimonious description of abilities; to identify an ability the levels on all three facets must be identified.

It can be noted, however, that Guilford (1980) now is admitting the possibility of higher-order factors along the lines of the levels of the facets in the SI-model. However, the analyses conducted by Guilford and associates do not provide a basis for identifying such factors of greater generality; to do that oblique factor analysis in which the correlations among the factors at the cell level can be determined would be necessary. Such reanalyses of the Guilford data have not yet been conducted to any large extent, even though they have been called for several times (e.g. Cronbach & Snow, 1977).

2.4.3 Discussion

Other attempts at organizing the results from Multiple Factor investigations have also been made (e.g. Horn, 1977; Pawlik, 1966). In each of these reviews it has been concluded that there are dozens and dozens of more or less clearly identified primary factors.

Table 1 presents a compilation of the most well established primary factors according to French et al. (1963) and Guilford (1967), along with examples of the type of items which measures the factor.

It would seem, however, that anyone interested in practical application of the knowledge represented in Table 1 is bound to sense a feeling of bewilderment. The multitude of factors, along with the lack of a clear

Table 1. Some of the most well established primary factors.

Factor label		
French et al.	Guilford	Types of tasks reflecting the factor
Induction (I)	CSS	Series and classification items in which a rule is to be found and applied
-	CFR	Figural analogies items, as in the Raven Progressive Matrices test
Syllogistic reasoning (Rs)	EMI,EMR	Syllogism items
General reasoning (R)	CMS	Complex arithmetic reasoning items, in which understanding the structure of the problem is crucial
-	CMR	Verbal analogies
Verbal comprehension (V)	CMU	Vocabulary
Mechanical knowledge (Mk)	-	Mechanical knowledge
Visualization (Vz)	CFT	Transformations of complex stimulus configurations
Spatial orientation (S)	CFS-V	Transformations of simple stimulus configurations as wholes
Flexibility of Closure (Cf)	NFT	Finding a simple figural pattern within a complex pattern
Speed of Closure (Cs)	CFU-V	Uniting disparate stimulus elements to a whole
Word fluency (Fw)	DSU	Rapid naming of words
Expressional fluency (Fe)	DMS	Fluency in composing connected discourse
Associational fluency (Fa)	DMR	Producing words from a restricted area of meaning
Perceptual speed (P)	EFU	Rapid judgement of perceptual identity
Memory span (Ms)	MSU,MSS	Remembering letters or digits in correct order
Associative memory (Ma)	MSR	Paired-associates learning
Number facility (N)	NSI	Performing simple arithmetical operations rapidly

organization, makes it quite difficult to compose a test battery for an application. If, for example, the aim is to predict achievement in a geometry course it may be asked if it is necessary to represent in the battery the whole array of spatial factors, and if not, which ones to select. In most cases such decisions would have to be made on the basis of judgement of the resemblance of the tests and the criterion which is to be predicted, in which situation the factor analytic findings do not contribute much.

Nor do the results summarized in Table 1 contribute much to our understanding of the organization of abilities; the list is just a compilation of low-level empirical findings, without a theoretical superstructure to aid interpretation and analysis.

Several critics have pointed at the limited utility of Multiple Factor analysis for describing the structu~~r~~e of ability. Humphreys (1962), for example, argued that the basic problem is "... the tendency to think of factors as basic or primary, no matter how specific, or narrow or artificial the test behavior may be that determines the factor" (p. 475). In the limit, each factor is identified by a set of parallel tests, which implies that the factor analysis does not effect any reduction of information. From a similar point of view, Undheim (1981) argued "that the widespread application of multiple factor analysis in research on abilities has carried factor analysis far beyond its descriptive and conceptual limitations as a research tool".

2.5 Hierarchical models

One way to counter-act the almost endless splintering of factors in Multiple Factor analysis would be to allow the factors to be correlated, and then analyze the correlations among the factors with factor analytic methods. Such higher-order analyses would yield hierarchical models, in which factors at lower levels are subsumed under factors at higher levels.

Even though this type of analysis was introduced already by Thurstone it has not, for different reasons, gained popularity until quite recently. For a substantial amount of time another research tradition has, however, been concentrated upon hierarchical models of the structure of human abilities. This tradition has been strongest in England as a continuation of the Spearman tradition, and before we describe the elaborations of Multiple Factor analysis into higher orders, we will attend to the British work on hierarchical models.

2.5.1 The Burt and Vernon models

In the British research on abilities Multiple Factor analysis has not had great impact and the g factor did not vanish altogether, as it did in the American research. In the post-Spearman research it was soon to be discovered, however, that in addition to g there are also group factors of great importance. Factor-analytic techniques were developed, which from a matrix of intercorrelations extract first the g factor, and then group-factors of successively smaller breadth. These hierarchical group-factor techniques (e.g. Burt, 1941; Harman, 1967) thus have the advantage of being able to supply information both about a general factor and group-factors.

The first hierarchical model was suggested by Burt (1949), who used the model to organize the findings of a comprehensive summary of factor analytic results. Below a general factor the Burt model includes 4 levels (see Figure 1): (1) sensation, representing simple sensory processes in different modalities; (2) perception, including motor-capacity and perceptual processes; (3) association, with memory and habit-formation factors; and (4) relation perceiving, involving the apprehension and application of relations. At each of these levels Burt (1949) reported evidence of group-factors, which in some cases also split into sub-factors.

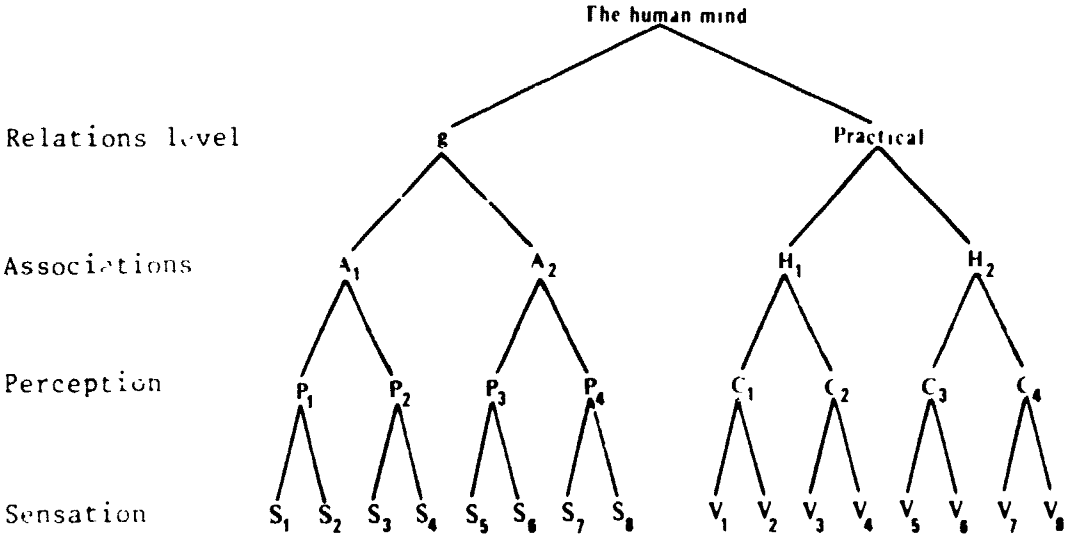


Figure 1. The Burt model.

The Burt model seems, however, to have been too much of a logically constructed classification scheme to earn any great impact. Instead a rather similar model, presented only slightly later by Vernon (1950, 1961, 1965) has received more widespread attention.

At the top of the Vernon model (see Figure 2) there is the g factor, and at the next level below there are two major group-factors: verbal-educational (v:ed) and spatial-practical-mechanical (k:m) ability. The v:ed factor subdivides into different scholastic factors, such as number factors and reading, spelling, linguistic and clerical abilities, and also into fluency and divergent thinking abilities. The k:m factor subdivides too and this complex includes minor group-factors such as perceptual, physical, psychomotor, spatial and mechanical factors. Each of these minor factors can then be subdivided by moie detailed testing.

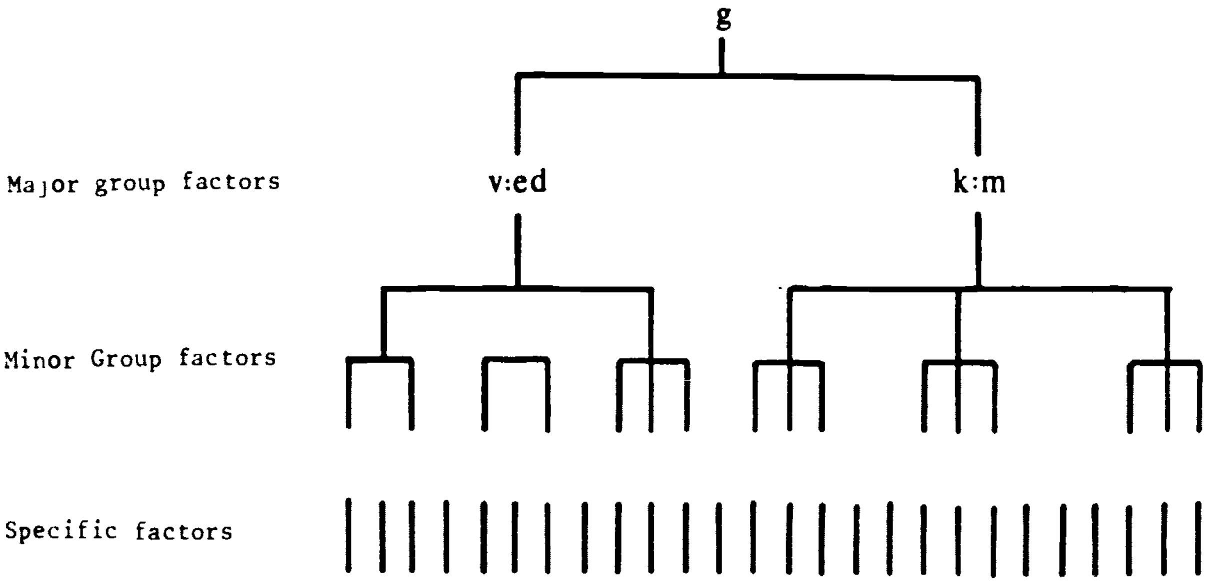


Figure 2. The Vernon model.

The Vernon model is, just like the Burt model, to a large extent a classificatory scheme summarizing the results from very many studies. Thus, with the factor analytic methodology employed by Vernon and others it is quite difficult to subject the model to encompassing and strict tests. It is, furthermore, quite difficult to use the model in such a way that individual scores on the factors at different levels are estimated and used as predictors; instead the model has most commonly been used for selecting and classifying tests.

2.5.2 The Cattell-Horn model

Another hierarchical model has been constructed by Cattell and Horn. The basic concepts in this model were developed by Cattell (1940, 1941, 1943), but the model was neither elaborated upon, nor put to empirical tests until considerably later (e.g. Cattell, 1963; Horn 1965, 1968; Horn & Cattell, 1966).

Methodologically the Cattell-Horn model is based upon oblique Multiple Factor analysis of several orders. Thus, in the first step an ordinary oblique Multiple Factor analysis is conducted, typically yielding a large set of primary or first-order factors. The correlations between the primary factors are then subjected to another factor-analysis which yields secondary or second-order factors. In principle this procedure of factoring at successively higher orders can be carried on until so few factors are obtained at a certain level that no factors can be identified. As we will see, however, Cattell and Horn have chosen to stop the factoring at the second-order level.

In the Horn and Cattell (1966) formulation the model includes 5 second-order or "general" factors:

Fluid intelligence (Gf) is reflected in tasks requiring abstraction, concept formation, and perception and eduction of relations, and in which tasks the stimulus material is either new or very familiar to the examinees. This kind of general intelligence is supposed to represent influences of biological factors and incidental learning on intellectual development. Gf is manifested in the primary abilities Induction (I), Cognition of Figural Relations (CFR), Cognition of Figural Classes (CFC), General Reasoning (R), Formal Reasoning (Rs), and also in primaries such as Memory Span (Ms), Associative Memory (Ma), and Number facility (N).

Crystallized intelligence (Gc) also is shown in tasks requiring abstraction, concept formation, and perception and eduction of relations, where, however, the stimulus material primarily is verbal-conceptual in nature. In contrast to Gf, Gc is is established through cultural pressures, education and experience. Gc is hypothesized to be manifested in primaries such as Verbal Comprehension (V), Mechanical Knowledge (Mk), Cognition of Semantic Relations (CMR), Cognition of Semantic Classes (CMC), R and Rs.

General visualization (Gv) is involved in "visualizing the movements and transformations of spatial patterns, maintaining orientation with respect to objects in space, unifying disparate elements and locating a given configuration in a visual field" (Horn & Cattell, 1966, p. 254). Gv is important in the primaries Visualization (Vz), Flexibility of Closure (Cf), Speed of Closure (Cs), Figural Adaptive Flexibility (DFT) and Spatial Orientation (S). It is also involved to a lesser degree in other primaries involving figural content, such as CFR, CFC and Mk.

General Fluency (F or Gr) is "reflected in tasks indicating flexibility in recalling and recognizing labels for cultural concepts irrespective of the subtlety of understanding of these" (Horn & Cattell, 1966, p. 254) and runs strongly through primaries such as Associational Fluency (Fa), Word Fluency (Fw), and Ideational Fluency (Fi).

General Speediness (Gs) is defined as quickness of performance, excepting the kind of performance defined by the Gr factor, and the quickness with which relations are perceived. Gs shows up most clearly in primary factors defined by very simple tasks, such as Perceptual Speed (P) and N.

The Cattell-Horn model has been subjected to empirical tests in large scale factor analytic studies by Cattell (1963), Horn (1972), Horn and Cattell (1966), Horn and Bramble (1967), and Undheim (1976, 1978), among others. These studies have in general confirmed the hypothesized structure. Humphreys (1967) concluded, however, from a reanalysis of the Horn and Cattell (1966) study that the model must be considered highly tentative. Lohman (1979) also reanalyzed the Horn and Cattell (1966) data, using nonmetric multidimensional scaling and hierarchical cluster analysis. These analyses failed to bring out Gf and Gv as separate factors. While Gs and Gr were traceable in some minor clusters of variables, Lohman considered these "general factors" as little more than "overblown primaries".

2.5.3 Comparisons between the models

As has been pointed out by Humphreys (1967) there are similarites between the Vernon and the Cattell-Horn models, but there also are important differences: The Cattell-Horn model lacks the g factor which has such a prominent place in the Vernon model and there are only two broad group-factors in the Vernon model, while there are 5 second-order factors in the Cattell-Horn model. These differences would seem so profound that if one of these models is accepted, the other must necessarily be refuted.

Cattell (1963) and Horn (1968), in particular, have argued in favor of their model. Horn (1968) stated:

> In Vernon's work, for example, a distinction is drawn between a broad "abstract" verbal-numerical-educational factor ... having properties similar to Gc, and an equally broad "practical" mechanical-spatial-physical factor..., which is somewhat similar to Gf.

and

> it is perhaps worth noting that the "abstract" versus "practical" distinction which is drawn to characterize the difference between v:ed and k:m is not used and is not appropriate for distinguishing between Gc and Gf.

Horn thus assumes that k:m in the Vernon must be some kind of mixture of the Gf and Gv factors in the Cattell-Horn model (cf Cattell, 1963). A similar assumption was made by Humphreys (1967) who argued that the answer to the question whether Gf and Gv are in fact distinct factors determines whether the Vernon or the Cattell-Horn models should be accepted. Many other researchers as well (e.g. Sternberg, 1980) have adopted the view that Gf roughly corresponds to k:m.

However, it would seem that the Cattell-Horn criticism of the Vernon model is founded on a misrepresentation of the model. Vernon himself does not accept the view that Gf should be more or less equated with k:m; he instead regards "... fluid ability as g with a slight mixture of spatial ability" (Vernon, 1969, p. 25).

Thus, an alternative view of the relationships between the two hierarchical models is that Gc corresponds to v:ed, that Gv corresponds to k:m, and that Gf corresponds to g. If it is in fact the case that Gf is more or less identical with g this would explain why there in the Vernon model is no major group-factor which corresponds to Gf: in the factoring technique used by Vernon the g factor is extracted first, and so much variance is extracted from the Gf tests that these fail to define a group factor in the next step of the analysis. The reason why Cattell and Horn have failed to notice that Gf corresponds to g is of course that they have stopped the factoring at the second-order level. If, however, a third-order factor (G) is determined it should be found that Gf has a loading close to unity in this factor.

If such relations between the Vernon model and the Cattell-Horn model can be substantiated, the remaining differences between the models seem quite minute and unimportant, and can most likely be settled in further empirical research.

2.5.4 Discussion

Hierarchical models offer several advantages as compared to models with abilities at one level only. They do allow a more parsimonious description of individual differences than do non-hierarchical models, since for any application the appropriate level of detail in measurement and interpretation can be chosen. For example, to predict overall school-achievement it would probably suffice to represent the top levels of the hierarchy, but if differential prediction between different tracks of study is wanted, measurement distinctions further down in the hierarchy can be made.

Even more important is the fact that hierarchical models contain more information than models with primary factors only, since in addition to the primary factors they contain the broader, more general, factors. From a theoretical point of view these broader factors are extremely interesting, and they may in fact be more amenable to analysis and understanding than are the primary factors (cf. Snow, 1978, 1980).

Given these advantages of the hierarchical approach it may be asked why it has been so slow to gain acceptance. One reason for this may be that there have been several, seemingly incompatible, models to choose among. Another reason may be that estimation of hierarchical models presents technical difficulties, which problem is aggravated by the fact that there are no "canned" computer programs specially designed to produce hierarchical solutions. Recently, however, progress has been made in the development of factor analytic techniques which offer, among other things, a greatly improved capability of handling hierarchical models. This new development is discussed in the next section.

2.6 Technical issues in factor analysis

As we have seen the combatants in the battles fought over different models of the organization of human abilties have been armed with different factor analytic procedures. The most conspicous difference between these factor analytic techniques is that while some readily produce a strong general factor, others can hardly even be forced to indicate the presence of a general factor. It is, of course, quite unsatisfying that choice of technique so much influences the pattern of results, so it may be worthwhile to investigate somewhat more closely how different factor analytic techniques deal with the general factor.

To make things concrete we will use an empirical example. From the empirical material presented later on in this report two tests, Opposites and Metal Folding, have been selected, and the items in each of these tests have been divided into 3 groups, thus yielding three sub-tests for each test (Op1, Op2, Op3 and MF1, MF2, and MF3).

If these 6 sub-tests are factor-analyzed we would expect the analysis to yield two factors, representing performance on each of the tests. What is more interesting, however, is the question how different techniques manage to represent the correlation between the tests. The observed correlation between Op and MF was .32, and after correction for attenuation the true correlation can be eftimated to be about .42. This correlation is due to the involvement of some general factor in both tests.

The correlations among the 6 sub-tests are presented in Table 2. We find that the sub-tests belonging to the same test have considerably higher correlations among themselves than they have with sub-tests from the other test; the latter correlations are always higher than zero, however.

Table 2. Correlations among the sub-tests derived from Opposites and Metal Folding (N=1224).

Test	Op1	Op2	Op3	MF1	MF2	MF3
Op1	1.00					
Op2	.58	1.00				
Op3	.57	.63	1.00			
MF1	.23	.29	.24	1.00		
MF2	.24	.30	.28	.74	1.00	
MF3	.23	.28	.25	.71	.69	1.00

Let´s first analyze this matrix witn the standard factor analytic procedure: principal component analysis followed by Varimax rotation. In the principal component analysis the 3 largest eigenvalues were 3.10, 1.52 and .44, which clearly indicates two common factors. The loadings of the variables in these two factors are presented in Table 3. The first factor has the appearance of a general factor wit positive loading on all variables; the second factor is bipolar with positive loadings on the Op-tests and negative loadings on the MF-tests.

The Varimax-results are also presented in Table 3, and these results are very easily interpreted: Factor I represents the MF-test, and Factor II represents the Op-test. The approximation to simple structure is quite good, and few would consider it necessary to perform an oblique rotation.

The fact that Op and MF have something in common is not easily seen from the Varimax results. However, all loadings in the Table are positive, even though some are very small, and these small positive loadings suffice to

Table 3. Results from a principal component analysis, followed by Varimax rotation, of the 6 sub-tests derived from Opposites and Metal Folding.

	Principal Components		Varimax factors	
	I	II	I	II
Op1	.63	.56	.11	.83
Op2	.69	.52	.19	.84
Op3	.66	.55	.14	.85
MF1	.77	-.48	.90	.14
MF2	.78	-.45	.88	.17
MF3	.76	-.46	.88	.15

account for the general factor. Consider, for example, the tests Op2 and MF2 which have an observed correlation of .30. Op2 has loadings of .19 and .84 respectively in the two factors, and MF2 has loadings of .88 and .17. Reproducing the correlation between the tests from these loadings we get a predicted value of .31, which is of course as close as one may expect to get.

We thus see that in orthogonal rotations the general factor is "rotated away" by being represented as small positive loadings in all factors. However, in interpretations of factor analytic findings, loadings which are lower than .30 are rarely attended to, and often not even presented. It may thus be claimed that orthogonal rotations to simple structure may be quite deceptive.

If an oblique rotation is carried out, the general factor is represented as the correlation among factors. There are two serious problems, however, with oblique rotations in exploratory factor analysis. The first problem is that there will almost always be small positive loadings scattered in the matrix, which tend to cause the true correlation between factors to be underestimated. The second problem is that most oblique rotational methods allow the researcher himself to determine the degree of obliqueness of the solution: in the Promax method (Hendrickson & White, 1964) this is governed by the parameter k; in the indirect oblimin method (e.g. Harman, 1967, pp. 325-326) this is governed by the parameter gamma, and so on. Oblique rotational methods can, therefore, not provide "objective" empirical information on the amount of actual correlation between factors.

This can be illustrated with our example. Applying oblimin rotation with gamma taken to be .25, .50 and 1.0, the estimated correlation between the factors was .26, .18 and .00, respectively. It can be noted that these are widely differing estimates, and that they all are considerably lower than the "true" correlation which is around .42.

The fact that oblique rotations are unable to estimate correctly the correlation between factors of course causes great problems when the purpose is to conduct a hierarchical analysis by Multiple Factor analysis of several orders and it can be expected that such analysis will fail to estimate properly the relative influence of factors at different levels.

Exploratory factor analysis also is fraught with other problems, such as deciding the number of factors to rotate, and determining statistical significance of factor loadings. It is no wonder, then, that scepticism has prevailed as to the possiblity of carrying out such analyses not only at the primary level, but also at higher levels.

Recently, however, factor analytic methods have been developed in which all the problems mentioned above have been solved. Jöreskog (1969) presented a method for estimating and testing confirmatory factor models, using maximum likelihood methods. In such models the number of factors, and the pattern of loadings is specified in advance, on the basis of whatever previous knowledge is available about the variables being measured. Estimates of parameters in such models are unique, so the problem of rotation is avoided altogether. Statistical tests are also available with which the fit of the data to the model can be determined.

We have applied confirmatory factor analysis to our illustrative data. The model used is shown in Figure 3. The model is a simple two-factor model in which the factor Op is assumed to affect performance on Op1, Op2, and Op3, and the factor MF is assumed to affect the observed variables MF1, MF2 and MF3 (see also Appendix 1).

In the Figure are also presented the estimates of the parameters in the model. It can be noted that the factor loadings (i.e. the relations between the latent variables and the observed variables) are rather similar to those obtained in the Varimax-rotation, although somewhat lower. The estimate of the correlation between the factors is .40, which is close to our expected value. The test of the fit of the data to the model is clearly insignificant (chi-square=9.25, df=8, p <.32), which thus implies that we can accept the model shown in Figure 3 as a proper representation of our observational data. In contrast with the exploratory techniques, confirmatory factor analysis thus provides us with reasonable results in the analysis of our illustrative data.

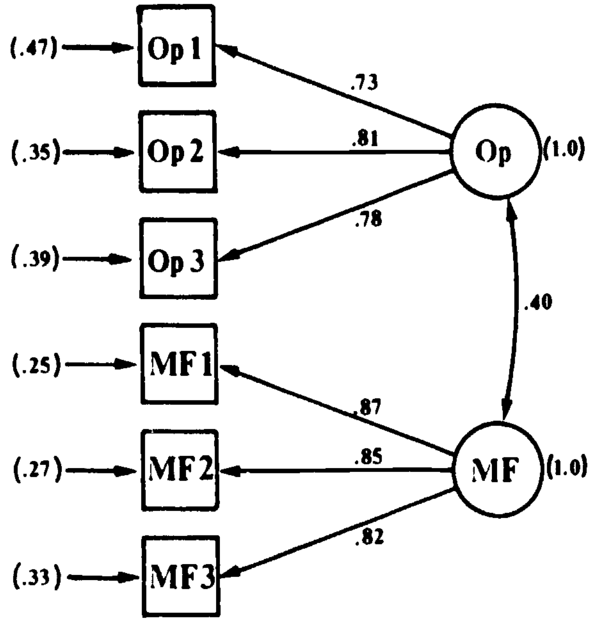


Figure 3. A confirmatory factor analytic model for the 6 sub-tests.

Jöreskog (1970) generalized the simple confirmatory factor analytic model to allow formulation of higher-order models, and in still further developments a model has been arrived at which, loosely stated, combines the factor analytic methods with path-analytic techniques (linear structural relations, LISREL; Jöreskog, 1973; Jöreskog & Sörbom, 1978). This latter model is a completely general linear model which contains all the earlier models as special cases. The LISREL model is presented in some detail in Appendix 1.

It would, thus, seem that the technical problems in estimating hierarchical models are now essentially solved. There is another prerequisite for applying confirmatory techniques, namely that there is a substantial amount of previous knowledge of the structure of the domain under study. However, within the area of cognitive abilities this is no great problems, since during the last 50 years thousands and thousands of exploratory factor analytic studies have been performed, and even with the imperfections of the exploratory techniques employed, a considerable amount of knowledge has been assembled.

2.7 Hierarchical, LISREL-based, models

In the discussion of the Vernon and Cattell-Horn models it was concluded that the difference between these models may be smaller than is evident at first sight. This suggests that it may be possible to construct another model which is a synthesis between the two models. This model would contain at the lowest level the primary factors in the Thurstone tradition; at the second level there would be the second-order factors of the Cattell-Horn model; and at the third and highest level would be found the G-factor of the Vernon model, which factor would also be identical with the Gf-factor of the Cattell-Horn model.

This model, which would contain most previously suggested models of the structure of human abilities as special cases, is as yet a hypothesis only. However, within the framework of the linear structural relations methodology this hypothesis can be confronted with observational data, and tests can be made both of the entire model, and of specific hypothesis derived from it. Most of this work remains as yet to be done, but some information on the validity of the model has been obtained from reanalyses of two published correlation matrices. One of these is a reanalysis of a study by Botzum (1951; see Gustafsson, 1980a); the other is a reanalysis of a study by Undheim (1978). Results from these reanalyses are briefly presented below.

2.7.1 The reanalysis of the Botzum study

The Botzum study was conducted to investigate more closely the reasoning and closure factors, which domains had previously been shown (e.g. Thurstone, 1938, 1940, 1944) to contain several factors. Botzum selected tests to measure different hypothesized reasoning and closure factors, along with tests for most of the Thurstone primaries. In all the test battery comprised 46 tests, which were administered to a sample of 237 male college students.

Botzum analyzed the matrix of intercorrelations between the tests with the Thurstone centroid method, and 10 factors were rotated graphically to oblique simple structure.

Among the factors there were two reasoning factors, interpreted as Induction and Deduction, and two closure factors, interpreted as Flexibility of Closure and Speed of Closure.

In the reanalysis a model containing primary factors only was first fitted for the 46 tests. In order to achieve a reasonable level of fit it was found necessary to include no less than 13 primary factors in this model. All these factors were easily interpreted within the framework of the Guilford and French et al. systems, however.

In the next steps of the reanalysis second-order factors were added, and tests of fit were made to see whether the second-order factors were able to account for the relations among the primary factors. These models did not encompass the entire test battery, however. For one thing tests were excluded because it was felt that for some primaries there were unnecessarily many tests, and for another thing the primaries Verbal Closure (Cv) and Word Fluency (Fw) had to be excluded because they failed to define the hypothesized F (or Gr) secondary.

The final model included 28 tests, 10 primary factor' and 5 secondary factors (see Figure 4). As hypothesized, Gf and Gc were clearly identified at the second-order level. Gf was loaded by I, CFC, Rs, and CMR (or analogical reasoning) and Gc was loaded by V, CMR and Rs. Gv could not be identified as a second-order factor, however. Instead no less than three second-order factors were found in the Gv-domain. One (Gvr) was loaded by Vz and Cf, and this factor was interpreted to represent the ability to retain images in the presence of distractions. Another factor (Gvt) was loaded by S and Vz, and this factor was interpreted as the ability to transform configurations into new positions. The third factor (Gvs) was loaded by P, Cs and Cf, and was interpreted as an ability rapidly to form unified percepts from unorganized stimuli.
Analyses at the third-order level were also conducted, but these were hampered by the fact that there were so few factors at the second-order level. It could be concluded, however, that the three second-order factors in the Gv-domain seemed to define a third-order Gv factor. When this factor was disregarded and a third-order G-factor was defined instead it was found that Gf had by far the highest standardized loading (.94) in this factor.

In summary, then, the reanalysis of the Botzum data allowed the following conclusions:

- At the primary level support was obtained for many factors in the Guilford and French et al. systems.

- The Cattell-Horn secondaries Gc and Gf were supported.

- No second-order Gv was found. Instead this factor seemed to splinter at the second-order level, to reappear at the third-order level.

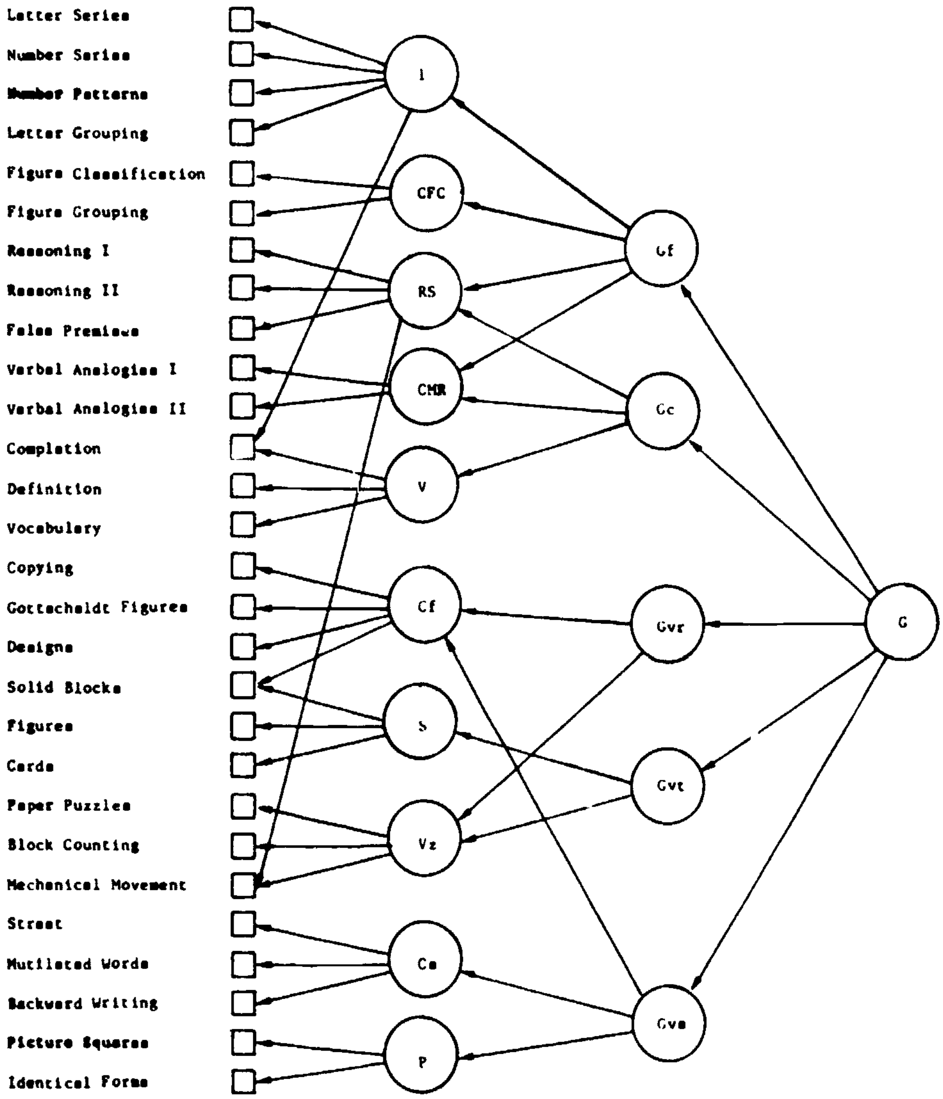


Figure 4. The model in the reanalysis of the Botzum (1951) study.

- Gf was the second-order factor most highly related to a third-order G-factor.

2.7.2 The reanalysis of the Undheim study

Undheim (1978) conducted a study with the explicit purpose to test the Cattell-Horn model. Above all interest was centered on the question whether Gc and Gf would be differentiable in a sample of 12-13 year old children.

A test battery comprising 30 tests was administered to a sample of 149 6th grade children. The tests in the battery were hypothesized to represent at least 12 primary factors. However, some primaries were represented by one test only, so an exploratory factor analysis could not aspire to identify that many factors.

Using principal factor analysis 5 factors were extracted from the matrix of intercorrelations. These factors were rotated to simple structure, using a variety of oblique and orthogonal methods of rotation. The results were consistent across the different methods, and the factors could be interpreted to represent Gf, Gc, Gv, Gs and Gr. The Undheim (1978) study thus provided very good support for the Horn and Cattell model.

Through the courtesy of Dr. Undheim the matrix of intercorrelations among the tests has been made available, and a reanalysis has been performed, using the same strategy as in the reanalysis of the Botzum study.

The final model, which is shown in Figure 5, had quite a good fit (chi-square=370.92, df=329, p <.06).
To obtain this good fit, it was necessary, however, to include no less than 9 factors at the primary level. Most of these factors correspond to the primary factors originally hypothesized by Undheim (1978). It was not possible, however, to make a distinction between the I-factor and the Figural Relations factor as was hypothesized by Undheim. Nor was it possible to make a distinction between the S, Vz and Cs factors, implying that Gv was identified as a primary factor.

As can be seen in Figure 5 there were three factors at the second level: Gr, defined primarily by Fw and Fi; Gs, defined primarily by P and N; and Gf, defined by R and the I,CFR factor. Since there was only one primary in the Gc domain, it was impossible to define this factor at the second-order level. For the analyses including a third level the V-factor was taken to represent the Gc-factor, however. Also loading the G-factor was the first-order Gv, along with the three second-order factors.

One third-order factor was found to be sufficient, and the most interesting result from the analyses at this level was the fact that Gf was found to have a perfect relation with G, the standardized loading being 1.0.

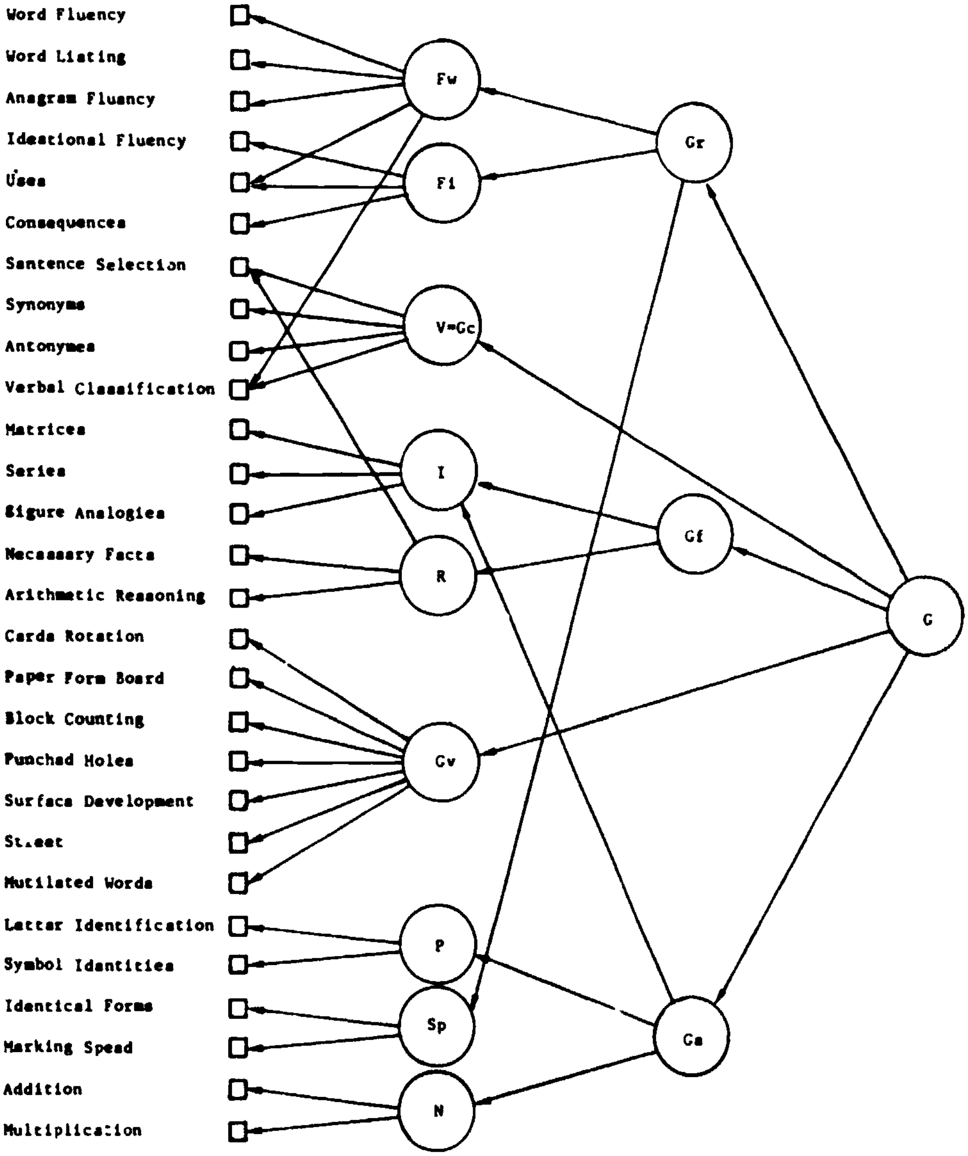


Figure 5. The model in the reanalysis of the Undheim (1978) study.

The results from the reanalysis of the Undheim study thus very clearly support the notion that Gf is identical with G, thereby bringing support to the idea of a synthesis between the Vernon and the Cattell-Horn models.

The results from the two reanalyses are quite consistent in this respect, and they also are consistent in showing that at the primary level very many factors can be identified. At one point the two reanalyses are quite inconsistent, however: While the Botzum data indicated that there in the Gv domain are no less than 5 primaries, 3 secondaries and one tertiary, the Undheim data indicated Gv to be an undifferentiated primary. It does seem clear, however, that the results in the Undheim study are at variance with findings from the majority of studies, which have shown beyond doubt the existence of several primary factors in the Gv-domain. One reason for why the studies show different results at this point may be that the Undheim sample consisted of a majority of young females, while the Botzum sample consisted of male college students. Further analyses of the Undheim data, with boys and girls kept separate, may bring additional information on this point.

2.7.3 Discussion

The studies presented above provide some support for a hierarchical model which is a synthesis of the Vernon and Cattell-Horn models. At the lowest level the very many primary factors in the Thurstone and Guilford traditions are supported. Some of these primaries are highly intercorrelated, however, and at the second-order level support is obtained for the factors hypothesized in the Cattell-Horn model, even though in one of the studies Gv was identified as a primary factor, and in the other more than one second-order factor appeared in the Gv-domain. The reanalysis of the Undheim study, finally, indicates that the second-order Gf-factor coincides with a third-order G-factor.

Each empirical study is by practical necessity restricted to a sub-set of the entire model. However, by fitting together the results in these studies with results from previous exploratory analyses it may be possible to construct a more comprehensive model. Starting from the primary factors listed in Table 1, such a model is outlined in Figure 6. In this model the Gf factor has been raised from the second-order level to the third-order level since if there is a perfect relationship between Gf and G there is no residual Gf-variance at the second-order level. Otherwise the second-order factors in the suggested model are identical with those in the Cattell-Horn model.

We will refer to the hierarchical, LISREL-based, model shown in Figure 6 as the HILI model.

The HILI model contains almost all the previously suggested models of the organization of abilities of special cases. Thus, the Spearman Two Factor model is represented in the sense that the G-factor has a prominent place

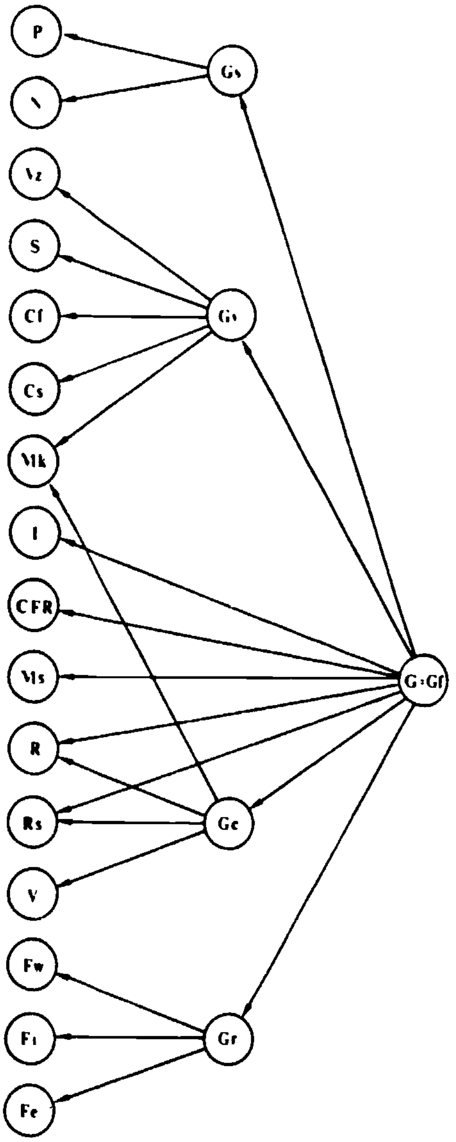


Figure 6. An hypothesized hierarchical model for some common primaries.

in the model, along with the fact that for each test a specific factor is hypothesized. The group-factors which Spearman admitted, but which he could not take into account with the technical apparatus available to him, are in the HILI-model represented at the first- and second-order levels. Parenthetically it may also be pointed out that while Spearman viewed his theory as an eclectic construction, which takes the best from each of the monarchic, oligarchic and anarchic positions, the HILI-model provides an even better synthesis of these positions, since it is better able to represent the oligarchic position than is the Two Factor theory.

The Thurstonian primaries are, of course, easily identified as a subset of the HILI-model. The G-factor identified by Thurstone and Thurstone (1941) is also represented, and just as was found by them, it appears that tests of non-verbal reasoning have the highest relation with this factor.

The Cattell-Horn model also forms a subset of the HILI-model in the sense that the first- and second-order levels in the two models coincide. The HILI-model goes one step further, however, by including also a third level. The fact that a G-factor is introduced has important implications for the interpretation of the second-order factors in the Cattell-Horn model. Discussion of this matter is post-poned, however, to the general discussion in section 7.1.

In the Vernon model the number of levels is the same as in the HILI-model. The G-factor coincides in the two models, and at the second-order level v:ed corresponds to Gc, and k:m corresponds to Gv. There is also a difference, however, since factors corresponding to Gs and Gr are by Vernon placed below v:ed. As was pointed out by Lohman (1979) these are, however, minor factors of lesser importance, and it may be a matter of taste whether they should be placed at the first- or second-order levels.

The Guilford SI-model is the only previously suggested model with which the HILI-model is clearly incompatible. This is because the abilities in the SI-model are taken to be orthogonal, which precludes reduction of primary factors to higher-order factors. It appears that Guilford (1980) now, however, admits the possibility of correlations between abilities, which paves the way for a reformulation of the SI-model in hierarchical terms.

It may, thus, be concluded that the suggested HILI-model is compatible with most previously suggested models of the structure of human abilities. However, while each of these concentrates upon variance at one or a few levels only, the present model simultaneously represents individual differences at several levels.

The key to the generality of the HILI-model lies in the identity of G and Gf, which resolves the essential difference between the Vernon model on the one hand, and the Cattell-Horn model on the other. Until further empirical evidence has been secured, this identity of G and Gf must, of course, be regarded an hypothesis only.

It should be mentioned, however, that other researchers too have formulated this hypothesis. Undheim (1981), in particular, has gone far towards a "restoration of general intelligence" on the basis of a reinterpretation of Gf in terms of G. Technically Undheim favors hierarchical Schmid-Leiman transformations of exploratory Multiple Factor analyses of several orders. It may be noted, however, that a reanalysis of the Undheim (1978) data with this technique (Undheim, 1979) failed to bring out the identity of Gf and G shown by the LISREL analysis. Instead Undheim found Gc to be the secondary factor most strongly related to G in those data. Results from some other reanalyses provided enough corroborative evidence, however, for Undheim to be able to conclude that Gf is equivalent with G.

It is difficult to explain exactly why the exploratory method of analysis applied by Undheim did not give the same results as the LISREL analysis of the Undheim (1978) data. For the reasons stated in section 2.6 it would seem, however, that the LISREL results are the most trustworthy. We will return to the Undheim model in the general discussion in section 7.1.

The empirical study presented next was designed to test the validity of the HILI-model. Further discussion of the interpretation of factors and implications of the model will, therefore, be postponed until the results from the study have been presented.

3 PROCEDURES OF THE EMPIRICAL STUDY

The major purpose with collecting the data in the present study was to obtain a reference material of tests and subjects, to be used in further empirical research on individual differences in learning. The test battery was assembled in such a way that enough primary factors would be represented to make possible identification of the second-order factors Gv, Gf and Gc. This allows a test of the HILI model, and if a reasonable level of fit is achieved this makes available for further empirical research a model for the representation of individual differences which is both parsimonious and general.

3.1 The test battery

In designing the test-battery care was taken to represent primary factors in the Gv-domain as fully as possible. One reason for this is that this domain is of special interest in relation to the planned studies of individual differences in learning; and another reason is that within this domain are found factors and tests which have been given a special place in process-oriented theories of individual differences. Thus, the cognitive style dimension field independence - field dependence (e.g. Witkin, 1950) comes very close to the Thurstonian primary Flexibility of Closure, and what Das, Kirby and Jarman (1979) refer to as simultaneous processing is measured by tests belonging to the Gv-domain.

The test-battery consists of two parts: one with 13 tests of ability, and one with 3 standardized achievement tests. The following tests of ability were included in the battery:

1. Number Series II. In the items in this test a series of 5 or 6 numbers are given, and the task is to add two more numbers to the series. Tests of this type have been shown to load the primary factor Induction (I), which in turn loads Gf.

2. Letter Grouping II. The items in this test consist of groups of letters, and the task is to decide which group of letters does not belong with the others. This kind of test too has been shown to load the I-factor.

3. The Raven Progressive Matrices. The items in the Raven test present a matrix of figures in which the figures change from left to right according to one principle, and from top to bottom according to another principle. One figure is missing, however, and the task is to identify this figure. It is not entirely clear to which primary factor the Raven test should be classified. French et al. (1963) would assign this test to the I-factor. Their I-factor is quite broad, however, and it has more the character of a second-order factor than a primary factor. Horn and Cattell (1966) use the Guilford notation (Cognition of Figural Relations, CFR) to classify this test and we will use the same notation here. CFR is hypothesized to load Gf.

4. Auditory Number Span. This is a conventional digit-span test, with digits in series of varying length being read for immediate reproduction. The test may be hypothesized to load the Memory Span (Ms) primary, which primary by Horn and Cattell (1966) is hypothesized to be weakly related to Gf.

5. Auditory Letter Span. This test is identical with the preceding test, except that letters are used instead of digits.

6. Metal Folding. In this test the task is to find the three-dimensional object which corresponds to a two-dimensional drawing. Metal Folding may be hypothesized to load the Visualization (Vz) primary, which in turn belongs with Gv.

7. Group Embedded Figures. This test consists of items in which the task is to find a simple figure within a more complex figure. The test has been shown to represent the Flexibility of Closure (Cf) factor, which in turn loads Gv.

8. Hidden Patterns. Each item consists of a geometrical pattern, in some of which a simpler configuration is embedded, and the task is to identify those patterns which contain the simple configuration. The test is similar to the Group Embedded Figures test and may be hypothesized to load the Cf-factor.

9. Copying. In this test each item consists of a given geometrical figure, which is to be copied onto a square matrix of dots. French et al. (1963) classify this test with the Cf-factor.

10. Card Rotations. Each item in this test gives a drawing of a card cut into an irregular shape, and the task is to decide whether other drawings of the card are merely rotated, or turned over onto the other side. This test is highly similar to the Thurstone tests Cards,

Figures and Flags, which have been shown to define the Spatial Orientation (S) primary. This primary loads Gv.

11. Disguised Words. In this test words are presented with parts of each letter missing, and the task is to identify the word. The test is highly similar to the Thurstone (1944) test Multilated Words, which by him was found to load the Speed of Closure (Cs) factor. French et al. (1963) mention, however, that tests like these may also have a loading on a Verbal Closure factor. According to Horn and Cattell (1966) Cs loads Gv.

12. Disguised Pictures. In the items of this test drawings are presented which are composed of black blotches representing parts of the object being portrayed, and the task is to identify the object. Tests similar to this one have been found to load the Cs-factor.

13. Opposites. In each of the items in this test the task is to select the word which is the antonym of a given word. The test may be hypothesized to load the Verbal Comprehension (V) primary, which in turn loads Gc.

These 13 tests comprise the tests of cognitive ability in the test battery and they were administered on one occasion. The tests are described and analyzed in detail in Chapter 4. In addition scores are available on Standardized Achievement tests in Swedish, mathematics and English. These tests are administered by the class teachers to most pupils in the 6th grade. The Standardized Achievement tests may all be hypothesized to represent Gc. These tests are described in greater detail in section 4.3.2.

3.2 Subjects

The study comprised 50 classes (or rather 51, since one class was divided into half-classes) in the 6th grade (i.e. the pupils are in their 12th year), in two different communities (Mölndal and Kungsbacka). In all 1254 pupils attended these classes, but for different reasons the battery of cognitive tests could not be administered to 30 pupils. The final sample thus comprised 1224 subjects (602 boys and 622 girls), with an attrition of only 2.4 per cent. For the Standardized Achievement tests attrition was greater, however, which is because it is up to the class-teacher to decide whether all, some or none of these shall be administered. Of the sample of 1224 subjects, 981 (or 80.1 per cent) had results on all the Standardized Achivement tests while 113 (9.2 per cent) had not taken any of these. For the remaining 130 subjects scores were available on one or two of the Standardized Achievement tests.

In Appendix 3 the correlation matrix for the test battery is given for the subset of subjects with complete data. Descriptive statistics for the 981 subjects with complete data, and for the 1224 subjects with results on the cognitive tests are also presented. The subjects who lack one or more of the Standardized Achievement tests seem to have a somewhat lower level of performance on the cognitive tests than the subjects with complete data.

3.3 Procedures used in the testing

All 13 tests were given at one single occasion for each class. In all it took about five 40-minute lessons to administer the tests to one class. The class had recesses and a lunchbreak in a regular manner. In most classes three lessons were used before lunch and two after the lunchbreak. One female administrator gave the tests to 25 classes in the community of Mölndal, and a male administrator gave the tests to 25 classes in the community of Kungsbacka. The class teacher was generally not present in the classroom during the testing.

The tests were in all classes administered in the same order and in a standardized manner (see Table 4). Before the administration all pupils were given a general introduction, and the purpose of the study was explained to them. They were told that they were participating in a large-scale experiment that would be followed by other experiments of less extent during the spring semester. The pupils also were told that their results on the tests would not be told to their parents or their teacher. However, the class would be informed of the results of the class as a whole in comparison with the other classes that participated. Furthermore, the pupils were told that the tests were of different levels of difficulty, and that no one would be able to solve all tasks. It was also pointed out that it was important that they listened to the instructions given for each test, and that they should ask questions, if they did not understand what was required of them.

Most pupils seemed to think that participation in the study was rather enjoyable and a nice break in regular school-work. However, all pupils were fairly tired at the end of the testing.

It seemed that some pupils were very disappointed not to be able to finish all the test-items, and those pupils were of course disturbed by the time-limit on each test. Nevertheless, the administrators´ overall

Table 4. Order of administration of the tests in the battery.

Test	Time for instruction	Time limit
Opposites	3	10
Raven Progressive Matrices	5	15
Auditory Number Span	3	12
Recess		
Hidden Patterns	5	3 + 3
Letter Grouping II	5	10
Card Rotations	5	4 + 4
Disguised Words	3	2.5 + 2.5
Recess		
Copying	5	3 + 3
Number Series II	5	10
Metal Folding	5	10
Recess		
Auditory Letter Span	5	12
Group Embedded Figures	5	2 + 5
Disguised Pictures	5	3 + 3

impression was that the pupils' attitude toward participation in the study was positive.

The administrators took notes of the climate in the class-room throughout the schoolday and the behavior of the class during the administration. A schedule of classification was used where information on motivation, concentration and relation to the administrator was gathered. Special incidents during each test and special circumstances prevailing in the class were noted. Besides, the teacher was interviewed about the history of the class in grades 4, 5 and 6 and information on motivation, concentration, willingness of cooperation, working-pace and school achievement was gathered.

3.4 Representation of the test information

In order to be able to carry out closer analyses of the psychometric properties of the tests in the battery, the examinnee responses were represented in machine-readable form in as great detail as was practically feasible.

For the multiple-choice tests (i.e. Letter Grouping II, Raven, Metal Folding and Opposites) the alternative chosen for each item was key-punched. The Group Embedded Figures test and Copying, which require the examinees to contribute responses in the form of drawings, were scored with a distinction being upheld between attempted and correct, and attempted but incorrect items.

The heavily speeded Hidden Patterns and Card Rotations tests, which contain very many items, were scored in such a way that the number of correct responses and the number of incorrect responses were determined.

In 5 of the tests (Number Series II, Disguised Words, Disguised Pictures, Auditory Number Span, and Auditory Letter Span) the answers are open-ended. These tests would normally require judgement of the correctness of each contributed response, which is a difficult and arduous task. However, for the present study a set of computer programs (the PRINS-system, developed by Jan-Gunnar Tingsell at the Department of Educational Research, University of Göteborg) was used to simplify this task. With the PRINS-system each examinee´s response to each of the items is recorded exactly as it is given, which is done interactively at a computer terminal. A frequency list of all contributed responses is then obtained, on the basis of which it may be decided which responses should be considered correct. In the final step the selection of correct responses is presented to the computer, and the scoring of the tests is done automatically. With this system it is, of course, also easy to study the properties of different scoring systems.

4 RESULTS FROM ANALYSES OF THE TESTS

In this chapter analyses of each test are presented. Since one of the purposes of the present report is to serve as a reference report on the reference material, quite detailed descriptions and analyses are presented.

The analyses at the test level are focused upon the internal consistency of the tests. In addition to measures from classical test-theory, such as the biserial correlation and the coefficient of reliability, Rasch model analyses have been relied upon. The Rasch model may not yet be well-known, sc a brief introduction to this item-response model is presented in Appendix 2.

4.1 Analyses of the Gf-tests

In the test battery there are three tests which are hypothesized to represent the second-order Gf-factor: Number Series II, Letter Grouping II, and the Raven Progressive Matrices test. In addition the Auditory Number Span and Auditory Letter Span tests may be hypothesized to belong with the Gf-complex.

4.1.1 Number Series II

Number Series II (NS) was newly constructed for this study, but it was, of course, modelled upon existing series tests and in particular upon the test Number Series constructed by Svensson (1964, 1971). To a certain extent selection of items for NS was based upon an analysis of the Svensson Number Series test with the Rasch model (Gustafsson, 1977). From that analysis it was concluded, among other things, that the original test, which had 40 items and a time-limit of 18 minutes, to some extent may have been speeded. The analysis also showed that for some items the fit to the Rasch model was very poor, which was interpreted as being due to the fact that these items put an extra demand on arithmetic skills. It was also observed that the reliability of the test was high even when several items were excluded.

In constructing the NS-test it was, therefore, decided to limit the test to 20 items, and the testing time to 10 minutes. Attempts were made to include items which demand only limited arithmetic skills. For certain types of series, however, it was impossible to avoid large numbers, and in some cases the number of terms in the presented part of the series was, therefore, limited to 5, while in most other items 6 terms were used. The items in the test, along with the algorithm upon which they are based are presented in Table 5. Some of the items in the test were taken from the Number Series test, but most were newly composed.

Each subjects´ response to each item was recorded as it was given, using the PRINS-system (see section 3.4). In the first step of the analysis the test was scored in such a way that only the logically correct responses as listed in Table 5 were judged correct. The proportion of correct answers from this strict (S) scoring are presented in Table 6, along with the biserial correlations. The range of variation of the proportion of correct answers is great: from a high of .95 down to .02. The biserial correlations are all fairly high and even in size, although there is a clear tendency towards a correlation between the difficulty of the item and the biserial correlation.

With this S-scoring the overall mean of the test is 7.8 with an sd of 3.75. The reliability (KR-20) is .84, which with only 20 items must be considered a high value.

The item responses have also been analyzed with the Rasch model, with the primary purpose of investigating the fit of the data to the model. The fit was poor, however: For the ML-ICCSL test (see Appendix 2) a very highly significant chi-square of 1381.6 with 342 df, was obtained. A closer analysis of the fit of each item indicated that the more difficult items tended to have too high a discrimination, while for several other items (items 7 and 9 in particular) too low a level of performance was observed for persons scoring high on the test as a whole.

Items 7 and 9 both involve quite difficult arithmetic computations, so it may be suspected that the reason why they show this pattern of misfit is that they put an extra demand on arithmetical ability. Also for other items showing the same pattern of deviation, but less markedly, it might be suspected that arithmetical ability influences the results.

Attempts, therefore, were made to score the responses to each item in such a way as to minimize this source of influence. For each item it was for each incorrect response decided whether it represented a logical error or a computational error. For some items these decisions were easier than for other items; in some cases the decision even was impossible to make on the basis of the available information. However, the general rule followed was to be liberal in the scoring. Table 6 presents descriptive statistics at the item level for the liberal (L) scoring. The first column shows the number of "incorrect" response categories judged correct in the L-scoring.

Table 5. The items in the Number Series II test, along with results from the strict scoring.

Item	Algorithm	Item						Correct answer		Prop. corr.	Biserial correl.
1	a_n+2	3	5	7	9	11	13	15	17	0.93	0.55
2	a_n-2	18	16	14	12	10	8	6	4	0.95	0.64
3	a_n+3	2	5	8	11	14	17	20	23	0.91	0.54
4	a_n-3	24	21	18	15	12	9	6	3	0.89	0.60
5	$a_n \cdot 2$	1	2	4	8	16		32	64	0.59	0.77
6	a_n-2n	60	58	54	48	40	30	18	4	0.19	0.77
7	a_n+2n	2	4	8	14	22	32	44	58	0.18	0.68
8	$a_n\|2$	320	160	80	40	20		10	5	0.46	0.76
9	$a_n \cdot 2$	3	6	12	24	48		96	192	0.43	0.73
10	a_n-n	31	30	28	25	21	16	10	3	0.36	0.79
11	a_n+5n	5	10	20	35	55	80	110	145	0.18	0.74
12	$a_n\|2$	448	224	112	56	28		14	7	0.41	0.88
13	$a_{n-1}+3$	7	8	10	11	13	14	16	17	0.51	0.78
14	$a_{n-1}-5$	19	17	14	12	9	7	4	2	0.35	0.83
15	$\begin{cases} a_{n-1}+2 & n=1,3... \\ a_{n-1}+4 & n=2,4... \end{cases}$	2	6	4	10	6	14	8	18	0.06	0.83
16	$a_{n-1}+2$	1	4	3	6	5	8	7	10	0.14	0.85
17	$\begin{cases} a_{n-1}-1 & n=1,3... \\ a_{n-1}-2 & n=2,4... \end{cases}$	8	12	7	10	6	8	5	6	0.07	0.90
18	$a_{n-1}-1$	12	13	11	12	10	11	9	10	0.15	0.80
19	$\begin{cases} a_{n-1}+1 & n=1,3... \\ a_{n-1}+3 & n=2,4... \end{cases}$	3	1	4	4	5	7	6	10	0.02	0.94
20	$\begin{cases} a_{n-1}-2 & n=1,3... \\ a_{n-1}-3 & n=2,4... \end{cases}$	15	16	13	13	11	10	9	7	0.02	0.85

This number ranges between 0 and 81, and it is particularly large for items 5, 6, 7, 9 and 11. These items all involve large numbers. The items early in the test were not much affected by the L-scc~ing for the simple reason that these items do not involve much of arithmetic skills. Nor were the items late in the test much affected by the L-scoring. This is due both to the fact that most of these items do not involve serious arithmetic difficulties and to the fact that in these items it was virtually impossible to make a distinction between computational and logical errors.

Table 6. Descriptive statistics for the items in Number Series II.

	Number of in-correct answers	Proportion correct			Biserial correlation		
Item	scored correct	Liberal	Strict	Diff. L-S	Liberal	Strict	Diff. L-S
1	3	0.97	0.93	0.04	0.62	0.55	0.07
2	6	0.96	0.95	0.01	0.71	0.64	0.07
3	6	0.97	0.91	0.06	0.72	0.54	0.18
4	5	0.94	0.89	0.05	0.64	0.60	0.04
5	17	0.64	0.59	0.05	0.81	0.77	0.04
6	22	0.31	0.19	0.12	0.78	0.77	0.01
7	36	0.32	0.18	0.14	0.76	0.68	0.08
8	1	0.46	0.46	0.00	0.77	0.76	0.01
9	81	0.59	0.44	0.15	0.85	0.73	0.12
10	9	0.44	0.36	0.08	0.84	0.79	0.05
11	29	0.35	0.18	0.17	0.82	0.74	0.08
12	7	0.43	0.41	0.02	0.89	0.88	0.01
13	2	0.53	0.51	0.02	0.80	0.78	0.02
14	1	0.36	0.35	0.01	0.83	0.83	0.00
15	8	0.09	0.06	0.03	0.80	0.83	-0.03
16	0	0.14	0.14	0.00	0.84	0.85	-0.01
17	0	0.07	0.07	0.00	0.86	0.90	-0.04
18	0	0.15	0.15	0.00	0.78	0.80	-0.02
19	2	0.03	0.03	0.00	0.91	0.94	-0.03
20	1	0.03	0.03	0.00	0.80	0.85	-0.05

Performance on item 11 was most affected by the L-scoring, the proportion of answers being .17 units higher than in the S-scoring. For the other items affected by the liberal scoring the increase in proportion correct was .12 - .15. The mean performance on the test increased from 7.82 to 8.75, implying that on average one extra point was awarded each subject in the L-scoring.

The L-scoring affected the biserial correlations in such a way that they became more homogenous, and the correlation between item difficulty and biserial correlation decreased. As might be expected, a considerably better fit to the Rasch model was obtained for the liberally scored test (chi-square=752.39, df=342). Since the test staticstic is more than twice the degrees of freedom it must be concluded that the fit is not perfect even under the L-scoring. However, considering the rather large sample of subjects the level of fit is not too poor.

It is quite impossible to present in detail for each item the responses judged correct in the L-scoring, so that will be done for one item only, no. 7.

For this item the simple principle was followed that all responses were judged correct in which 44 was given as the continuation of the series, irrespective of the next term given. In all, 116 subjects gave such responses. In addition the following responses were judged correct (the number after the answer indicates how many subjects gave that response): 34 44 (1); 34 48 (3); 35 58 (1); 40 54 (2); 42 54 (13); 42 56 (3); 43 55 (17); 46 58 (1); 46 60 (2); 46 62 (3); 46 64 (1); 52 90 (1); 52 92 (1); 54 68 (2); 54 86 (6); and 64 78 (3). While for some subjects some of these responses undoubtedly reflect logical mistakes, it is also conceivable that they are due to fairly simple errors of computation.

In all 161 different responses were contributed to this item, and most of these were judged incorrect even in the L-scoring. The most common incorrect answer was 42 52 (164). This answer reflects application of the simple algorithm "last number plus 10" to the last two terms in the given series, and represents an undue simplification of the problem. Other common incorrect answers were 34 26 (27) and 34 38 (20) which both reflect application of too rudimentary an algorithm to capture more than the variation of the last two termc in the given series. Another frequent incorrect answer was 64 128 (89), which represents the algorithm "last number times 2" applied to the last term in the given series.

Table 7 presents for each of the items the two most common incorrect responses, along with the frequencies with which they occured. A close analysis of incorrect responses has not yet been performed, but some preliminary observations can be made on the basis of the information presented in the table.

Some of the incorrect responses listed in Table 7 are likely to be the effects of computational errors, such as the answer 16 19 to item 1 (i.e. application of the algorithm "last number plus 3" instead of the correct algorithm "last number minus 3"). However, most of the incorrect responses represent errors of logic, such that the subject has failed to take into

Table 7. The most common incorrect responses to the items in the Number Series test. Frequencies in parentheses.

Item	Correct answer	Incorrect answers
1	15 17 (1142)	16 19 (29); 14 15 (15)
2	6 4 (1165)	20 22 (15); 7 6 (6)
3	20 23 (1119)	19 21 (40); 20 22 (9)
4	6 3 (1094)	7 5 (29); 27 30 (23)
5	32 64 (719)	24 32 (48); 20 24 (45)
6	18 4 (230)	20 10 (198); 28 24 (74)
7	44 58 (215)	42 54 (164); 64 128 (89)
8	10 5 (561)	10 0 (227); 640 1280 (16)
9	96 192 (532)	56 112 (27); 54 60 (23)
10	10 3 (444)	10 4 (57); 11 6 (49)
11	110 145 (220)	110 151 (92); 85 90 (55)
12	14 7 (500)	14 0 (10); 16 8 (10)
13	16 17 (628)	15 16 (88); 9 12 (66)
14	4 2 (428)	5 3 (139); 4 1 (56)
15	8 18 (72)	8 12 (39); 8 16 (24)
16	7 10 (167)	2 7 (119); 7 9 (57)
17	5 6 (82)	9 11 (63); 12 7 (18)
18	9 10 (185)	12 13 (48); 14 15 (40)
19	6 10 (30)	2 6 (75); 8 8 (24)
20	9 7 (30)	12 14 (67); 9 9 (50)

account the full complexity of the presented information. There are several interesting types of such errors. One type, for example, is represented by the answer 20 22 to item 2, the response 27 30 to item 4, and the answer 640 1280 to item 8. The series in these items are all descending, but these responses are ascending. They are, furthermore, correct when appended to the beginning of the presented series. Thus, in this type of incorrect response the subject has failed to operate with decreasing numbers, but has still, by redefining the problem as it were, managed to produce a response.

Another type of incorrect answer is represented, for example, by the response 24 32 to item 5, the response 20 10 to item 6 and the response 42 52 to item 7. In all these incorrect responses too simple an algorithm has been applied, which fits the last few terms in the presented series, but which is not able to generate the entire series. This type of incorrect answer was by far the most frequent one, which indicates that the major source of difficulty in the items is to infer the rules underlying the presented series.

4.1.2 Letter Grouping II

The test Letter Grouping II (LG) was constructed on the basis of two predecessors: Letter Grouping in the DBA-battery (Härnqvist, 1968), and the Letter Sets test in the ETS-battery (French et al., 1963). One of the reasons for developing a new test instead of using the DBA-test is that the items in this test include 4 response alternatives only. With more response alternatives a higher reliability may be secured with the same number of items. From the ETS-test the ideas were to some new items borrowed.

The LG-test consists of 20 multiple-choice items, each with 5 response alternatives. Table 8 presents the algorithms from which the groups of letters were constructed. As can be seen the items range considerably in complexity. No attempts were made, however, to vary systematically the level of complexity of the items, since this was judged an impossible task with such a short test, and with so many factors affecting item difficulty.

Descriptive item statistics are also presented in Table 8. The proportion correct ranges from a high of .98 to a low of .24, with a fairly even distribution between these extremes. The biserial correlations are rather high and even in size, with somewhat higher values being observed for items 13, 14 and 15.

It can only be conjectured as to why these items are especially good measures of the latent variable underlying performance on the items. It would seem, however, that in these items the structure of the outlying group of letters comes quite close to the structure of the other groups.

The test of fit to the Rasch model proved significant (chi-square=619.9, df=342 with the ML-ICCSL test). However, the ratio of the chi-square sum to the degrees of freedom is only about 1.8, which with the quite large sample of persons being analyzed here indicates a reasonably good fit. The test can, thus, be considered quite homogenous.

The mean of raw scores on the test is 11.7, with an sd of 4.1. The reliability is .82, which is a large coefficient for a multiple choice test with just 20 items.

Table 8. The algorithms of the items in Letter Grouping II, along with descriptive item statistics.

Item	Structure of[1] correct answer	Structure of[1] incorrect answer	Proportion correct	Biserial correlation
1	AAAA	AABA	.98	.31
2	1234	1243	.94	.48
3	D11X	D11Z	.83	.53
4	1233	1222	.86	.63
5	1213	1215	.68	.64
6	$1_1 1_2 2_2 2_1$	$1_1 UZ2_1$	.53	.60
7	1221	2112	.73	.59
8	1234	1345	.77	.63
9	1212	2121	.79	.66
10	1231	1232	.65	.67
11	1345	1456	.45	.64
12	1357	14710	.41	.67
13	2112	1221	.53	.74
14	3412	3421	.45	.74
15	1232	H2N2	.52	.78
16	$1_1 3_1 1_2 3_2$	$1_1 1_2 3_1 3_2$	.34	.62
17	2341	3561	.42	.67
18	$1_1 2_1 1_2 2_2$	$1_1 3_1 5_1 6_1$	.40	.69
19	1345	1456	.24	.59
20	$2_1 1_2 1_1 2_2$	$4_1 1_2 1_1 2_2$	.24	.50

1) Note. In the coding of the structure of the items the following rules were used: (1) A letter indicates a particular letter as presented in the item; (2) the number 1 indicates any letter in the alphabet, and higher numbers the ordinal position of other letters counted from this letter; and (3) when the numbers are indexed the second rule applies for each indexed series separately.

4.1.3 The Raven Progressive Matrices

The Raven Progressive Matrices test (RA) was constructed by Raven (1938), on the "... assumption that if Spearman's principles of noegenesis were correct, it should provide a test suitable for comparing people with respect to their immediate capacities for observation and clear thinking" (Raven, 1960). The test systematically exploits the noegenetic principles of eduction of relations and eduction of correlates on meaningless figures.

Each item presents a logical analogy problem, in which a matrix of figures changes from left to right according to one principle and from top to bottom according to another principle. The figure in the right-most/bottom corner is missing, however, and the task is to select the one among 6 or 8 presented designs which completes the pattern.

The test consists of five parts (A-E), containing 12 items each. In the manual it is recommended that the test is administered without a time limit, and it is mentioned that 40-60 minutes may be typical amounts of time spent on the test. For the group-testing situation in the present study it would be practically impossible, however, to let everyone spend unlimited time with the test.

The decision was made, therefore, to administer only parts B-E, with a time limit of 15 minutes, and then decide with statistical analysis which of the items late in the test provide useful information. Items B1 to B3 were used as common practice items.

The test is organized in such a way that within each set of 12 items the items increase in difficulty according to a certain principle.

Most items in set B involve change from one geometric form to another, along with a change of the "content" (black/white, striped/not striped, dotted/not dotted and so on) of the geometric form. These items are all rather easy. In the present sample the most difficult item in this set (B12) was answered correctly by 44 per cent of the sample. In most of the items in set C factors such as size, location, and number of attributes are varied simultaneously along the two dimensions. The items in this set show a wide range of difficulty: in the present sample the observed range was from 94 per cent correct to 6 per cent correct answers.

In set D the combination of geometrical forms is systematically exploited. The items in this set too vary much in difficulty: between 95 and 8 per cent correct answers were observed in the sample.

The E-set items, finally, operate with geometrical forms in such a way that these combine according to algebraic rules (i.e. addition and subtraction).

Since many persons in our sample did not have time enough to attempt the items in this set it is hard to tell about the relative difficulty of the items.

Table 9 presents descriptive item statistics. From the figures presented it can be seen that the range of difficulty within each subset of items is great. The range of variation of the biserial correlations also is great, implying that the test may be fairly heterogenous. For this reason no closer analysis has been made of the test items using the Rascn model.

Table 9. Descriptive statistics for the items in the Raven test.

Item	Proportion correct	Biserial correlation	Item	Proportion correct	Biserial correlation
B4	.92	.63	D1	.95	.86
B5	.89	.65	D2	.88	.88
B6	.77	.53	D3	.85	.81
B7	.66	.39	D4	.76	.76
B8	.64	.70	D5	.85	.87
B9	.71	.75	D6	.75	.77
B10	.74	.77	D7	.64	.61
B11	.64	.71	D8	.68	.65
B12	.44	.54	D9	.60	.60
C1	.94	.61	D10	.52	.63
C2	.91	.59	D11	.19	.34
C3	.80	.58	D12	.08	.33
C4	.74	.63	E1	.45	.56
C5	.79	.67	E2	.36	.49
C6	.66	.49	E3	.31	.41
C7	.77	.74	E4	.22	.53
C8	.57	.56	E5	.20	.53
C9	.19	.18	E6	.16	.48
C10	.39	.49	E7	.16	.26
C11	.22	.39	E8	.08	.29
C12	.06	.18	E9	.05	.18
			E10	.03	.14
			E11	.03	-.12
			E12	.03	-.32

The first 6 items in set D have the highest biserial correlations, while at the same time they display high proportions of correct answers. The reason why these items are good measures of the ability underlying performance on the test may be that in them the two factors vary in such a way that only a careful logical analysis brings out the correct answer; it cannot be found by simple "pattern-perception". Of course many other items in the test have this quality as well. However, it may be suspected that the test was too speeded for the E-items to function properly in this sample. It can also be observed that the more difficult items within each sub-set have lower biserial correlations, which is most likely an effect of guessing. It can be shown (see Gustafsson, 1980b) that random guessing causes overestimation of the discrimination of easy items, and underestimation of the discrimination of difficult items.

Item C9 provides an example of an item having a very low discrimination, along with a proportion of correct answers which is not too low. The reason why this item has so low a discrimination may be that it suffices to consider the left-right variation; the top-bottom variation consists in a change from circle to square to triangle, which hardly requires a close analysis to be identified.

The items in the E-set have considerably lower proportions of correct answers than have the items in the other sets. The biserial correlations also tend to be lower, and they decrease systematically so that for the two most difficult items the biserials even are negative. These poor properties of the E-set are more likely to have been caused by an insufficient time allowance, however, than by any property immanent in the items themselves.

The number of contributed answers, rights as well as wrongs, varies between 8 and 48, with about 40 per cent of the examinees answering all items. For 45 per cent of the subjects in the sample item E12 was the last answered one and for an additional 40 per cent the last answered item was one in the E set. It can thus be concluded that an absolute majority of the subjects in the sample had sufficient time to answer the B, C and D sets, but that many did not have sufficient time for the E set.

It thus seems that with the time limit of 15 minutes insufficient time was available for the items in the E set, thereby destroying the measurement properties of these items. This is also indicated by the fact that the reliability of the test with the E items included is .868; excluding the E items the reliablity increases to .870.

It may be asked, of course, whether the items in the different sets measure different abilities. One way to investigate this is to use the ML-PCC test (see Appendix 2) in pairwise comparisons of the sets. The results from such comparisons are presented in Table 10, along with the observed correlations. All test statistics are highly significant, indicating that each of the item sets measures a distinct latent variable.

The correlations among the item sets are quite low, and especially so those with the E set. These correlations are attenuated by errors of measurement, however, which with such short scales have considerable impact. The true correlations are, therefore, underestimated by a large amount.

Table 10. Results from pair-wise comparisons with the ML-PCC test of the item sets in the Raven test.

	C		D		E	
	Chi-square	Corr.	Chi-square	Corr.	Chi-square	Corr.
B	429.8	.52	626.3	.54	805.7	.19
C			467.6	.55	467.3	.29
D					624.9	.31

Note. The chi-square statistics all have 143 df.

To estimate the true correlation among the scales the items within each set were divided into two subsets according to whether the ordinal position was odd or even. In this way 8 subsets with 6 items each were obtained, which were than subjected to a sequence of LISREL-analyses.

In the first step it was tested whether one factor suffices to account for the intercorrelations among the 8 subsets. This test investigates, of course, the same hypothesis as does the ML-PCC test, and it gave the same answer: a very highly significant value was observed for the test statistic (chi-square=931.6, df=20). In the next step a simple oblique 4-factor model was specified, in which the half-tests derived from the same set were assumed to be caused by the same latent variable. This model fitted well (chi-square=32.3, df=16), and Table 11 presents estimates of the correlations among the latent variables in this model. Sets B, C and D have the highest correlations among themselves; the correlations are so far from unity, however, that it must be concluded that in each of the sets there is a substantial amount of specific variance. The correlations with the E set are very low indeed.

Table 11. LISREL estimates of the true correlations among the sets of items in Raven.

	B	C	D	E
B	1.00			
C	.69	1.00		
D	.65	.73	1.00	
E	.25	.40	.40	1.00

It may thus be concluded from the analyses of the Raven test that sets B, C and D in the present study provide useful information about individual differences, while the E-set does not. In the final scoring of the test the results on the E-items were, therefore, not included. It can also be concluded, however, that the sets measure different abilities. Even though these abilities are so highly correlated that it may be meaningful to talk about a general Raven factor, the nature of what is measured by each of the sets will be investigated more closely. This will be done in LISREL analyses, which are presented in section 5.1.

4.1.4 Auditory Number Span

The test battery includes two memory span tests, which most likely represent the Ms-factor. This primary, in turn, seems to belong with the Gf-domain. The major purpose for including these tests in the battery was, however, that they have a prominent place in the Das, Kirby and Jarman (1979) model of simultaneous and successive processing, in which these and similar tests are used to represent successive processing.

The Auditory Number Span test (ANS) is a conventional digit span test in which series of digits are read at a speed of one digit per second. The task is to write down the series when it is completed. The test consists of 19 items, with the number of digits varying between 4 and 10. The test is almost identical with the Auditory Number Span test in the ETS battery, but the items with 11 or 12 digits were not used, since it was felt that they would be too difficult for the present sample.

The subjects' answers were recorded exactly as they were given within the PRINS-system to allow a comparison of different scoring models. In the basic scoring model (BSM) a response was judged correct if it was an exact reproduction of the presented series. In the sequential scoring model

(SSM) the subjects were allowed to miss one digit (in the series with 5 to 7 digits) or two digits (in the series with 8 or more digits) as longs as the correct order was preserved for the remaining digits. In the liberal scoring model (LSM) not only one or two missing digits were allowed but up to two reversals of order as well. These different scoring models will be studied more closely in relation to analyses of the simultaneous/successive processing model, which will be presented in another context.

Table 12. Means and standard deviations for the Auditory Number Span test.

Scoring model	$\bar{x}$	sd
BSM	4.41	2.68
SSM	5.55	2.98
LSM	6.70	3.10

Table 12 presents descriptive statistics for the ANS test under the 3 scoring models. As may be expected the mean is highest under the LSM, and lowest under the BSM, the difference between each of the scoring models being somewhat more than one point. The scoring models all yield a reliablity around .71.

4.1.5 Auditory Letter Span

The Auditory Letter Span (ALS) test is identical with the ANS test, except that letters are used instead of digits. The test is highly similar to the Auditory Letter Span test in the ETS battery but the series of letters had to be newly constructed. This is because the Swedish pronounciation caused problems differentiating the aurally presented letter in many of the series. A new set of 19 items was constructed, with the length of the items varying between 3 and 9 letters. Within a series letters which sound alike were avoided; nor was any letter repeated within a series.

The subjects´ answers were treated in the same way as with the ANS test, but only two scoring models were used; a basic scoring model (BSM) and a liberal scoring model (LSM). The LSM was essentially the same as the LSM for the ANS test, but answers were judged correct also if they contained letters with a pronounciation similar to the ones in the presented series.

Table 13 presents descriptive statistics for the two different scoring models. The mean in the LSM is somewhat more than one unit higher than the mean under the BSM. The reliability under the strict scoring model is .62. In the liberal scoring model it increases to .66.

Table 13. Means and standard deviations for the Auditory Letter Span test.

Scoring model	$\bar{x}$	sd
BSM	4.52	2.17
LSM	5.68	2.45

4.1.6 Discussion

The tests NS, LG and RA have in common that they present quite complex items, in which the major source of difficulty is to find relations and apply relations. The tests also have in common that they do not to a large extent rely upon a previously acquired store of knowledge. These facts make it reasonable to hypothesize that they belong in the Gf-domain. It should be pointed out, however, that the three tests represent three different content areas, and secondary loadings in other factors cannot be ruled out.

The analyses at the item level of these three tests indicate that NS and LG in particular are homogenous tests. RA appears to be more heterogenous, however, the sub-sets of items reflecting distinct aptitudes. Further analyses of the RA-test are presented in section 5.4.

Just as the other tests in this group, the ANS and ALS tests have quite simple content, and in the present sample individual differences in degree of familiarity with numbers and letters cannot be expected to account for a large proportion of the variance in test scores. Horn (1968) instead saw performance on tests such as these as reflecting fairly directly the efficiency of the nervous system, and presented evidence that they do load Gf. The loadings tend to be modest in size, however, which is most likely due to the fact that the complexity of the items is not very great.

4.2 Analyses of the Gv-tests

In the test battery there are seven tests which are hypothesized to represent the second-order Gv-factor: Metal Folding, hypothesized to load the primary Vz; Group Embedded Figures test, Hidden Patterns and Copying, hypothesized to load Cf; Card Rotations, hypothesized to load S; and Disguised Word and Disguised Pictures, hypothesized to load Cs. These tests are subjected to a more detailed scrutiny below.

4.2.1 Metal Folding

The Metal Folding (MF) test was constructed by Svensson (1964, 1971). The task presented is to find the three-dimensional object among four alternatives that can be made from a flat piece of metal with bending lines marked on the drawings. In its original version the test consists of 40 items. For the present study, however, the test was shortened to 30 items, by excluding the last 10 items. The time limit for taking the test was reduced from 15 minutes to 10 minutes. The reason for shortening the test was that it was judged important to keep the administration time for the whole battery within reasonable limits.

On the 30-item test the mean is 18.5, with a standard deviation of 6.1, and a reliability coefficient of .87. This shows that even after the shortening, the test has quite good psychometric properties.

Descriptive item statistics are presented in Table 14. The proportion of correct answers ranges from .93 to .15, but for most items the proportion of correct answers exceeds .50. The biserial correlations vary greatly, with the lowest observed value being .38 and the highest value being .89. This indicates that the test may be quite heterogenous.

An analysis of the items with the highest biserial correlations shows that these items have some common characteristics, which may capture some of the essence of the Vz-factor. Items 17, 24, 26, 15, 19, 13, 18 and 20 are comparatively simple in the sense that there are not many foldings to be made. It also seems that it should be possible to complete the folding in a holistic way; that is, in one step. There are, furthermore, similarities in the global shape of the unfolded figure and one or more of the alternative objects to be choosen among. An analysis of the distributions of answers over the alternatives shows that it is precisely these alternatives that have attracted many incorrect answers.

An analysis of the items with low biserial correlations shows a more mixed picture. Many of the items with low correlations, however, do not require

Table 14. Proportions of correct answers and biserial correlations for the 30 items in the Metal Folding test.

Item	Proportion of correct answers	Biserial correlation
1	.89	.51
2	.93	.55
3	.85	.57
4	.90	.51
5	.80	.52
6	.75	.39
7	.83	.47
8	.72	.44
9	.74	.60
10	.66	.64
11	.68	.38
12	.35	.44
13	.70	.79
14	.61	.68
15	.66	.80
16	.62	.67
17	.66	.89
18	.67	.75
19	.60	.79
20	.66	.72
21	.27	.55
22	.50	.63
23	.54	.58
24	.55	.83
25	.36	.63
26	.57	.81
27	.42	.45
28	.50	.63
29	.41	.38
30	.15	.41

any folding at all since it is possible to conclude which answer is correct simply by noticing that some details are present in the unfolded figure and in the alternative, or by comparing the shape of a side in the unfolded figure and in the alternative objects. These items are often more complex, and the folding is not easily done in a holistic manner. One of the items with a low biserial correlation is, however, an exception. In item 11 the correct answer is a "house" with a flat roof and one of the alternatives is a house 'ith an ordinary leaning roof. This alternative was choosen by almost 23 per cent of the subjects. The familarity with this alternative may have caused it to be so attractive.

The test of fit to the Rasch model proved highly significant. The ML-ICCSL test gave a chi-square value of 1805.5 with 812 degrees of freedom. The ratio of the chi-square statistic to the degrees of freedom is about 2.2. Even with the size of the present sample this indicates substantial violations of the model assumptions, and provides further evidence that the test is quite heterogenous.

4.2.2 Group Embedded Figures

The Group Embedded Figures test (GEFT) was developed by Witkin et al. (1971). Here it is taken to represent Cf but within the framework of cognitive-style theory it is also used to measure field dependence - field independence. The subject's task is to locate a previously seen simple figure within a larger complex figure which has been so organized as to obscure or embed the sought-after simple figure.

The test used is a translation into Swedish of the GEFT. The test was translated and reproduced so as to deviate as little as possible from the original.

There are three sections in the test. The first section contains 7 very simple items which serve the purpose of making the subjects familiar with the test. Sections 2 and 3 each contain 9 more difficult items. In the present study, however, only sections 1 and 2 were administered. The time allowed for section 1 was 2 minutes while 5 minutes were given for section 2. It is recommended by Witkin, Oltman, Raskin and Karp (1971) that the scores on the first section are not included in a total score and this has not been done. In fact 1066 out of the 1224 subjects answered all items in the first section correctly.

For the 9 items in section 2 the mean is 3.88, with a standard deviation of 2.48. The reliability of the test is .76, which must be considered acceptable for a 9-item test.

Descriptive item statistics are presented in Table 15. The proportion of correct answers varies between .81 and .28. For most items, however, the proportion correct is in the .30 to .50 range.

Table 15. Proportions of correct answers and biserial correlations for the 9 items in section 2 of the Group Embedded Figures test.

Item	Proportion of correct answers	Biserial correlation
1	.81	.67
2	.41	.77
3	.44	.81
4	.31	.84
5	.35	.65
6	.29	.72
7	.52	.86
8	.49	.73
9	.28	.82

The biserial correlations vary between .65 and .86. Items 5 and 1 have the lowest correlations and items 4 and 7 the highest. In a Rasch analysis item 5 was identified as having too low a discrimination. For item 1 there was, however, no sign of any violations to the model assumption. The low correlation may instead be explained by the high proportion of correct answers for this item.

The task in item 5 is to find an embedded "T" in the more complex figure. The major source of error was that subjects failed to recognize the correct length of the bars in the "T". Most subjects, in fact, found a "T-like" simple figure in the more complex one. There were errors both with regard to the horizonatal and vertical bars, but especially for the length of the vertical bar. This indicates that the ability to retain an image of the exact perceptual features of a figure is not central to the ability measured by the test. Further evidence of this is provided by the items 6 and 8. These items also give room for identifying a simple figure not identical to the original sought for but which has the same overall shape, and they too have comparatively low biserial correlations. Items with high biserial correlations, items 4 and 7, on the other hand demand subjects to identify a three-dimensional object -- a rectangular block -- in the more complex configuration. Here there are limitied possibilities to make mistakes, either you identify the object or not. There are no perceptual details to be wrong about. It thus seems as if it is the ability to identify a figure or object, without regard to the exact perceptual details, that is essential.

According to Witkin et al. (1971) the test should be speeded. However, there is no evidence of speededness present in the data. Nor did observations during the administration of the test indicate that section 2 of the test is speeded with the present sample.

4.2.3 Hidden Patterns

The Hidden Patterns (HP) test, too, is intended to measure Cf. Each item consists of a geometrical pattern, and in some of these a single given configuration is embedded. The task is to mark each pattern in which the configuration occurs. These items closely resemble those of GEFT, but they are much easier, since there is only one simple configuration to look for and since the distracting pattern is less complex. The test is heavily speeded, there being 200 items in each of the two parts, with only 2 minutes allowed for each part.

The number of correct and incorrect identifications was counted, along with the ordinal number of the final item attempted. The test was scored in such a way that from the number of correct answers was subtracted the number of incorrect answers.

No subject attempted all items, and the maximum number of correct answers is 98 and 104 for part I and II, respectively. This indicates, along with the fact that few errors were made, that the test was indeed speeded.

The mean score on Part I is 32.9, with an sd of 12.7, and for Part II the corresponding figures are 37.7 and 15.3. A higher mean is thus observed for Part II. Since it is unlikely that the items of Part II are easier than the items of Part I, this higher mean can be attributed to learning effects. When taking Part II the subjects had more experience with the simple figure, and during the work with Part I they may have been able to develop strategies for coping with the items.

The correlation between the two parts is .72. Even though this coefficient is somewhat inflated by the fact that the two parts were administered in immediate succession it does indicate that a score based on both parts has a satisfactory reliability.

4.2.4 Copying

The Copying (CO) test is another test of the Cf-primary. The items in this test consist of a four-line geometrical figure, to be copied onto a square matrix of dots. It is believed that the test requires Cf in the act of superimposing the geometrical configuration onto the visual field defined by the matrix of dots.

The test is taken from the ETS Kit, and was originally suggested by McQuarrie's "Test for Mechanical Ability" and Thurstone's adoption of it. The test consists of two separately timed (3 minutes) parts, each of which has 32 items. The test is designed to be speeded.

Descriptive item statistics are presented in Table 16. From the figures it can be seen that the test was indeed speeded in the present sample: The proportion of correct answers diminishes as a function of the ordinal position of the items, and there is a positive correlation between item difficulty and item discrimination, which is typical of speeded tests (Gustafsson, 1980b).

Some items, however, have a medium proportion of correct answers, along with a low biserial correlation (items 5 and 6 in Part I, and items 1, 5 and 7 in Part II). Most of these items have the characteristic that in them the subjects must end the drawing within the dotted pattern, i.e. they must not finish a line at an outer row or column of the matrix of dots. One possible explanation why these items do not seem to be central to what is measured by the test, is that the able subjects develop strategies in which the global characteristics of both the figure to be copied and the dotted pattern are important. To minimize the confusion introduced by the matrix of dots, they have to have some point of reference, and the dots in the outer "frame" of the matrix may serve as such reference. This would make it easy to end the drawing of a line at the outermost dots, but at the same time the details of the figure may easily be overlooked.

To some extent these results are similar to the results obtained in the analyses of GEFT and MF. In these tests too items with low biserial correlations were found to have the characteristic that attention to details and exact configuration is important.

The mean number of correct answers is 11.3 on Part I, with a standard deviation of 4.0, and for Part II the corresponding figures are 11.1 and 4.9. The mean level of performance is thus the same on the two parts, but the variation in level of performance is somewhat larger for Part II. This could be an effect of some subjects developing appropriate strategies to cope with the task, while other subjects were frustrated and exhausted.

Table 16. Proportion of correct answers and biserial correlations for the two parts of the Copying test.

	Part I		Part II	
Item	Prop.corr	Biserial corr.	Prop. corr	Biserial corr.
1	.86	.44	.34	.39
2	.92	.46	.86	.39
3	.92	.58	.55	.52
4	.88	.57	.72	.51
5	.48	.24	.45	.42
6	.60	.33	.83	.42
7	.83	.52	.14	.42
8	.69	.50	.53	.49
9	.62	.59	.80	.53
10	.81	.77	.73	.60
11	.75	.77	.73	.56
12	.63	.82	.56	.68
13	.63	.78	.56	.71
14	.42	.72	.63	.69
15	.21	.65	.52	.70
16	.24	.73	.50	.77
17	.19	.80	.28	.77
18	.08	.74	.29	.81
19	.15	.76	.27	.74
20	.09	.76	.23	.74
21	.08	.64	.11	.69
22	.09	.81	.11	.78
23	.05	.82	.08	.83
24	.04	.86	.07	.78
25	.02	.68	.04	.79
26	.01	1.00	.02	.98
27	.02	.88	.03	.92
28	.01	.79	.03	.70
29	.01	.73	.03	.93
30	.01	1.02	.02	.58
31	.01	.69	.02	.54
32	.003	1.15	.01	.41

The two parts correlate .68, which indicates a satisfactory reliability for a combined score.

4.2.5 Card Rotations

The Card Rotations test (CR) is a test in the ETS battery. It is highly similar to the Thurstone Cards test, which along with Figures and Flags function as markers for Spatial Orientation (S) factor. This factor has been interpreted as an ability to perceive spatial patterns or to maintain orientation with respect to objects in space.

Each item in the CR-test presents a drawing of a card cut into an irregular shape. The task is to indicate which of eight other drawings of the same shape are merely rotated and not turned over onto the other side. The test is divided into two separately timed parts (4 minutes) with 14 items each.

There is no item level information for this test. The number of correct and incorrect answers was counted, and the last item attempted was recorded. The number of correct answers minus the number of incorrect answers is taken as the subject's score on the test. The maximum score which can be obtained on each part thus is 112.

Only 47 and 41 subjects attempted the last item on each of the two parts. This indicates that the test was speeded as planned. The mean score on Part I is 52.4, with an sd of 22.0, while the mean on Part II is 44.0, with an sd of 18.6. The mean on Part II thus is lower, which is probably due to fatigue.

The correlation between scores on the two parts is .75, which is the highest value observed among the tests with two parts.

4.2.6 Disguised Words

The Disguised Words (DW) test is constructed to measure the Cs factor. The test is highly similar to the test Mutilated Words used by Thurstone (1944) in the first study in which the Cs-factor could be identified. In the test words are presented with parts of each letter missing, and the task is to identify the word. The test consists of two separately timed (2 minutes and 30 seconds) parts, with 12 items each.

The subjects' responses were recorded as they were given, using the PRINS-system, in order to allow consistent judgement of the answers.

Table 17 presents descriptive item information. There is a clear tendency for the distribution of item difficulties to be bimodal, i.e. the items are either quite difficult or quite easy. The biserial correlations also vary greatly, indicating a substantial amount of heterogeneity among the items. Item 1 in Part I has the lowest biserial correlation. The reason why this item is so poor is probably that there was a competing response ("ska") to the correct answer ("eka"), and it does seem necessary to make a rather close analysis of the presented information to arrive at the correct answer. Since the essence of the Cs-factor is the ability rapidly to achieve closure this requirement of close attention to detail probably ruined the power of this item to reflect Cs.

Table 17. Item statistics for the Disguised Words test.

	Part I			Part II		
Item	Correct answer	Prop. correct	Biserial corr.	Correct answer	Prop. correct	Biserial corr.
1	eka (punt)	.26	.32	apa (monkey)	.82	.81
2	jag (me)	.44	.43	del (part)	.62	.48
3	katt (cat)	.53	.44	hund (dog)	.53	.65
4	skola (school)	.96	.48	sång (song)	.58	.80
5	stol (chair)	.96	.35	bord (table)	.83	.74
6	svar (answer)	.65	.57	folk (people)	.37	.72
7	hylla (shelf)	.87	.67	lampa (lamp)	.98	.45
8	morgon (morning)	.58	.75	frihet (liberty)	.12	.57
9	tidning (newspaper)	.19	.72	apelsin (orange)	.70	.69
10	misstanke (suspicion)	.03	.81	skandal (scandal)	.19	.52
11	papper (paper)	.15	.73	kartong (box)	.18	.75
12	tradition (tradition)	.09	.65	hemlighet (secret)	.08	.64

In Part I there is a tendency for the higher biserial correlations to be found for the items late in the test, but there is no such tendency for Part II. It can probably be assumed, therefore, that most subjects had sufficient time to make an attempt at answering most items in the test.

The reliability of Part I is .43, while for Part II a higher reliability of .66 is observed. For the combined score the reliability coefficient is .71. The reliability thus is rather low, which tends to be typical of tests of the Cs-factor. The mean of the combined test is 11.7 with an sd of 3.4.

4.2.7 Disguised Pictures

The Disguised Pictures (DP) test is another test purporting to represent Cs. The DP test is a shortened version of the original Street Gestalt Completion test described by Thurstone (1944, p. 8), mixed with items from the Concealed Pictures test from the ETS Kit. Each item presents a picture, parts of which are missing, and the task is to identify the object depicted. The test consists of two separately timed (2 minutes and 30 seconds) parts, with 12 items each.

In order to facilitate judgement of the subjects' answers, they were recorded as they were given in the PRINS-system. For several items very many responses were judged correct: for item 10 in Part I, for example, no less than 141 different responses (but including different spellings) were judged correct.

Table 18 presents descriptive item statistics. As was the case for the Disguised Words test, there is a clear tendency for the proportions of correct answers to the items to be either quite high or quite low. The biserial correlations tend, however, to show less variation for the DP test than was the case for the DW test. There is in the item statistics no indication that the test parts were speeded for the present sample.

The reliabilities are .52 and .57 for Parts I and II, respectively. For the combined score a reliability coefficient of .67 is obtained. The reliability of DP is thus even lower than the reliability of DW. The mean of the combined scores is 12.7, with an sd of 3.3.

Table 18. Item statistics for the Disguised Pictures test.

		Part I			Part II	
Item	Correct answer	Prop. correct	Biserial corr.	Correct answer	Prop. correct	Biserial corr.
1	car	.97	.68	dog	.997	.83
2	baby, bear	.96	.44	tricycle	.33	.66
3	skatingshoe	.71	.76	soldier	.79	.56
4	bird	.98	.61	tennisplayer	.80	.65
5	dog	.66	.57	engine	.13	.57
6	telephone	.31	.65	hen	.46	.45
7	house	.76	.69	shoe	.93	.72
8	faucet	.10	.57	power shovel	.52	.69
9	runner	.18	.57	horses & carriage	.09	.61
10	archer	.37	.58	boat	.66	.69
11	cart	.13	.56	flight of birds	.24	.56
12	rabbit	.29	.55	dancing couple	.30	.61

4.2.8 Discussion

The tests presented in this section are supposed to cover the primaries Vz, S, Cf, and Cs. This has not yet been demonstrated, however, and a discussion of the factorial structure thus has to be postponed till later. Here we will only discuss some of the results of the item analyses.

The tests vary in complexity. Metal Folding, which is supposed to define Vz, is the most complex test in the set. Generally, more complex spatial tests are supposed to contain a large amount of Vz (Lohman, 1979). In the item analyses it was, however, found that the "best" items were of lower complexity. Complexity is then defined as the number of foldings to be made.

This, of course, is a very limited definition, since it does not say much about the complexity of processing. However, in the "good" items is seems possible to arrive at a solution in a holistic manner, i.e. it seems to be possible mentally to transform the unfolded figure into an object in one step. Items which allow subjects to arrive at a solution by comparing some detail or details have a low discrimination. As a general tendency the more complex items did show low biserial correlations, but it is not possible to make any far-reaching interpretations from this fact. This is because the test can be suspected to be influenced by guessing, which would influence the discrimination in the way found (Gustafsson, 1980b).

Even though the "good" items seem to have the common characteristic that they are possible to solve by holistic processing, it is not possible to rule out other explanations on the basis of the item analysis. One of the problems with spatial tests is that they may readily be solved in different ways, for example by resorting to reasoning or other analytic means (cf. Lohman, 1979).

Group Embedded Figures, Hidden Patterns and Copying are supposed to define Flexibility of Closure (Cf). Among these tests, HP and CO are speeded. Witkin et al. (1971) recommended that also GEFT should be administered under speeded conditions, but there is not indication that this is the case in the present sample.

In the item analyses one interesting observation was made. Items in GEFT which seem to function well do not demand of subjects to pay attention to details. The two items in which subjects have to identify a three-dimensional object appeared especially good. In the CO-test it was found that for items with low discrimination subjects also have to pay attention to details, namely the exact position where a line ends. This indicates that a common characteristic of these tests is the holistic character of the processing involved in solving them. It is the ability to create and retain an image in the presence of distraction that matters, more than the vividness and exact shape of this image.

The Disquised Words and Disguised Pictures tests are supposed to measure the ability rapidly to achieve perceptual closure. These tests deviate from the other tests in the hypothesized Gv-cluster. While the other tests are rather abstract with regard to the figural content, the DP test is concrete in the sense that concrete objects are pictured. The DW test may also be suspected to have a verbal component since in this test subjects have to identify words.

4.3 Analyses of the Gc-tests

In the battery of cognitive tests there is only one test representative of the second-order factor Gc, the vocabulary test Opposites. However, the Standardized Achievement tests are hypothesized to contribute to the identification of Gc. Since no item level information is available for the tests, analyses cannot be carried out at the same level of detail as for the other tests. Results at the sub-test level are available, however, and the psychometric properties of the sub-tests can more or less be taken for granted since they have been carefully tried out.

4.3.1 Opposites

The Opposites (Op) test was constructed by Svensson (1964, 1971). It consists of 40 items in which the task is to select which word among four choices is the antonym of a given word. The test thus measures vocabulary and loads the V-primary. It was administered with a time limit of 10 minutes.

Table 19 presents the words tested (in translation) along with descriptive item statistics. There is a wide range of item difficulties, the proportion of correct answers being fairly evenly distributed between .99 and .15. The biserial correlations also vary greatly. There is, however, a tendency for the difficult items late in the test to have lower biserial correlations than the other items. This is most likely due to the fact that the test is a multiple-choice test which gives ample opportunities for guessing (cf Gustafsson, 1980b).

The mean of raw scores on the test is 21.8, with an sd of 5.7. The reliability is .80.

4.3.2 Descriptions of the Standardized Achievement tests

For most of the pupils in the sample results are available at the sub-test level for the Standardized Achievement tests. This presents an opportunity to analyze the factorial structure of these tests. Such analyses are presented in the following section, while the sub-tests are described in the present section.

The Standardized Achievement test in Swedish (SA) consists of 6 sub-tests:

1. Spelling contains 25 items in which the task is to correctly spell dictated words.

2. Reading Comprehension attempts to measure the pupils´ ability to understand texts written in different styles and with different contents. The test presents 6 different texts of 100-200 words in relation to which 2 to 5 multiple-choice questions with 5 alternatives are asked. In all there are 21 items in the test, which is administered with a time-limit of 35 minutes.

Table 19. Item statistics for the Opposites test.

Item	Word tested	Prop. correct	Biserial corr.	Item	Word tested	Prop. correct	Biserial corr.
1	Beautiful	.99	.22	21	Petty	.48	.54
2	Open	.99	.33	22	Anonymous	.60	.58
3	Cold	.98	.23	23	Reckless	.51	.56
4	Glad	.95	.13	24	Noble	.31	.41
5	Destroyed	.92	.59	25	Separate	.53	.51
6	Rare	.90	.63	26	Ambitious	.31	.37
7	Clear	.74	.44	27	Impulsive	.37	.47
8	Spurt	.79	.53	28	Agreeable	.22	.40
9	Wag	.76	.63	29	Reject	.39	.34
10	Generous	.81	.55	30	Permanent	.40	.55
11	Permanent	.60	.59	31	Humble	.16	.21
12	Light	.75	.62	32	Hilly	.33	.47
13	Recommend	.76	.65	33	Destitution	.21	.39
14	Smooth	.70	.49	34	Mannerly	.31	.28
15	Desert	.67	.50	35	Fool-hardy	.28	.57
16	Assent	.52	.51	36	Concrete	.26	.27
17	Idler	.58	.57	37	Important	.29	.57
18	Gay	.50	.54	38	Ample	.15	.55
19	Attack	.77	.37	39	Melancholic	.24	.28
20	Depressed	.55	.57	40	Intolerant	.23	.19

3. Words of Relation tests the pupils´ ability to use conjunctions and adverbs. An 8-sentence text is presented in which 12 words are missing and the task is to select these words from a list of 28 words. The time limit is 12 minutes.

4. Vocabulary consists of items in which the synonym of a word presented in a one-sentence context is to be selected from a list of 5 choices. There are 25 items in the test and the time limit is 12 minutes.

5. Word List tests the pupils´ ability to use a word list to find the meaning, spelling and flexion of words. The test consists of 11 questions to be answered by use of a word-list covering the letter N. It is administered with a time limit of 10 minutes.

6. Sentence Construction presents a text lacking punctuation, and the task is to add the 18 punctuation marks which are missing. The time limit is 15 minutes.

The Standardized Achievement test in mathematics (MA) is composed of 5 sub-tests:

1. Numerical Calculations presents 20 items testing understanding of the number line, the ability to carry out additions, subtractions, multiplications, divisions and calculations with fractions. The time limit is 35 minutes.

2. Per Cent Calculations tests the ability to carry out calculations involving the per cent concept. There are 16 items in the test, which are to be solved during at most 25 minutes.

3. Estimates tests the ability to make rapid estimates of the approximate result of an expression. The test is a multiple-choice test with 21 items and a time-limit of 10 minutes.

4. Geometry and Diagrams consists of 8 geometry items presenting tasks such as computing the area of rectangles, and 6 items assessing the ability to understand information presented in graphs and tables.

5. Applied Computations presents 12 verbally stated problems, most of which require a mixture of the rules of arithmetic.

In the Standardized Achievement test in English (EA) there are 4 sub-tests:

1. Vocabulary consists of 40 multiple-choice items which present a 1- or 2-sentence context (and for some items a picture). One word is missing and the task is to indicate this word. The time limit is 30 minutes.

2. Listening Comprehension presents via tape-recorder brief pieces of information, in relation to which questions are asked. The questions are answered by indicating the appropriate alternative among 4 given ones. For 15 questions the alternatives are verbal, and for 20 questions the alternatives are pictorial. This test takes 30 minutes to administer.

3. Forms and Structures tests the knowledge of grammatical rules, such as the do-construction, flexion of verbs, and so on. The items are presented in groups of 2 to 4 within a context of a few sentences. Each group of items has 3 to 5 alternatives in common, one of which is to be selected for each item. In all there are 40 items in the test and the time limit is 30 minutes.

4. Reading Comprehension consists of three different types of items. In one type, of which there are 9 items, a sentence is presented in which a word is missing. This word is to be identified in a list of 4 alternatives. In another type, of which there are 5 items, a one-sentence question is asked, and the task is to select the appropriate answer from a list of 4 alternatives. In the third type of items 5 texts of 75-200 words are presented, in relation to each of which 3 to 5 multiple-choice questions are asked. In all there are 15 items of this type. The time limit for the test is 30 minutes.

4.3.3 Analyses of the Standardized Achievement tests

Table 20 presents descriptive statistics for the sub-tests of the Standardized Achievement tests. For most of the sub-tests the mean is close to about half the maximum score but some of the subtests appear to be somewhat too easy for the present sample. The distributions of raw scores for these tests have been investigated, and they do not appear to deviate too much from normality, however.

In order to investigate the factorial structure of the sub-tests a LISREL-model was first fitted. The model tried was a simple 3-factor model, with the sub-tests of each Standardized Achievement test taken to define one factor. This model fitted poorly, however (chi-square=304.5, df=87).

Since the sub-tests of the Standardized Achievement tests have not previously been investigated with factor-analytic methods, there is no source of hypotheses about alternative models. Therefore, instead of trying more or less blindly to modify this LISREL model, exploratory factor analysis was resorted to.

Maximum likelihood factor analysis (Jöreskog & Sörbom, 1976) was applied under the assumption of 1 to 5 common factors, yielding the following results for the test of the number of common factors:

Number of factors	chi-square	df	p-value
1	1634.1	90	.000
2	469.5	76	.000
3	202.0	63	.000
4	88.6	51	.001
5	62.5	40	.013

Table 20. Descriptive statistics for the sub-tests in the Standardized Achievement tests. N=981.

Test	Maximum score	mean	sd
<u>Swedish</u>			
Spelling	25	17.8	4.5
Reading Comprehension	21	10.7	3.7
Words of Relation	16	7.6	2.7
Vocabulary	25	12.6	5.0
Word List	11	7.3	2.1
Sentence Construction	18	13.3	3.9
<u>Mathematics</u>			
Numerical Calculations	20	13.5	4.0
Per Cent Calculations	16	11.8	4.1
Estimates	21	12.0	3.7
Geometry and Diagrams	14	7.8	2.9
Applied Computations	12	6.8	2.6
<u>English</u>			
Vocabulary	40	27.7	8.1
Listening Comprehension	35	24.7	5.7
Forms and Structures	40	27.1	7.0
Reading Comprehension	30	21.5	6.3

Even in the 5-factor solution the chi-square is somewhat too high in relation to the number of degrees of freedom, which indicates that as many as 6 common factors may be needed to account for the intercorrelations among the sub-tests. However, technical problems appeared already in the 5-factor solution, which did not converge properly, and in which one factor was loaded highly by one test only. Thus it seems that with exploratory methods at most 4 factors can be extracted. The Promax-rotated loadings (with k=3) in the 4-factor solution are shown in Table 21, along with the results from the 3-factor solution. In the 3-factor solution each factor essentially corresponds to one of the Standardized Achievement tests, and most sub-tests have loadings in one factor only. In the 4-factor solution there is one Mathematics and one English factor, but the sub-tests in the Swedish test have loadings in two factors. One of these is highly loaded by Vocabulary and Reading Comprehension, and the other by Sentence Construction, Spelling, Words of Relation, Word List, and by Forms and

Structures from the English test. This latter factor seems to reflect facility with the formal, grammatical aspects of language, while the former factor seems to reflect vocabulary and comprehension.

Table 21. Promax-rotated factor loadings in the 3- and 4-factor solutions.

	3 factors			4 factors			
Test	1	2	3	1	2	3	4
Swedish							
Spelling	.23		.51	.20			.54
Reading Comprehension			.63			.49	.22
Words of Relation			.77			.28	.54
Vocabulary			.59			.67	
Word List		.21	.58				.52
Sentence Construction			.48				.63
Mathematics							
Numerical Calculations		.79			.79		
Per Cent Calculations		.77			.76		
Estimates		.68			.68		
Geometry and Diagrams		.73			.73		
Applied Computations		.67	.24		.67		
English							
Vocabulary	.79			.76			
Listening Comprehension	.72			.75			
Forms and Structures	.61		.24	.60			.35
Reading Comprehension	.68		.29	.67			

Note. Only loadings .20 or higher are shown.

4.3.4 Discussion

With the exception of the Swedish test it appears that a sum of the sub-test scores to form an overall score on each test implies a very small loss of infcrmation. However, also for the Swedish test a sum of the sub-test scores may represent the important source of individual differences in spite of the fact that the sub-tests were found to load two factors. This is because one of the factors found appears to be more or less identical with the V-factor, which in the test battery is already represented by the Op test. The V-variance in the Swedish test may, therefore, perhaps be accounted for by a loading in the V-factor. In the model analyses to be presented in the next section the Standardized Achievement tests will therefore not be represented at the sub-test level but as sums.

4.4 Conclusions

The primary aim of the analyses reported here has been to investigate the psychometric properties of the tests. For almost all the tests it may be concluded that these are good: The reliabilities tend, with the exception of DW and DP, to be in the 80´s to 90´s; the mean score on the unspeeded tests tends to be close to about half the maximum possible score; and most tests show a high degree of internal consistency.

It may be noted that none of the tests fits the Rasch model according to the statistical tests of model fit. However, the fairly large sample of subjects analyzed here makes even quite minor deviations appear highly statistically significant. According to graphical analyses of fit (see Appendix 2) some of the tests (RA, LG and GEFT in particular) seem to fit quite well. It should also be pointed out that in applications such as the present study the most profitable use of the Rasch model appears to be in comparative studies of the degrees of fit obtained, for example, with different scoring systems. However, requirements that the test-statistics should be non-significant would have absurd implications, forcing the test developer to spend large amounts of time on revision of tests and items, and it might imply an undue narrowing of the scope of the test (cf Gustafsson, 1980b).

5 MODELS FOR THE TEST-BATTERY

In order to test the validity of the HILI model developed in chapter 2, the LISREL-technique is relied upon. In applying this method of analysis it is often, however, a useful stra'egy to develop and test a model by first analyzing sub-sets of the variables in sub-models, which are then pieced together into one model. This strategy has been adopted here, and we will begin by considering models for the Gf- and Gv-tests.

5.1 Models for Gf and Gv

As has been described in previous chapters the tests in the battery are hypothesized to load several different primary facors. Some primary factors are represented by one test only, however. In those cases this single test is used to represent the primary factor. This is done by splitting the test into half-tests, which are both entered into the model. With this procedure, the primary factor also represents the unique variance of the test, but for most of the tests analyzed here the amount of unique variance is likely to be small.

In the Gf- and Gv-domains the following primary factors have been hypothesized:

- Induction (I), meaaured by LS and LG.
- Cognition of Figural Relations (CFR), measured by RA. The test RA is split into half-tests, by assigning odd numbered items to one test, and even numbered items to the other (RA-O and RA-E).
- Visualization (Vz), measured by MF. The MF-test is split into half-tests, according to the odd/even rule to give the half-tests MF-O and MF-E.
- Spatial Orientation (S), measured by the two parts of CR (CR-I and CR-II).
- Flexibil. of Closure (Cf), measured by GEFT, HP and CO.

- Speed of Closure (Cs), measured by DP and DW.

The factors I and CFR are hypothesized to load Gf, while the primaries Vz, S, Cf and CS are hypothesized to load the secondary factor Gv. It will be remembered that the primary factor Ms, represented by ANS and ALS, has also been hypothesized to have a weak relation with Gf. Investigation of this hypothesis, is, however, post-poned till later in this section.

The hypotheses expressed verbally above are represented graphically in the LıSREL-model depicted in Figure 7 (see also Appendix 1).

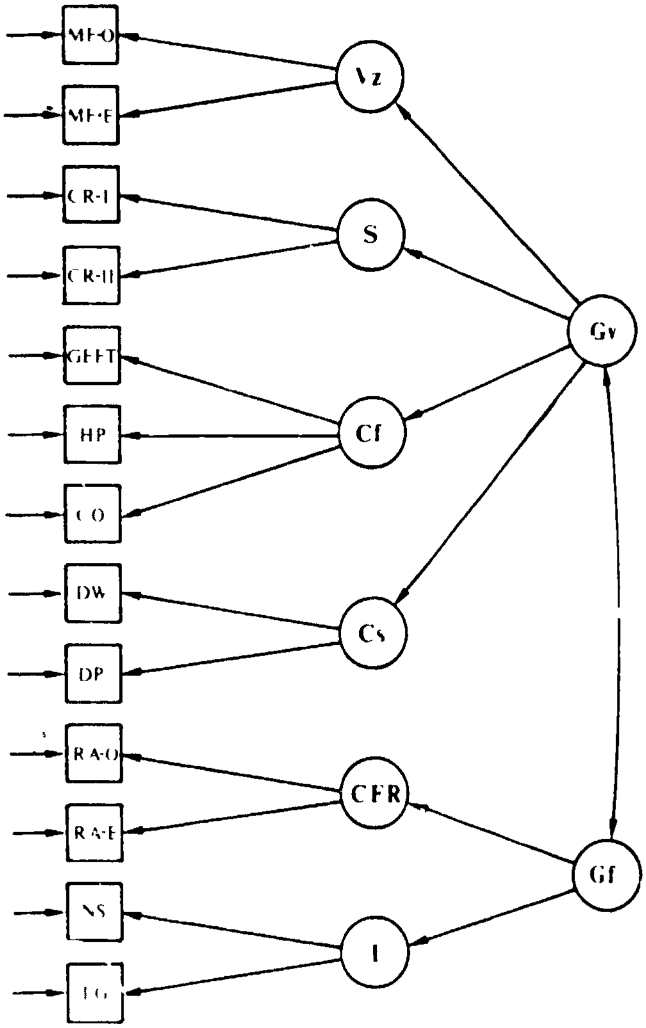


Figure /. The hypothesized model for the Gf- and Gv-tests.

Estimating this model from the correlation matrix for the sample of 1224 subjects a rather poor fit was obtained (chi-square=154.1, df=58). Even though the large sample of subjects necessarily inflates the chi-square statistic even when there are only minor deviations from the model, there seems to be room for improvement of fit. A series of modifications of the model were, therefore, introduced.

The final model is shown in Figure 8. This model has a very good fit (chi-square=53.7, df=51, p <.37).

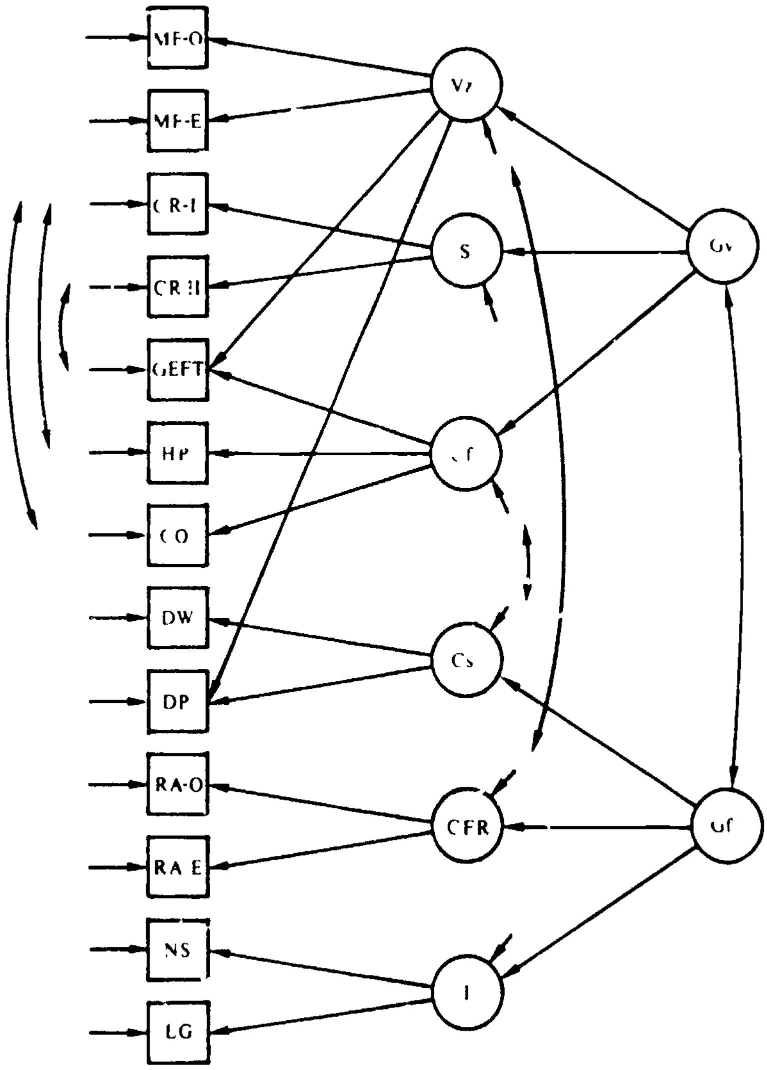


Figure 8. The final model for the Gf- and Gv-tests.

In this model there are several significant relations which were not explicitly hypothesized:

- The DP-test was found to have a small but significant loading in the Vz-factor, along with the loading in Cs. This may be attributed to the fact that the DP-test involves figural content, while DW involves verbal content. As a consequence of this loading of DP in Vz the Cs-factor is poorly identified, however, with a very high loading in DW and a very low loading in DP. This implies that in the present model the Cs-factor is virtually identical with the true variance of the DW-test. It should also be pointed out that the true variance of DW is probably overestimated since this estimate is considerably higher than the reliability of the test.

- The model fits better when Cs is taken to load Gf (chi-square=53.7, df=51) than when it is taken to load Gv (chi-square=62.5, df=51). This indicates that once the variance due to figural content of Cs-tests is accounted for by relations at the primary level, the Cs-factor may not belong with the Gv-complex, but with Gf (cf Thurstone, 1944). It must be stressed, however, that in the present study the Cs-factor may not be properly identified.

- There is a significant (t=3.23) covariance between the residuals of Cf and Cs which indicates that it may be possible to define a second-order closure factor.

- There is also a significant (t=4.83) covariance between the residuals of CFR (or the RA-test) and Vz (or the MF-test) which shows that RA, along with the Gf-variance in the tests, taps an ability in the Gv-domain.

- GEFT has a significant loading in the Vz-factor (t=4.38). This may be because GEFT is the only unspeeded test defining the Cf-factor.

- There are significant covariances between the errors of CR-I and CO (t=3.19), CR-I and HP (t=3.61), and CR-II and GEFT (t=-3.81). The positive correlations all involve speeded Gv-tests, while the negative correlation involves one unspeeded and one speeded Gv-test. This indicates that there may be another factor in the Gv-domain, representing the ability rapidly to perform relatively simple spatial tasks.

Before these results are discussed at greater length, some further empirical results will be reported. It will be remembered that the analysis of the RA-test in section 4.1.3 showed the sub-sets of items to be heterogeneous. The heterogeneity is not evident in the model shown in Figure 8, and it may be asked whether performance on the sub-sets in RA has variance in common with the other tests in the battery.

In order to answer this question, a special model was constructed. This model, which is shown in Figure 9 (see also Appendix 1), contains two hierarchies: one hierarchy orming a set of independent variables and the other forming a set of dependent variables. Independent variables are all tests except RA the primary factors I, Vz, S, Cf and Cs, and a second-order G-factor. Dependent variables are the RA sub-sets which are divided into half-tests to define the latent variables accounting for sub-set performance. These latent variables in turn define an RA-factor. Here the relations among the independent and dependent variables are of primary interest.

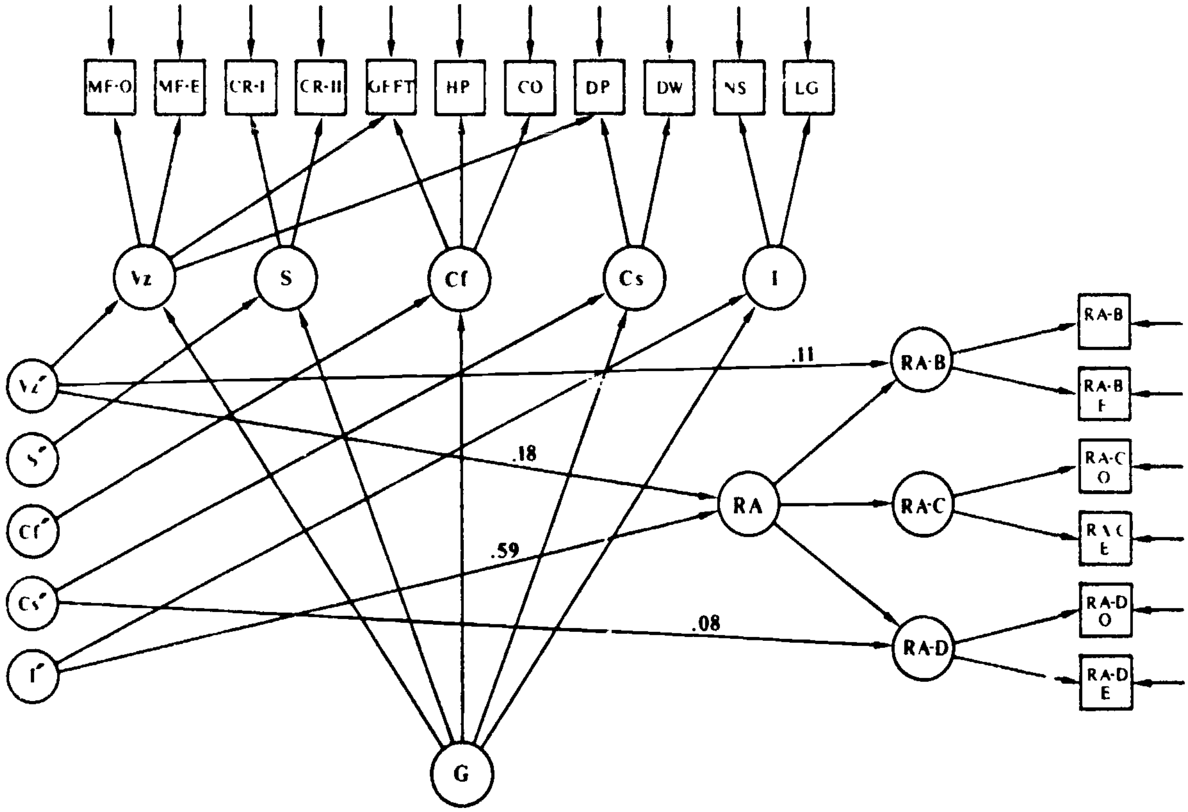


Figure 9. The model used for the Raven Progressive Matrices test.

The model shown in Figure 9 has a non-significant test-statistic (chi-square=123.9, df=102, p <.07). In this model the dependent RA-factor is predicted by I and Vz, which corresponds to the results obtained in the overall model (Figure 8). However, as can be seen in Figure 9 there also are relations between Vz and RA-B and between Cs and RA-C. These relations indicate that there is indeed within the sub-sets of items systematic variance which may be analyzed in terms of primary factors. In some items in the B-set (e.g. B5, B6 and B7) the rule defines a change of position of geometrical forms, which may accour for why there is Vz-variance in this sub-set beyond that in the othe -sets. It is, however, more difficult to explain the Cs-variance set, and until replicated this result has better be left uninterp

The Ms-factor is left out of the model shown in Figure 8. Adding, however, the tests ANS and ALS to define the primary factor Ms, which in turn is hypothesized to have a loading in Gf, the fit of the model does not deteriorate appreciably (chi-square=87.0, df=75, p <.15) The standardized loading of Ms in Gf is .45. This result thus supports the Cattell and Horn (1966) hypothesis that Ms has a weak relationship with Gf.

The primary purpose of the model tests performed in this section has been to see whether the hypothesized structure in the Gf- and Gv-domains receives support or not. Even though it has been necessary to modify the original model (Figure 7) it may be concluded that most hypotheses have been supported: All primary factors, except perhaps Cs, have been identified, and these first-order factors seem to load the two second-order factors Gf and Gc according to the hypothesized pattern.

Most modifications of the originally hypothesized model had to be made in the Gv-domain. Several of these modifications indicate the presence of further factors at both the first- and second-order levels. It will be remembered that the reanalysis of the Botzum data (see section 2.7.1) yielded no less than 3 second-order factors in the Gv-domain, and some of the present findings indirectly support these results. Thus, the covariance found between the residuals of Cf and Cs supports the second-order closure factor (Gvs) found in the Botzum data, and the relations between Vz, CFR and GEFT strongly resemble the Gvr-factor, which was interpreted as an ability to retain images in the presence of distractions. In the present study these are minor factors, however, and with 3 or 4 primary factors only it is impossible to define more than one second-order factor. Further study of the hierarchical structure of the Gv-domain will, therefore, have to await analysis of larger test-batteries.

The Cs-factor is the only primary factor which is not clearly identified in the model. The loading of DP in Vz causes DW to take on a very high loading in the Cs-factor, and moves this factor from Gv to Gf.

That measurement of Cs is difficult is also seen from the low reliabilities of the DW- and DP-tests. It may, indeed, be that the present approach of trying to measure Cs with group-administered paper- and pencil tests is technically deficient. When Thurstone (1944) first identified the factor he used similar items, but these were displayed one at a time individually to the subjects, with strict control of item exposure time and viewing distance. Furthermore, the response measure used by Thurstone was the number of items requiring less than 3 seconds for correct identification. With the present approach such tight control is impossible, and the "speed"-aspect of Speed of Closure may have to yield for other ways of solving the items, such as by application of analytic reasoning strategies (cf Botzum, 1951; Lohman, 1977).

5.2 Models for Gc

It might be possible to construct a confirmatory factor-analytic model for the sub-tests of the Standardized Achievement tests on the basis of the exploratory analyses presented in section 4.3.3. However, such a model would be an ad hoc model, and the large number of sub-tests would make it unwieldy. The unweighted sums of the sub-test scores will, therefore, be used in the modelling.

The exploratory analyses tests showed the Swedish test to be affected by two factors. However, one of these -- the Verbal Comprehension factor -- is in the test battery represented by the vocabulary test Opposites as well. It may be possible, therefore, to represent the V-variance in the Swedish test by a loading in the V-factor.

This line of reasoning suggests a model for the Gc-tests which contains one general Scholastic Achievement factor, loaded by the three Standardized Achievement tests, and one Verbal Comprehension factor, loaded by Opposites (represented by half-tests formed according to the odd/even rule) and the Swedish Achievement test.

Estimating this model from the matrix of intercorrelations for those 981 subjects which have results on all the Standardized Achievement tests a very good fit was obtained (chi-square=4.1, df=3, p <.25). The loading of the Swedish Achievement test in the V-factor is significant (t=3.11). Even though the factors in this model are highly correlated (.80) these results indicate that in the present battery there are at least two primary factors in the Gc-domain.

5.3 The model with three second-order factors

The model devloped for the Gv- and Gf-tests shows a good fit, and so does the model developed for the Gc-tests. However, when these models were combined into one model with three second-order factors, a very poor fit was obtained (chi-square=529.4, df=150). A partial explanation for this poor fit is that this result was obtained with the number of subjects taken to be 1224, which is not quite correct since 343 subjects lack results on one or more of the Standardized Achievement tests. This explains, however, only a small part of the poor fit, since when the model was estimated for the 981 subjects with a complete set of results the fit was not much better (chi-square=431.0, df=150).

The major reason for the poor fit of this model is that the Mathematics Achievement test has variance in common with Number Series II. This made it possible to introduce yet another primary factor (Num-Ach), loaded by the Mathematics test and Number Series II. This modification, along with a few others, improved fit to an acceptable level (chi-square=184.5, df=144, p <.013, N=981).

The final model is shown in Figure 10. As can be seen from the figure the introduction of the Num-Ach factor has turned the general scholastic achievement factor into a pure Verbal Achievement (Ve-Ach) factor, loaded by the Swedish and English Achievement tests. The Num-Ach factor has its highest loading in Gc, but there is also a loading in Gv which may be due to the fact that at least one sub-test (Geometry and Diagrams) in the Mathematics Achievement test involves figural content.

As can be seen from Figure 10 some other modifications have been introduced as well. Covariances are allowed between the residuals of Cf and Ve-Ach, and between the residuals of Ms and Ve-Ach. Covariances also are estimated between the specific factors of Disguised Words and Swedish Achievement, and between the specific factors of the Auditory Letter Span test and Swedish Achievement. Even though these covariances are all statistically significant, the estimates of the parameters are all rather small, and they have to be considered relatively unimportant. It should be pointed out, however, that the covariance found between Cf and Ve-Ach brings some support for the Witkin et al. (1977) claims of the importance of field independence (or Cf) in school-learning. Furthermore, the covariance between Ms and Ve-Ach supports the Das et al. (1979) argument that successive processing is important in language acquisition and production.

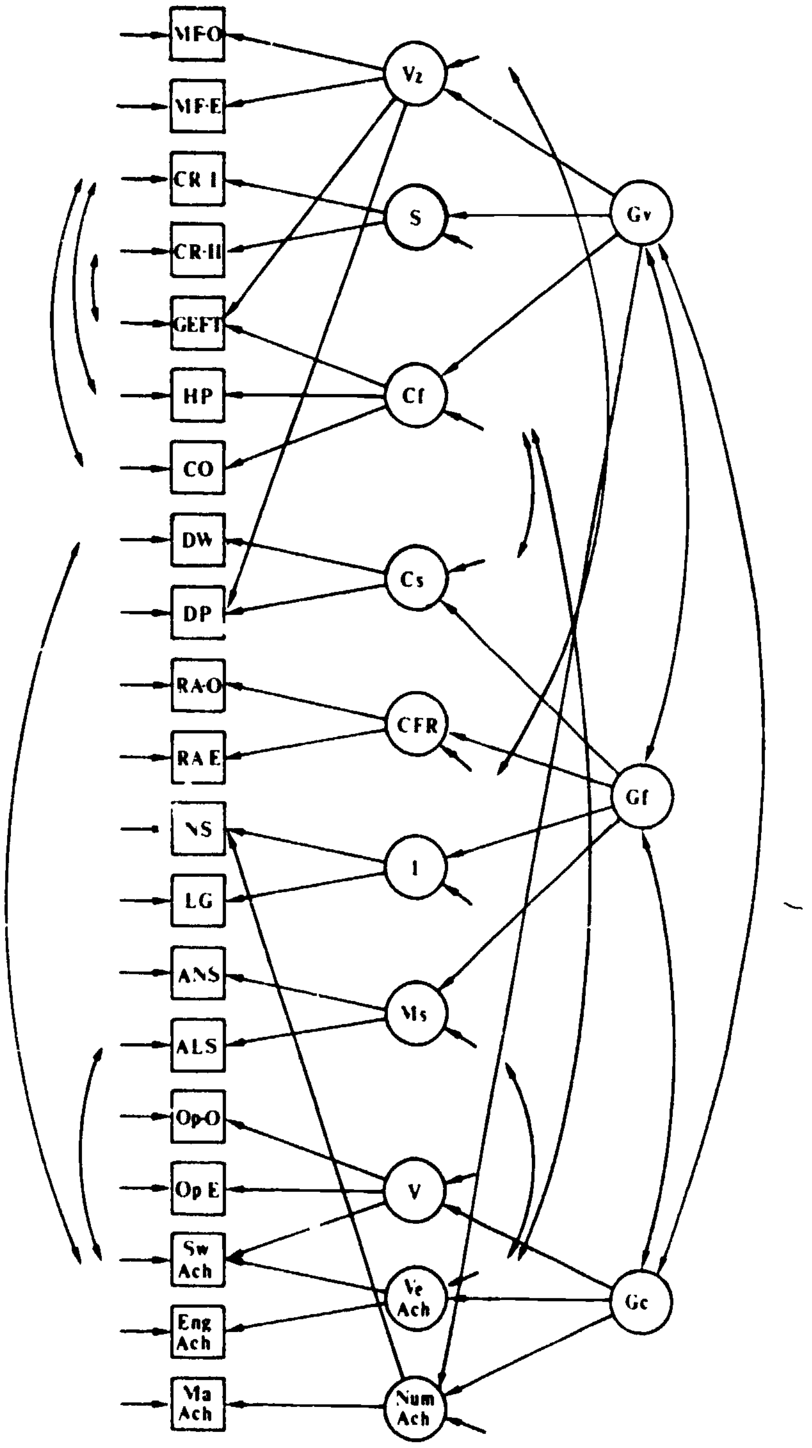


Figure 10. The model with three second-order factors.

5.4 The model with a third-order factor

The model with a third-order G-factor is identical with the model shown in Figure 10, except that another factor is introduced to account for the intercorrelations among the three second-order factors.

With three second-order factors a model with one third-order factor is "just identified", i.e. it has the same chi-square and the same degrees of freedom as the second-order model. By imposing further restrictions it is, however, possible to define a fully identified model. One such restriction is, of course, given by the hypothesis of primary interest here, namely that Gf is identical with G. This hypothesis implies that the residual variance in the second-order Gf-factor is zero. Estimating the model with this constraint imposed the test of fit yielded a chi-square of 187.7 with 145 degrees of freedom. The difference between this statistic and the statistic for the model without a third-order G-factor is not significant (chi-square=3.2, df=1), which result strongly supports the hypothesis that Gf is identical with G.

An alternativ? explanation may of course be that the test lacks statistical power. This alternative hypothesis can be ruled out, however, since the estimate of the loading of Gf in G was slightly above unity, and since a very poor fit was obtained when the restriction was imposed for Gv (chi-square=98.2, df=1) and Gc (chi-square=126.6, df=1).

Before some descriptive results from this final model are presented, there is reason to present results from one further investigation. It will be remembered that in the scoring of Number Series II two different scoring systems were applied. The results so far have been achieved with the test scored according to the Strict scoring system, but it is of course of interest to investigate whether the validity of Number Series II is improved with the Liberal scoring scheme. Estimating the model with the liberally scored version of Number Series II included instead, a slightly higher test statistic (chi-square=190.3, df=145) was obtained. Only few of the parameter estimates were affected at all. The loading of I in Gf increased, however, marginally from .987 to 1.000. It would thus appear that the liberally scored version is slightly more valid than the version with strict scoring. The difference seems so slight, however, that the considerable amount of extra effort involved in the liberal scoring does not seem worthwhile.

The final model is presented in Figure 11, .d Table 22 presents estimates of correlations between the observed variables and each of the higher order factors.

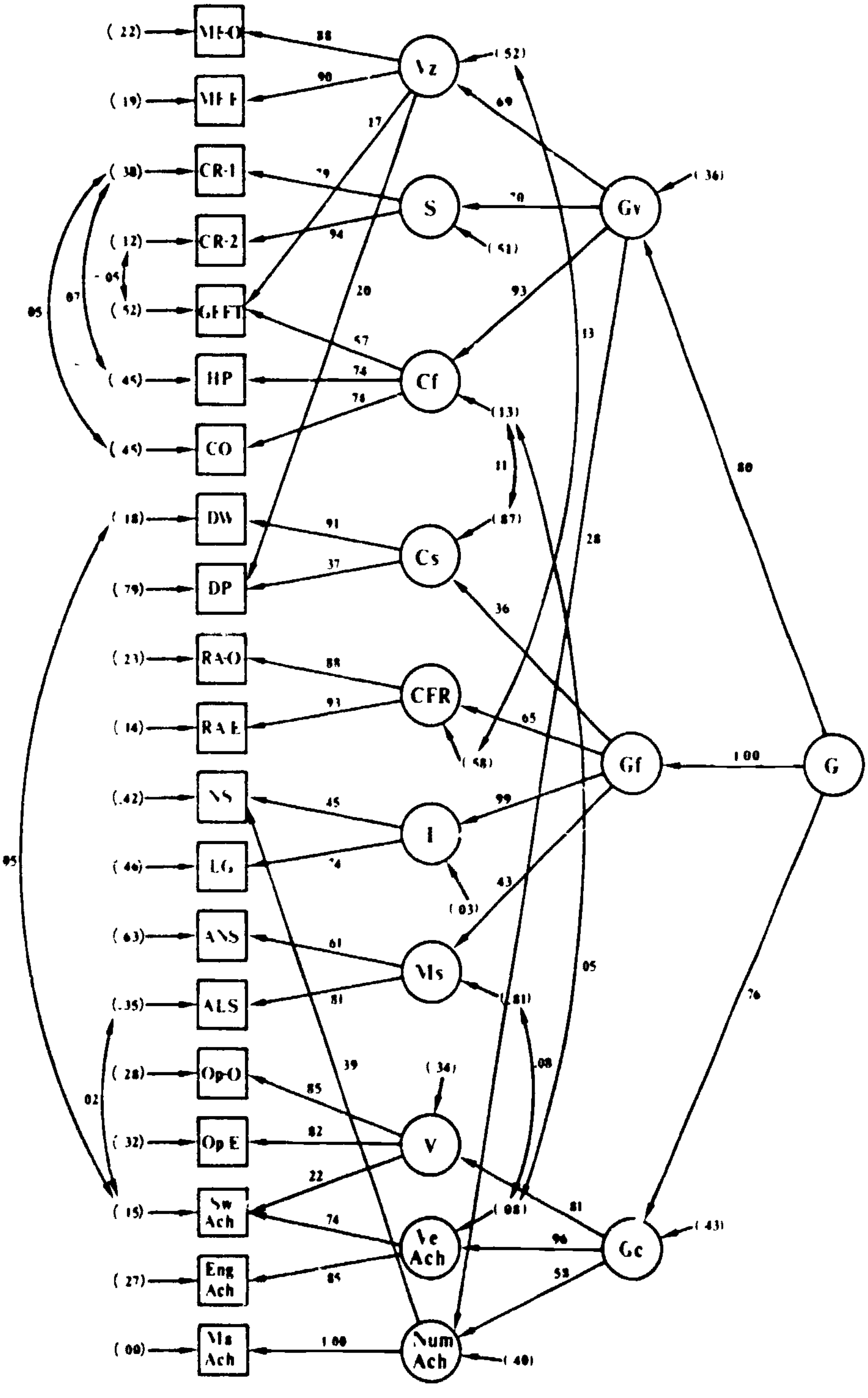


Figure 11. The model with a third-order factor

Among the primary factors I has the highest correlation with G, and this correlation is virtually perf ct. The next highest correlations between G and primaries are observed for Cf and Ve-Ach. It may also be observed that

Table 22. Correlations between the higher-order factors and the tests and primary factors.

Variable	G	Gv	Gf	Gc
Vz	.55	.69	.55	.42
S	.56	.70	.56	.42
Cf	.74	.93	.74	.56
Cs	.36	.29	.36	.27
CFR	.65	.52	.65	.49
I	.99	.79	.99	.75
Ms	.43	.35	.43	.33
V	.61	.49	.61	.81
Ve-Ach	.73	.58	.73	.96
Num-Ach	.66	.63	.66	.74
MF-O	.24	.37	.28	.25
MF-E	.24	.38	.28	.26
CR-I	.21	.34	.25	.23
CR-II	.25	.40	.30	.27
GEFT	.25	.40	.30	.25
HP	.27	.42	.29	.27
CO	.27	.42	.31	.27
DW	.16	.16	.19	.17
DP	.12	.15	.14	.13
RA-O	.28	.28	.33	.30
RA-E	.29	.29	.34	.31
NS	.34	.36	.40	.43
LG	.35	.35	.42	.38
ANS	.13	.13	.15	.14
ALS	.17	.17	.20	.18
Op-O	.25	.25	.30	.47
Op-E	.25	.25	.29	.46
Sw-Ach	.33	.33	.38	.61
Ma-Ach	.32	.38	.38	.51
Eng-Ach	.30	.30	.35	.56

all the primary factors have rather high correlations with all the secondary factors, which, of course, reflects the substantial involvement of G in the secondaries.

The correlations between the tests and G all are considerably lower than the correlations between the primary factors and G, which is because the tests are affected by errors of measurement and test specificity. The highest correlations are obtained for the two I-tests, but the highly reliable Standardized Achievement tests involve almost the same amount of G-variance. The low and even correlations seem to indicate that to estimate properly the higher-order factors, it may be necessary to use a rather large and representative test battery.

5.5 Discussion

The LISREL analyses of the test battery provide strong support for the HILI model. Not only does the pattern of loadings in the primary and secondary factors in most respects conform with expectations, but support is also obtained for the hypothesis that Gf is identical with G.

It must of course be realized that the present study tests only a sub-set of the model since it is restricted to Gf, Gc and Gv. However, the reanalysis of the Undheim (1978) study included the Gs and Gr factors as well, and taken together these studies very strongly support the HILI model as sketched upon in Figure 6.

6 ANALYSES OF SEX DIFFERENCES

Sex differences in level of performance on mental tests is a problem to which much attention has been devoted (for reviews see e.g. Anastasi, 1958; Maccoby, 1966; Maccoby & Jacklin, 1974), and some tentative generalizations have been reached. Thus, during adolescence boys tend to score higher than girls on tests with numerical and spatial content, while girls tend to score higher on tests representing verbal abilities (Maccoby & Jacklin, 1974). The differences are not large, however, and contradictory results are legion.

But the sexes may differ not only in level of performance on tests, but also in other, and perhaps more interesting respects, such as the strategies relied upon in solving the test items, test reliability, variability of ability, and so on. Such types of sex differences have not been much studied. One of the reasons why the study of sex differences has been concentrated upon comparisons of level of performance is undoubtedly that technical problems have prevented a closer analysis of other types of sex differences.

The LISREL technique represents, however, an improvement in this respect since with this methodology it is possible to study not only mean differences in tests and factors, but also differences in factor loadings, factor variances and error variances. While LISREL models for one group of persons are most commonly estimated from the correlation matrix, models involving two or more groups of persons are best estimated from the moment matrix around zero, which makes it possible to take into account differences in test and factor means, as well as differences in variances.

Table 23 presents means and standard deviations on the tests for the boys and girls in the sample. The boys have a higher mean on three tests (Number Series II, Card Rotation, and Mathematics Achievement), while the girls have a higher observed mean on all other tests. The fact that the girls tend to score higher is most likely an effect of the mental growth spurt for girls aged 12 to 13 identified by Ljung (1965). An alternative hypothesis may be that the sample is less representative for one of the sexes, but this hypothesis seems unlikely, given the very low rate of attrition for the battery of cognitive tests.

Table 23. Means and standard deviations for the tests in the battery for boys and girls.

	Boys			Girls		
	x	s	N	x	s	N
Number Series II	7.95	4.00	602	7.69	3.50	622
Letter Grouping II	10.03	3.48	602	11.87	3.51	622
Raven	23.42	6.34	602	24.88	5.49	622
Auditory Number Span	4.31	2.91	602	4.51	2.43	622
Auditory Letter Span	4.40	2.22	602	4.64	2.12	622
Metal Folding	18.29	6.41	602	18.75	5.84	622
Group Embedded Figures	3.78	2.51	602	3.98	2.45	622
Hidden Patterns	66.42	23.10	602	74.28	24.22	622
Copying	21.91	8.21	602	22.91	8.15	622
Card Rotation	98.91	37.43	602	93.96	38.51	622
Disguised Words	11.71	3.50	602	11.73	3.34	622
Disguised Pictures	12.86	3.53	602	12.50	3.14	622
Opposites	21.26	5.74	602	22.34	5.71	622
Swedish Achievement	64.83	18.25	541	71.71	16.30	558
English Achievement	96.98	26.32	521	104.13	22.68	537
Mathematics Achievement	52.39	15.18	503	50.81	13.92	527

6.1 Overall tests of sex differences

In the first step of the analysis the third-order model (see section 5.3.1) was fitteu with the sample divided according to sex, and with every parameter constrained to be equal for the sexes. This model fitted very poorly (chi-square=760.5, df=375), which indicates the presence of sex differences. When the model was estimated separately within the groups of boys and girls, a good fit was obtained for the boys (chi-square=166.1, df=144, $p < .10$) but not for the girls (chi-square=218.6, df=144).

The model which was judged acceptable for the total sample thus fits better for boys than for girls. Some modifications were, therefore, made of the model in order to improve fit for the girls. The final model (chi-square=175.0, df=138, $p < .02$) includes 6 additional free parameters. Among these are loadings of Card Rotation II and the Swedish and Mathematics Achievement tests in the CFR-factor, which indicates that fcgirls the non-verbal reasoning factor is of greater scope than it is for boys.

This model fits well for the boys too (chi-square=158.9, df=138, p <.11) so it was used for all further investigations of sex differences. When the model was estimated simultaneously for boys and girls with all parameters constrained to be equal over the sexes a highly significant test-statistic was obtained (chi-square=747.05, df=369); when no constraints of equality were imposed the test-statistic obtained a much lower value (chi-square=338.2, df=278, p <.01). The difference between these two test-statist ; forms an overall test of the sex differences, and this test is very highly significant (chi-square=408.8, df=91).

It may thus be concluded that the overall test indicates sex differences with respect to the size of the estimated parameters in the HILI model. There are very many parameters in this model, however, and the overall difference may be accounted for by differences with respect to one, some or all of the parameters. In the following sections the sources of the significant overall statistic are localized.

6.2 Differences in means

The descriptive statistics presented in Table 23 indicate sex differences in level of performance, which is confirmed by a highly significant value of the test of the difference between the vectors of observed test means (chi-square=252.9, df=20).

The differences in test means cannot, however, be accounted for by a sex difference in the mean of the G-factor (chi-square=26.5, df=1), nor can they be accounted for by differences in the means of the second-order factors (chi-square=42.0, df=3). When the primary factors are allowed to take on different means for the sexes a substantial proportion of the difference in test means is, however, accounted for (chi-square=183.7, df=10). The fit of this model is still not entirely satisfactory and a major reason for this is that among the tests representing the I-factor, boys have a slightly higher mean on Number Series II, while girls have a higher observed mean on Letter Grouping II. This indicates that it is not meaningful to investigate a sex difference with respect to the mean of the I-factor, but that the specific test means must be concentrated upon.

The difference in means on Letter Grouping II is highly significant (chi-square=55.2, df=1), while the difference in means on Number Series II is not (chi-square=2.2, df=1). It may thus be concluded that there is a difference in favor of girls on Letter Grouping II, but that there is no overall sex difference in level of performance on the I-factor.

With this effect taken into account the model fits quite well, buι the fit still is significantly worse than the fit of the unconstrained model (chi-square=29.9, df=10). Further tests show this to be due to two other differences in specific test means: a difference in favor of girls on the Hidden Patterns test (chi-square=12.8, df=1) and a difference in favor of boys on part II of the Card Rotation test (chi-square=6.0, df=1). In the final test the difference between the completely unconstrained model and the partly constrained model proved to be non-significant (chi-square=11.1, df=8).

Table 24 presents the estimated means on the primary factors. The t-values, which are presented in Table 24 as well, show that for the primary factors Cf, CFR, V and Ve-Ach the girls have a significantly higher mean, while for Num-Ach the mean is significantly higher for the boys.

Table 24. Estimated means on the primary factors for girls with the means for boys taken to be zero.

Factor	Estimate	Standardized estimate	t-value
Vz	.21	.08	1.25
S	-.85	-.05	-.64
Cf	.24	.17	2.51
Cs	-.15	-.05	-.75
CFR	.67	.26	4.35
I	.00*	.00*	-
Ms	.18	.11	1.39
V	.48	.18	2.71
Ve-Ach	4.46	.35	5.62
Num-Ach	-2.04	-.15	-2.48

Note. * denotes a fixed parameter.

This pattern of sex differences conforms in most respects with previous findings, with the male superiority on numerical tests being as well established a fact as the female superiority in the verbal domain. It may be noted, however, that in the Gv domain no overall male superiority is found, as might be expected from much previous research. Instead the different primary factors in this domain exhibit a most varied pattern of sex differences: There is no difference on Vz or S, and a difference in favor of girls on Cf and CFR. In addition to these effects there is a difference in favor of boys in Part II of the Card Rotation test, and a difference in favor of girls on the Hidden Patterns test. It thus appears that on simpler items which require mental rotation the boys have an advantage, while the girls have an advantage on the more complex, analytic tasks.

While these results are inconsistent with the generalizations arrived at by Anastasi (1958) and Maccoby and Jacklin (1974) they conform quite well with results reported by Gustafsson (1976). In that study a sizeable difference in favor of boys was found for an S-test, while no difference was found with respect to the test Metal Folding, which here is taken to represent Vz. The absence of a difference in favor of boys on Vz does not appear, then, to be a chance finding restricted to the present study.

There are, in fact, indications that the traditionally found male superiority on spatial tests is now disappearing, at least for certain types of spatial tests. Evidence of this was presented by Härnqvist and Stahle (1977), who showed that for two representative samples of 13-year olds tested in 1961 and 1966 there was a diminishing sex difference on the Metal Folding test. Even more interesting, however, is their finding that the sex difference in performance was a function of the degree of equality of treatment of boys and girls in school, as codified in the curriculum.

The notion that the pattern of sex differences on spatial tests is related to the sex differentiation of society is also supported by cross-cultural research. Fairwheather (1976) reviewed cross-cultural studies of sex differences in field-independence (i.e. Cf) and concluded that the male superiority only appears in cultures which are stratified according to sex, while it is not found in cultures where females are encouraged to be more exploratory, e.g. among Eskimos.

With this line of reasoning it cannot, of course, be explained why the girls in the present study achieved a higher level of performance on Cf and CFR. The most likely explanation for this female superiority is, however, a general maturational lag for boys at 12 to 13 years of age.

The difference in favor of girls on the Hidden Patterns test is most likely due to there being a certain amount of variance from the factor Perceptual Speed (P) in this test (cf. Thurstone, 1944). A higher level of performance for girls on P-tests is a frequently reported finding (cf. Maccoby & Jacklin, 1974).

It also remains to be explained why the boys are superior on the second part of the S-test, when they are not on the Vz-factor. Gustafsson (1976) noted that for girls there were higher correlations between reasoning tests and Metal Folding than for boys, and argued that this may indicate that the girls successfully employ reasoning strategies in solving items in the Vz-type of tests. No such correlational differences were observed for the S-test and it was suggested that the speededness of this type of tests makes them less amenable to such non-spatial solution strategies. If this hypothesis is correct it suggests that the disappearing sex difference on the Vz-factor may at least partly be due to attitudinal factors: Girls may nowadays approach such test items in a more spirited manner, not being deterred by a strange-looking type of tasks, but they solve the items by relying on a partly different set of processes than boys do.

This hypothesis implies that the factorial structure might be different for boys and girls, with higher loadings of the complex, analytic Gv-tests in Gf- and Gc-factors for girls than for boys.

6.3 Differences in factorial structure

The overall test of the equality of the loadings of the tests in the primary factors is significant (chi-square=29.8, df=17, p <.01), but there is no significant sex difference with respect to the loadings of the first-order factors in the second-order factors or in the loadings of the second-order factors in the third-order factor (chi-square=16.0, df=10).

The loadings of the tests are thus not invariant over the sexes even though the chi-square statistic is not large in relation to the degrees of freedom. Further analyses showed the overall statistic to be accounted for by differences with respect to the loadings of Card Rotation II and the Swedish Achievement test in the CFR-factor (chi-square=5.3 and chi-square=7.4 with 1 df, respectively), the loadings being higher for the girls. The loading of the Mathematics Achievement test in the CFR-factor, which was introduced in the separate analysis for the girls, is not significantly different for boys and girls (chi-square=2.4, df=1), which is due to the fact that for boys too the loading is positive, even though it is not significant.

In the preceding section it was hypothesized that for girls the Vz- and Cf-tests would involve Gf and Gc variance. This hypothesis does not receive much support, however. To an extent the results even contradict the hypothesis since it is found that an S-test loads higher for girls than for boys in the CFR-factor. Looked at from a slightly different point of view, however, the results bring some support for the attempted line of reasoning, since the loading of Card Rotation II in CFR for the girls is accompanied by a lower level of performance, which indicates that the girls may have adopted a less successful reasoning strategy on part II of this test. The fact that for the girls there is a higher loading of the Swedish Achievement test in CFR also indicates that the girls may employ a verbal reasoning strategy on the Raven test.

6.4 Differences in variances

Differences in the variance/covariance of tests and factors is another possible source to account for the highly significant overall sex difference. An overall test of the difference in variance of the G-factor and the variances/covariances of the residuals of the factors of lower orders yields a significant test statistic (chi-square=58.9, df=19), so it is worthwhile to carry out more specific tests of this type of sex differences.

Such tests show that the variance of G does not differ for the sexes (chi-square=0.3, df=1), and nor do residual variances of the second-order factors (chi-square=1.7, df=2). However, for the residual variances of the primary factors a significant test-statistic is obtained (chi-square=36.2, df=10), indicating that one or more of these differ for boys and girls. Further tests show the difference to be accounted for by differences in the residual variances of S (higher for girls; chi-square=5.5, df=1), Vz (higher for boys; chi-square=15.5, df=1) and CFR (higher for boys; chi-square=4.9, df=1). Since the loadings of the primary factors in the second-order factors are the same for boys and girls these results reflect differences in the variance of the primary factors.

Some of the covariance terms also differ. Thus, the covariance between the residuals of I and Ve-Ach is higher for girls than for boys (chi-square=5.0, df=1), as is the covariance between the residuals of Cf and Ms (chi-square=7.0, df=1). Such differences in the size of the covariance of residuals are not easily interpreted, however.

Maccoby and Jacklin (1974) reviewed some studies of sex differences in variability of ability and concluded that there is "...some evidence for greater male variability in numerical and spatial abilities, but not consistently in verbal ability" (p. 118). The present results only partly conform to this generalization, but it is important to keep in mind that the Maccoby and Jacklin conclusion refers to the variability of observed scores, while our results pertain only to the variability of true scores.

The findings of greater male variability in Vz and CFR agree, however, with the conclusion that there is a tendency towards greater male variability in spatial ability. Maccoby and Jacklin interpreted this as being a consequence of an over-representation of males at the high end of the scale due to the male superiority in spatial ability, and also an over-representation of males at the low end of the scale, due to the greater vulnerability of males to anomalies of prenatal development, birth injury and childhood disease. While this interpretation seems reasonable it fails to account for the higher female variance in the S-factor. We will return to this question in the discussion in the next section.

Differences in the variability of observed scores may be due either to differences in the variability of true scores, or to differences in the variances of errors, or, of course, to different combinations of such differences. The overall test of the equality of the error variances of the tests proves significant (chi-square=39.2, df=25). Further tests show this overall effect to be accounted for by differences in the error variances for two of the tests: the Auditory Number Span test (chi-square=4.2, df=1) and the English Achievement test (chi-square=5.4, df=1), the error variance in both cases being higher for boys.

6.5 Discussion

The analyses have shown that there are in fact sex differences in most respects. However, it does appear that the most important differences are those obtained with respect to level of performance; these differences account for a very substantial proportion of the overall effect. It has already been concluded that the pattern of sex differences in level of performance rather closely matches the classical pattern of higher performance for girls in the verbal domain, and higher performance for boys in the numerical and spatial domains. It is interesting to note, however, that among the primary factors belonging with Gv boys have an advantage only on an S-test, while there is no difference on Vz and a higher performance for girls on Cf. It has already been surmised that this rather mixed pattern of sex difference may reflect a trend of disappearing sex differences in spatial ability, as a function of a lesser degree of sex differentiation in society.

It was hypothesized that the disappearing sex difference on the more complex, unspeeded, spatial tests may be due to the fact that girls tend to adopt non-spatial reasoning strategies in solving these items. It was also hypothesized that this would show as differences in the factor loadings for boys and girls. However, even though some such differences have been localized, the results do not unambiguously support the hypothesis. The findings that for girls Card Rotation II and the Mathematics and Swedish Achievement tests load CFR does indicate that this non-verbal reasoning factor is broader and of greater importance for girls than boys. However, loadings of Vz and Cf in CFR for the girls would have been more in line with the hypothesis. Further research will have to show whether there is any truth in the hypothesis of differences in factorial structure as a function of sex differences in strategies of solving spatial tests.

It is interesting to note that there are significant differences in the variances of the primary factors in the Gv domain. The girls have a higher variance in S, while the boys have a higher variance in Vz and CFR. There is thus a tenaency for the group with the lower mean to have a larger variance. This may indicate that the difference in mean performance is accounted for by rather small groups of persons at the extremes of the distributions, which in turn may be due to choice of either an appropriate or inappropriate strategy in solving the test items. Thus the differences in variance may be interpreted to provide indirect support for the hypothesis of strategy differences, even though it is difficult, for lack of suitable data, to pinpoint the exact nature of such differences.

In addition to these sex differences, the covariances between the residuals of some of the primary factors were also found to differ, as were some of the error variances of the tests. These findings seem, however, extremely difficult to interpret and until replicated they had better be left aside.

7 DISCUSSION AND CONCLUSIONS

It has already been concluded that the findings of the empirical study rather strongly support the hypothesized HILI model. So far, however, we have not discussed the interpretations and implications of the findings. This is done below.

7.1 Interpretations of the higher-order factors

Much of the factor analytic research on ability is characterized by rather superficial interpretations of the factors. This is partly due to the fact that the empirical results are obtained at a fairly high level of abstraction, so the factors are not easily identified with specific processes. In higher-order factor analyses the level of abstraction is carried one step further, which makes it even more difficult to interpret the higher-order factors in psychological terms. Some suggestions have been offered, however, for interpretations of these sources of individual differences, and these will be discussed next.

7.1.1 Interpretations of G

The practical utility of the G factor is evident from the wide-spread use of tests of general intelligence (see section 2.2). In such tests, it will be remembered, tasks are sampled from a broad range, which makes the G factor the dominant one, at the expense of more specific factors. These tests thus rely on the principle that Spearman called "the indifference of the indicator", i.e. given a sufficiently broad sampling of test items the G factor will be the most important source of individual differences on the test (see also Cronbach, 1951, and Härnqvist, 1977, for mathematical explanations of this fact.)

However, while the G factor has more than proven its utility in practical applications, it has bee remarkably absent from theoretical discussions since the pioneering work of Spearman. At best it might be said that attention has now and then been called to the fact that the G factor exists (e.g. McNemar, 1964; Humphreys, 1962). During the last few years, however, it appears that interest in the general factor has been revitalized, as may be seen in the writings of for example Humphreys (1979), Snow (1977, 1978, 1980, in press), Sternberg (1980) and Undheim (1981).

As has already been mentioned Undheim (1979, 1981) has proposed ideas almost identical with those espoused here, and from a series of studies it was concluded:

> Although Cattell´s hypothesis of two intelligence factors, fluid and crystallized intelligence, is seemingly supported by the simple-structure factor analytic distinctions of two such factors in several studies, hierarchical order analysis indicate that these findings may support an alternative hierarchical model of intelligence where fluid tasks are central to the definition of intelligence and group factors of crystallized ability or verbal-educational knowledge, visualization and speediness emerge. Thus the results are consistent with a more parsimonious neo-Spearman structuring of broad intelligence factors (Undheim, 1979, p. 11).

This conclusion thus conforms with the conclusion drawn from the present study.

In his interpretation of general intelligence Undheim (1980) stressed that G is a consequence of learning, and that the nature of intelligence is determined by cultural values: "... general reasoning *is* good reasoning with the contents of our culture" (Undheim, 1980, p. 12). This line of reasoning led Undheim to suggest a very broad definition of G, namely that it represents the entire repertoire of knowledge, skills and strategies.

From this definition of G also follows that "a measure of general intelligence should sample achievements in many subject matters -- some of which are tied to the academic curricula that subjects are exposed to, others tied to intellectual achievements acquired out of school -- and should include recent as well as not-so-recent acquisitions" (Undheim, 1980, p. 14).

It would seem, however, that the Undheim interpretation is fraught with problems. For one thing it is framed in very general terms, and it does not forward our understanding of general intelligence much. For another thing it appears that Undheim in his sampling model of intelligence disregards the most essential of his findings, namely that Gf coincides with G. Formulated in simple terms this result implies that a score obtained from the broadest possible and most representative sampling of tasks is virtually perfectly correlated with scores obtained on a small set of Gf tasks, such as a letter series test, a number series test, and a figural analogies test. The most interesting question must then be why the Gf tests have such power.

Reasoning tasks such as those in Gf tests have been very closely investigated in computer simulation studies, by, among others, Simon (1976). With respect to a sequence extrapolation task, highly similar to the items in the Number Series test, it was concluded that a program to solve such tasks must have the following capabilities:

1. Ability to detect relations of same, next and complement between pairs of symbols.

2. Familiarity with the symbols useu, and knowledge of their alphabets, stored in long-term memory.

3. Ability to hold and accumulate in relational structures newly-acquired information about the sequence, and finally to represent the pattern of the sequence in a relational structure, stored in memory.

4. Ability to keep one's place in a system of processes (a program), and to keep track in short-term memory of information needed as inputs to processes (Simon, 1976, p. 72).

Individual differences in all these capabilities could account for individual differences in performance on the task. It would seem, however, that the relation perceiving abilities described in 1 are fairly low-level abilities not likely to account for a substantial proportion of individual differences in performance, and the knowledge requirements specified in 2 are very limited (i.e. knowledge about the alphabet or the basics of the number system). To the extent that the list of capabilities is exhaustive it would, therefore, seem that the major sources of individual differences on such a task should derive from the ability to accumulate and retrieve newly-acquired information, and to administer and keep track of processes and the information they operate upon.

Simon (1976) also speculated about possible interpretations of Spearman's g. As one possibility it was suggested that g may reflect individual differences in the efficiency of a relatively small number of basic processes and structures which have been shown to be involved in the programs for a broad range of tasks. But Simon also pointed out that "it is not certain to what extent g is to be attributed to common processes among performance programs, or to what extent it derives from individual differences in the efficacy of the learning programs that assemble the performance programs" (Simon, 1976, p. 96). In terms of the computer analogy, then, individual differences in ability to perform Gf tasks would be more an effect of the efficiency of the applications programmer, the operating system and the data base mangagement system, than an effect of the speed and efficiency of specific instructions.

Snow (1977, 1980, in press) too has adopted a hierarchical view of the structure of ability, and has developed the Simon interpretation one step further. In the Snow model G is placed at the highest level with Gf, Gc and Gv as the most important abilities below G. While this model is partly based upon factor analytic findings, other techniques too have been relied upon. Thus Snow (1980) presented results from a multidimensional scaling of a large test battery (see also Marshalek, 1977). In this analysis the higher-order factors Gf, Gc and Gv were clearly discernable as clusters of tests. What is more interesting, however, is that a graph of the results clearly brought out the "centrality" of the clusters of tests. The central tests are those that correlate with a wide range of other tests, and it was found that the Gf tests, such as Raven and Letter Series, were the most central ones. These results thus strongly support the HILI model.

In an attempt to formulate a process theory of intelligence Snow has made a distinction between three kinds of processes: performance processes, control processes and assembly processes. Performance processes are psychological processes which perform a specific task, such as retreiving information from long-term memory. Control processes administer and keep track of the activities of the performance processes, and the assembly processes are those processes which with a particular goal in mind select and combine a certain set of performance processes. Snow pointed at the fact that most research on information processing has concentrated on the performance processes "so most cognitive theories look like performance programs. More recently, attention has been turning to the executive functions, but these are thought of mainly as control processes. The primary executive function, however, would appear to be assembly; the computer program analogy has far too long left out the programmer" (Snow, 1980).

Snow suggested that tests of general ability in particular may pose demands for new assembly of performance processes:

> Perhaps they represent to a greater degree the kinds of assembly and control processes needed to organize on a short term basis adaptive strategies for solving novel problems. The more complex and varied the sequence of novel problems, the more adaptive the processing needs to be. Raven Progressive Matrices is perhaps the archetypical example of such a test... (Snow, 1980)

According to this interpretation the most important features of Gf tests are that they present novel and complex tasks. Their novelty forces the examinee to find new ways of solving the tasks, and their complexity ensures that this is not a simple task: the examinee must always be prepared to find new modes of attack, and with a greater complexity follows that the number of steps and intermediary results to keep track of increases rapidly.

It is quite interesting to compare Snow's interpretation of G with Spearman's interpretation of g, which factors may be assumed to be identical. In Spearman's interpretation, it will be remembered, the concepts eduction of relations and correlates are important in the sense that tests which require such processes provide the best measures of g. The items in such tests tend to be quite complex, so Spearman and Snow seem to agree that complexity of the task is an important aspect in the measurement of general intelligence.

Spearman saw g as an expression of mental energy, while Snow sees it as being accounted for by the efficiency of assembly and control processes. At first sight these interpretations would appear to be quite incompatible, since they are framed in very different language. There is an important similarity, however, in the sense that both Snow and Spearman argue that individual differences in general ability are not due to individual differences in the efficiency of performance processes.

It would, in fact, appear possible to carry the comparison between the Snow and Spearman interpretations one step further. The primary basis for Spearman's formulation of g in terms of mental energy was the observation of "universal mental competition", i.e. that cognitive acts tend to interfere with each other. But the phenomenon of mental competition may just as easily be accounted for in information processing terms: the cognitive system is a system with a limited central resource which poses high demands on the efficiency with which processes are selected and combined when complex tasks are solved. It would thus seem that the Snow approach to understand general intelligence is able to account also for the phenomenon of mental competition.

Spearman's interpretation in terms of mental energy, and Snow's interpretation in terms of assembly and control processes are both formulated at a very high level of abstraction, and they both have more a character of being metaphors, than being full-fledged and worked out theories. It is quite natural, then, that the metaphors are heavily influenced by the ideas and the language of the time when they were formulated: physics and mechanics in the case of Spearman, and computers and computer programming in the case of Snow.

At the present state of knowledge it seems quite impossible to carry interpretations much further than Snow has done. What is important, however, is that the metaphors are well chosen, so that they may be developed into more specific formulations. In retrospect it seems clear that the Spearman metaphor in terms of energy represents a dead end, since it did not stimulate any further research or thinking. The Snow approach, may, however, be more profitable: it relates directly to flourishing research on information processing, computer simulation, and artificial intelligence, and it already operates with powerful concepts. An interpretation of general intelligence along the lines suggested by Simon and Snow seems, therefore, to provide a useful framework for further research.

7.1.2 Interpretations of Gc and Gv

The HILI model hypothesizes several second-order factors, but since the present study is restricted to Gc and Gv we will concentrate upon those.

It must be stressed that the factors labeled Gc and Gv in the HILI model are not directly comparable with the factors with same labels in the Cattell-Horn model. This is because in the HILI model there is no variance from the G factor at the second-order level, while in the Horn-Cattell model the G-variance is included in Gc and Gf. In order to separate clearly these two ways of representing general intelligence, we will refer to the residual variance in the factors after extracting G as Gc′ and Gv′ respectively, while the factors with G variance included are referred to as Gc and Gv. In the present data the residuals account for about 40 per cent of the variance in the second-order factors, while the remainder of the variance is due to G.

Horn and Cattell see both Gc and Gf as representing kinds of general intelligence, the major difference between the factors being that Gf is found in more or less culture-free tasks, while Gc is found in tasks with culture-bound content. Since our culture is strongly dominated by verbal content, Gc may even be regarded as a "blown up V" (Horn, 1976).

Horn and Cattell have not, however, gone very far towards an interpretation of Gv. In their 1966 paper they define Gv as being involved in "visualizing the movements and transformations of spatial patterns, maintaining orientation with respect to objects in space, unifying disparate elements and locating a given figuration in a visual field" (Horn and Cattell, 1966, p. 254). This, however, is just a list of the definitions of the primary factors Vz, S, Cs and Cf, and it does not specify what is common to these primary factors.

At the most superficial level, however, it seems clear that the common denominator of these factors is that they all deal with figural or spatial tasks. One way to account for Gc′ and Gv′ would then be to identify the processing requirements of verbal and figural information, respectively.

In an analysis of the correlational literature on spatial ability Lohman (1979) concluded:

> the crucial component of spatial thinking may be the ability to generate a mental image, perform various transformations on it, and remember the changes in the image as the transformations are performed. This ability to update the image may imply resistance to interference, both externally and internally generated. Further, it implies that one of the crucial features of individual differences in spatial ability may lie not in the vividness of the image, but in the control the imager can exercise over the image (Lohman, 1979, p. 116).

Lohman's interpretation thus stresses mental imagery in the processing of figural information.

Olson (1975), too, discussed the problem of what characterizes the processing of different types of information. He pointed out that

> what makes a problem spatial is not its surface properties but rather the structure of the symbol system or mental representation employed in obtaining a solution ... symbol systems are basically of two sorts: notational (linguistic) and non-notational (pictorial, spatial). The primary distinguishing feature of these systems is that the first is contrastive, often binary; the latter is continuous such that for any two positions there is a further point between them. (Olson, 1975, p. 76)

Olson also suggested that there may be individual differences both in the facility with which persons deal with the symbol systems, and in the amount and structure of knowledge being available in them.

It may be noted that the Olson distinction between notational and non-notational systems has its counterparts in other theoretical formulations. Thus, in the literature on brain laterality (e.g. Bock, 1973; Harris, 1975; Nebes, 1974; Wittrock, 1978) it is claimed that there are two broad modes of processing information, one associated with the left hemisphere, and the other with the right hemisphere. The left hemisphere processes information in a linear, sequential fashion and is specialized upon verbal information processing, while the right hemisphere processes information in a parallel, holistic, or synchronous fashion, and is specialized upon visual, figural information.

A very straightforward interpretation of the Gc′ and Gv′ factors could thus be that they reflect the efficiency of the left and right hemispheres, respectively (cf Bock, 1973). However, even though such a biologically based interpretation may be correct it does not contribute much psychological knowledge.

What seems important, however, is that there appears to be some consensus in the characterization of two fundamentally different modes of processing. One is described as analytic, linear, binary, serial, or successive, and the other is described as global, parallel, holistic, synchronous, simultaneous or continuou_. The first type of processing seems best suited for verbal content and for the rationalistic kind of thinking upon which premium is put in industrial societies, while the second type of processing is best suited for figural content and for visual imagery. One possibility, then, is that Gc´ and Gv´ express the facility with which these types of processing, respectively, are performed.

However, individual differences in performance need not only reflect the ability to perform certain types of processes but may also be due to differences in the organization, structure, and quantity of stored information of different types.

A concrete example of this type of individual differences is provided by perception in chess. It has been shown (e.g. Chase & Simon, 1973; Simon, 1976) that a chess master is very much superior a weak player in reconstructing from memory the position of the pieces in a game, but that there are no differences between strong and weak players in their ability to recall the positions of pieces placed at random on the board. Simon (1976) suggested that this is due to the chess master having a very large amount of standard configurations of pieces stored in a semantic memory. In reconstructing the positions of the pieces the chess master only has to remember a small set of standard configurations, from which the positions of the individual pieces are easily derived. Thus, "how well a player can reconstruct a briefly viewed position is mainly a function of how large a vocabulary of familiar configurations of pieces is stored in his long-term memory, and accessible through his discrimination net" (Simon, 1976, p. 94).

In a similar way it may be assumed that the structure and content of verbal semantic memory is of great importance in determining acquisition of new information. This suggests that in addition to proficiency in performing the verbal, sequential and the visual, imagery types of processes, individual differences in Gc´ and Gv´ may be due to differences in previously acquired knowledge.

Undheim (1979) argued that the verbal-educational group factor identified by him, which factor may be assumed to be identical with Gc´, represents a rather narrow achievement factor:

> it may be related to opportunity, interest and effort in verbal-educational achievement in school as well as out of school -- reflecting engaged time in school learning, in reading books

> more generally, reading newspapers and magazines, watching "educational" programs on TV, etc. (Undheim, 1979, p. 12).

Thus, Undheim sees Gc´ as being the accumulated result of choice of verbally oriented acitivities, and Gv´ may parallelly be viewed as the result of choice of spatially oriented activities. Such a theoretical position comes close to the "transfer" theory proposed by Ferguson (1954), and is supported for example by findings that choice of educational and occupational tracks does affect the relative stength of verbal and spatial abilities (e.g. Balke-Aurell, 1973, in press)

Snow (1980, in press) also sees Gc as being the result of prior learning and argued that it:

> represents the long term accumulation of knowledge and skills, organized into functional cognitive sytems by prior learning, that are in some sense crystallized as units for use in future learning. Since these are products of past education, and since education is in large part accumulative, transfer relations between past and future learning are assured. The transfer need not be primarily of specific knowledge but rather of organized academic learning skills. Thus Gc may represent prior assemblies of performance processes retrieved as a system and applied anew in instructional situations not unlike those experiences in the past ... (Snow, 1980).

A similar line of reasoning could easily be constructed to account for Gv.

It thus seems that there are two different explanations of individual differences in Gc´ and Gv´, one that takes its starting point in the different processing characteristics of verbal and figural information, and one that takes its starting point in differences in long-term memory as a consequence of prior learning. These interpretations are of course not mutually exclusive and they may both be true. There may also be quite intricate relationships between these mechanisms. Thus, small initial differences in proficiency in a certain type of processing may affect interests and preference which in turn may cause large differences in acquired knowledge. It is also conceivable that the availability of a large knowledge base enhances and expediates the type of processes which operate on that knowledge base.

Given all these possibilities it is not easy to make a choice between the interpretations. It would seem, however, that the interpretation in terms of knowledge may be the more potent one, for the reason that even if there are individual differences in power to perform the different types of processes, such differences may be expected to bring about differences in acquired knowledge.

7.1.3 Interpretations of primary factors

In previous research primary factors have been concentrated upon, at the expense of factors of higher orders. Still, however, interpretation has often been cursory and has most often consisted of an analysis of what is common to the tests that load a factor. Interpretations at this level thus result in descriptions of tasks characterizing the factor (see Table 1), and it does indeed seem quite difficult to carry interpretations of primary factors much further on the basis of factor analytic findings only.

In the HILI model the primary factors represent the variance which is left after the variance from the higher-order factors has been partialed out, and for many factors this is only a rather small fraction of the total variance. Furthermore, assuming that the variance of greatest psychological interest is represented by the G factor and the factors at the second-order level, the residual of the primary factors may be of limited interest. It does not seem worthwhile, therefore, to discuss in this context each of the primary factors.

7.2 The generality of the HILI model

In section 2.7.3 the HILI model was compared with other suggested models and it was concluded that the model is a very general model, which encompass most previously suggested models of the structure of human abilities. It is quite interesting, therefore, to compare the HILI model with yet another model, suggested by Sternberg (1980), in relation to which similar claims have been made.

The "componential theory of intelligence" proposed by Sternberg is based on the concept of component, which is defined as "... an elementary information process that operates upon internal representations of objects or symbols". (Sternberg, 1980, p. 6). On the basis of function, components are classified into five different kinds: meta-components, performance components, acquisition components, retention components, and transfer components. Meta-components "are higher order control processes that are used for executive planning and decision making in problem-solving" (p. 7), while performance components represent processes actually used in task performance.

Sternberg also classifies components according to level of generality into three categories: general components, class components and specific components. General components are processes used in all tasks within a given universe; class components are processes used within a sub-set of tasks; and specific components are used in the accomplishment of single tasks.

The classification of components according to generality is utilized in an assumed hierarchical organization of tasks. For each task in a hierarchy the same general components are used, and for each task different specific components are used. The level in the hierarchy at which a task is placed is determined by the class components: tasks at the lowest level each require one set of class components, while tasks at higher levels require all the class components of tasks at lower levels within the same branch of the hierarchy.

Sternberg confronted several of the factor analytic models of the organization of human abilities with this componential conception of task performance. With respect to the Spearman Two Factor theory it was argued that the g-factor comprises a set of general components that is common to a wide variety of intellectual tasks, while the s-factors correspond to spe.ific components. It was, furthermore, argued that the meta-components have a much higher proportion of general components among them, since for almost every task executive routines for planning and monitoring performance must be invoked. It was, thus, concluded that "individual differences in meta-componential functioning will be primarily responsible for the appearance of individual differences of a general nature" (p. 10).

The Thurstone PMA´s were by Sternberg interpreted to reflect individual differences in class components, while the correlation among the primary factors is accounted for by general components. As an example, Sternberg ment'oned the I-factor, which appears to involve a relatively small set of class components (i.e. infe..nce, mapping, application, and justification).

The concepts of fluid and crystallized intelligence within the Catell/Horn theory were also discussed. Tests of crystallized intelligence were interpreted to reflect "the products of acquisition, retention and transfer components, whereas fluid ability tests seem to involve the execution of performance components".

Sternberg thus argued that "... the componential theory can provide at least a tentative and sketchy account of how different forms of factor analysis and rotation can support different factorial theories of intelligence. On this view, each "theory" can be viewed as a special case, or subtheory, of a single theory." (p. 11).

Sternberg went on to argue that componential and factor theories of intelligence are differentially useful for different types of application:

For purposes of prediction the factorial type of approach is best suited; for purposes of diagnosis of individual strengths and weaknesses both factorial and componential approach may be useful; and for purposes of training the componential approach is best suited because as processes, the components may be directly trainable.

Sternberg concluded that "factor theories of intelligence are all right almost. What this means is that almost all factor theories of intelligence are right in the sense of being special cases of a more general psychometric theory, but that they are not quite all right when considered in isolation. They need to be supplemented by componential theories..." (p. 12).

While there is no need to challenge the conclusion that componential theories are complementary to factorially based models of the structure of abilities, it would not seem that the componential theory outlined by Sternberg is able to provide a psychometric super-theory, within which the different suggested models of the structure of abilities are contained as special cases. This can be seen if the specific interpretations proposed by Sternberg are scrutinized.

It is argued that the g-factor in Spearman's Two Factor model represents individual differences in metacomponents; that the Thurstonian I-factor represents individual differences in the performance components inference, mapping, application, and justification; and that Gf reflects individual differences in the execution of performance components generally. We have shown, however, that g is identical with Gf, and the empirical evidence also indicates that I is virtually identical with these higher-order factors. Sternberg thus proposes a set of three different explanations for the same individual difference variance. While these explanations are not mutually exclusive, this indicates that the componential theory is much to loose to function as a general psychometric theory.

Even more important, however, is the fact that while the factorial models identify and structure systematic sources of individual differences at different levels of generality, the componential theory models performance on intellectual tasks. This very fundamental difference between the factorial and componential approaches to the study of intelligence is seen if the content at the different levels of the two models is scrutinized. In the componential approach the hierarchy is a hierarchy of tasks, while in the factorial approach the hierarchy is a hierarchy of sources of individual difference variance. In fact, the componential approach could be perfectly valid even if there were no individual differences at all, while in such a case the factorial approach would break down because there is no covariance to analyze.

This difference in focus of attention makes the factorial and componential approaches complementary, but it also implies that the componential approach cannot provide a theory under which the factor-analytic models may be subsumed. Instead it seems that among the factor-analytic models the HILI model is the most general one. It also seems, however, that the HILI model is the factor-analytic model most clearly compatible with Sternberg's componentıal theory, and since these models are complementary they might profitably be used in conjunction.

7.3 Uses of the HILI model

The HILI model not only is a very general model, but the fact that it is formulated within the LISREL framework implies that it may be used as a so called measurement model in investigations of the relations between factors of ability and other variables. Such use of the HILI model offers several important advantages in comparison with traditional methods for analyzing relations betwen sets of variables:

- It makes it possible to study the relations between the factors and other variables, without the results being contaminated by error variance and specific variance in the psychological tests (e.g. Jöreskog, 1978; Gustafsson & Lindström, 1979).

- The fact that the model is hierarchical makes it possible to formulate extremely parsimonious models for the relations with other variables, by invoking first the G-factor, and then invoke only as many of the lower-order factors as may be necessary. It is not immediately obvious how such hierarchical measurement models may be formulated in LISREL (cf Bentler, 1980); Appendix 1 indicates, however, how this may be done.

- In most studies it will be impossible to include tests to represent all factors and all levels in the model. However, even in those cases when relatively few tests are used the hierarchical approach and mode of thinking may be utilized, and the factors may be interpreted within the framework of the HILI model. For example, if in a study interest is centered on the G factor a selection of three or four tests representing e.g. Gc, Gf and Gv may yield one common factor. This factor should come very close to the third-order G in the HILI model. The results in such a study may thus be compared to results obtained in another study with a much larger test battery, even though it will of course not be possible to separate error variance, test specificity, and the residuals of primary and second-order factors. The HILI model thus provides a framework for relating results obtained in studies in which different tests have been employed.

These advantages of the model might make it extremely useful in all branches of research on individual differences.

APPENDIX 1: The LISREL model

The abbreviation LISREL stands for linear structural relations and it is a model of high generality in which many other statistical models can be found as special cases (Jöreskog & Sörbom, 1978, pp. 2-3). LISREL was introduced by Jöreskog (1973, 1977), and a description of the model, and a computer program with the same name (LISREL IV) is given by Jöreskog and Sörbom (1978, cf. also Jöreskog & Sörbom, 1976, 1977). LISREL includes as special cases the methods for analysis of covariance structures developed by Jöreskog (1969, 1970, 1971, 1974). Here only a very sketchy description of LISREL can be given, and for a full account the reader should consult the references.

The LISREL model consists of two parts: the measurement models for the dependent and independent variables, in which latent variables (common factors) are defined in terms of observed variables, and the linear structural equation model, in which the relations between the latent variables are specified.

The measurement models are factor analysis models in which a smaller set of latent variables (factors) are supposed to account for the relations between the observed variables, and which are used to describe the measurement characteristics of the observed variables. There are two sets of observed variables $\underset{\sim}{y}'=(y_1,y_2,\ldots,y_p)$ and $\underset{\sim}{x}'=(x_1,x_2,\ldots,x_q)$, corresponding to dependent (outcome) and independent (aptitude) variables respectively and two sets of latent variables $\underset{\sim}{\eta}'=(\eta_1,\eta_2,\ldots,\eta_m)$ and $\underset{\sim}{\xi}'=(\xi_1,\xi_2,\ldots,\xi_n)$, corresponding to dependent and independent latent variables, respectively. There also are vectors specifying the unique parts (errors of measurement and specificity) of the $\underset{\sim}{y}$ and $\underset{\sim}{x}$ variables, $\underset{\sim}{\varepsilon}'=(\varepsilon_1,\varepsilon_2,\ldots,\varepsilon_p)$ and $\underset{\sim}{\delta}'=(\delta_1,\delta_2,\ldots,\delta_q)$.

The relations between the latent and the observed independent variables are specified in $\underset{\sim}{\Lambda}_x$, which is a factor loading matrix of order q x n for the regression of the $\underset{\sim}{x}$ variables on the $\underset{\sim}{\xi}$ variables, and the relations between the latent and the observed dependent variables are specified in the corresponding factor loading matrix $\underset{\sim}{\Lambda}_y$, of order p x m.

The measurement model for the $\underset{\sim}{x}$ variables is written:

(1) $\underset{\sim}{x} = \underset{\sim}{\Lambda}_x \underset{\sim}{\xi} + \underset{\sim}{\delta}$,

and for the $\underset{\sim}{y}$ variables it is written:

(2) $\underset{\sim}{y} = \underset{\sim}{\Lambda}_y \underset{\sim}{\eta} + \underset{\sim}{\varepsilon}$.

The structural equation model specifies the causal relationships among the latent variables and to represent these, two parameter matrices are used: $\underset{\sim}{\Gamma}$ which is a coefficient matrix of order m x n for the structural relations between the $\underset{\sim}{\xi}$ and the $\underset{\sim}{\eta}$ variables; and $\underset{\sim}{\beta}$ which is a coefficient matrix of order m x m for the structural relations among the $\underset{\sim}{\eta}$ variables. The residuals (disturbance terms or errors in equations) in the dependent variables are represented with the vector: $\underset{\sim}{\zeta}'=(\zeta_1,\zeta_2,\ldots,\zeta_m)$.

The system of linear structural relations has the form:

(3) $\underset{\sim}{\beta}\underset{\sim}{\eta} = \underset{\sim}{\Gamma}\underset{\sim}{\xi} + \underset{\sim}{\zeta}$.

The following covariance matrices must also be defined:

$\underset{\sim}{\theta}_\delta$ is a diagonal or symmetric matrix of order q x q containing the covariance matrix for the unique parts of the $\underset{\sim}{x}$ variables.

$\underset{\sim}{\Theta}_\varepsilon$ is a diagonal or symmetric matrix of order p x p containing the covariance matric for the unique parts of the $\underset{\sim}{y}$ variables.

$\underset{\sim}{\Phi}$ is a diagonal or symmetric matrix of order n x n containing the covariance matrix of the $\underset{\sim}{\xi}$ variables.

$\underset{\sim}{\Psi}$ is a diagonal or symmetric matric of order m x m for the covariance of the residuals.

Thus, in LISREL it is not necessarily assumed that the errors of measurement in the independent and dependent variables are uncorrelated with each other. It should also be pointed out that it is possible to specify LISREL models which allow estimation of covariances between errors of measurement in the independent and dependent variables; this can be effected through specifying a model in $\underset{\sim}{y}$ variables only. It is assumed, however, that the errors of measurement are uncorrelated with $\underset{\sim}{\xi}$,$\underset{\sim}{\eta}$ and $\underset{\sim}{\zeta}$.

It can be shown (cf. Jöreskog & Sörbom, 1978, p. 5) that if a set of observational data can be described with the equations (1), (2) and (3), and if the other assumptions are fulfilled, the covariance matrix Σ of order (p+q) x (p+q) of the observed dependent and independent variables is:

$$(4) \quad \Sigma = \begin{pmatrix} \Lambda_y(\beta^{-1}\Gamma\Phi\Gamma'\beta'^{-1}+\beta^{-1}\Psi\beta'^{-1})\Lambda_y+\Theta_\varepsilon & \Lambda_y\beta^{-1}\Gamma\Phi\Lambda_x' \\ \Lambda_x\Phi\Gamma'\beta'^{-1}\Lambda_y' & \Lambda_x\Phi\Lambda'+\Theta_\delta \end{pmatrix}$$

In specifying a LISREL model it is necessary to specify the nature of each element in the matrices $\Lambda_x, \Lambda_y, \Gamma, \beta, \Phi, \Psi, \Theta_\delta$ and Θ_ε (the elements will be referred to with small Greek letters). The elements can be of three different kinds: a fixed parameter, i.e. the parameter is assigned a given value; a free parameter, i.e. the parameter is to be estimated; and a constrained parameter, i.e. the parameter is to be estimated but it is constrained to be equal to one or more other parameters.

From the relations (1) - (3) it would seem that LISREL is subject to a major limitation -- the means of the latent or the observed variables are not included in the model. In terms of regression analysis this would correspond to regression models without the intercept parameter, and to make a complete evaluation of effects it is necessary to include the intercept as well. Sörbom (1974, 1976, 1978) has formulated models which do allow hypotheses on the means, and which come very close to the LISREL model. As has been shown by Sörbom (1979) these models can in fact be estimated with the LISREL program, using a special specification, so in reality LISREL does allow estimation and testing of the intercept parameter. A new version of the LISREL program (LISREL V) is now being released in which this capability is obtained automatically, so there is no reason to describe how it can be done with LISREL IV (see, however, Gustafsson & Lindström, 1979).

So far LISREL has been presented as if there was one group (population) of persons only, but often there are two or more groups of persons. LISREL handles any number of groups, however, and the presentation given above is easily generalized, through adding a superscript (i, i=1,...,g) indicating to which of g groups a parameter or a matrix of parameter refers. Thus, for example, the matrix of coefficients of structural relations between independent and dependent latent variables in the ith group is referred to as $\Gamma^{(i)}$. If a parameter or a matrix of parameters is constrained to be equal in all groups an asterisk (*) is used to denote that, i.e. $\underset{\sim}{\Gamma}^{(*)}$.

The values of the non-fixed parameters in the LISREL model can be estimated from the sample covariance matrices. However, to obtain any estimates it is necessary that the model is identified. The problem of identifiability can be defined in the following way:

> Identifiability depends on the choice of the model and on the specification of fixed, constrained, and free parameters. Under a given specification, a given structure $\underset{\sim}{\Lambda}_y, \underset{\sim}{\Lambda}_x, \underset{\sim}{\beta}, \underset{\sim}{\Gamma}, \underset{\sim}{\Phi}, \underset{\sim}{\Psi}, \underset{\sim}{\Theta}_\varepsilon, \underset{\sim}{\Theta}_\delta$ generates one and only one $\underset{\sim}{\Sigma}$ but there may be several structures generating the same $\underset{\sim}{\Sigma}$. If two or more structures generate the same Σ, the structures are said to be equivalent. If a parameter has the same value in all equivalent structures, the parameter is said to be identified. If all parameters of the model are identified, the whole model is said to be identified. (Jöreskog & Sörbom, 1978, p. 9).

For some special cases there are general rules for determining whether a specific model is identified or not (e.g. Werts, Jöreskog & Linn, 1973; Wiley, 1973) but in most instances that is not the case. The LISREL IV program has, however, the capability of detecting if a model is not identified (cf. Jöreskog & Sörbom, 1978, pp. 10-11).

In an identified model the values of the non-fixed parameters can be estimated with maximum likelihood methods. It is assumed that the distribution of the observed variables is sufficiently well described by the moments of the first and second orders, which in particular holds true when the observed variables have a multinormal distribution.

Each analysis of a fully identified model not only yields estimates of parameters but also an overall chi-square test of the goodness of fit of the model, along with standard errors of the estimated parameters. As a help in modification of a poorly fitting model, the first order derivatives with respect to the fixed parameters are also computed (cf. Sörbom, 1975).

Through computing the differences between the values of the test statistics obtained with more and less constrained models, i.e. models differing as to the number of parameters estimated, it is also possible to test the significance of subsets of parameters. Consider the following concrete example: A model is estimated for two groups in which $\underset{\sim}{\Gamma}^{(1)}$ and $\underset{\sim}{\Gamma}^{(2)}$ are not constrained to be equal. The test of fit gives χ^2_1 with df_1 degrees of freedom. Then a model is specified in which $\underset{\sim}{\Gamma}^{(*)}$ is estimated instead, which will have χ^2_2 with df_2 degrees of freedom. The test statistic $\chi^2_2 - \chi^2_1$ then is chi-square distributed with $df_2 - df_1$ degrees of freedom, and the test is, of course, a test of the equality of the coefficients of structural relations within the groups. In the same way other parameter matrices, or subsets of parameter matrices, can be tested.

Estimating hierarchical models

The specification of LISREL models is most easily shown in graphical presentations. Figure 12 shows the models used in section 2.6 to illustrate a simple measurement model (cf Figure 3). There are 6 observed y-variables (enclosed in squares), and 2 η-variables (enclosed in circles). The η-variables affect the y-variables, which is shown by the straight arrows from η to y. In addition, the y-variables are affected by error (ε). A correlation is assumed between the two η-variables, which is indicated by the curved bidirectional arrow between the two latent variables.

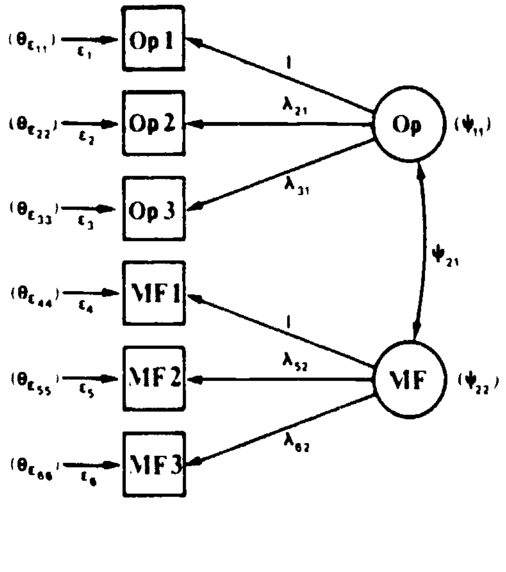
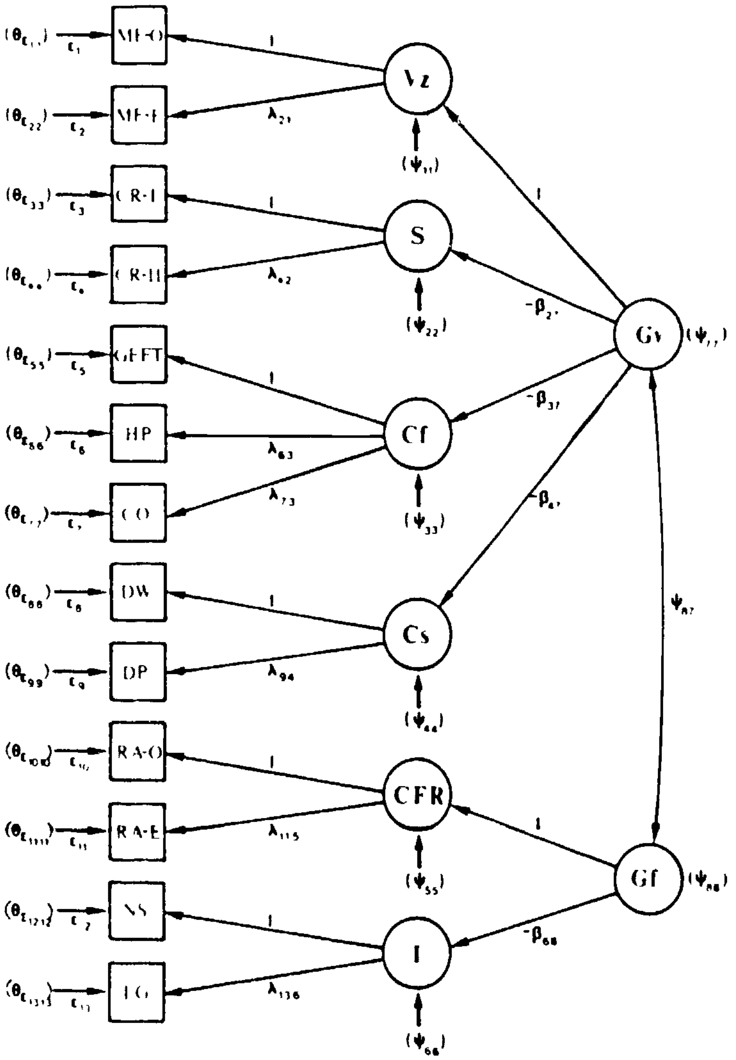


Figure 12. An example of a simple measurement model.

An example of a higher-order measurement model is shown in Figure 13 (cf Figure 7). In this model one set of latent variables (the second-order factors) affect another set of latent variables (the first-order factors). In addition each first-order factor is affected by a disturbance (specific) factor (ζ). A correlation is also hypothesized between the second-order factors.



$(\theta_{\epsilon_{11}})$ ϵ_1 MF-O
$(\theta_{\epsilon_{22}})$ ϵ_2 MF-F
$(\theta_{\epsilon_{33}})$ ϵ_3 CR-I
$(\theta_{\epsilon_{44}})$ ϵ_4 CR-II
$(\theta_{\epsilon_{55}})$ ϵ_5 GEFT
$(\theta_{\epsilon_{66}})$ ϵ_6 HP
$(\theta_{\epsilon_{77}})$ ϵ_7 CO
$(\theta_{\epsilon_{88}})$ ϵ_8 DW
$(\theta_{\epsilon_{99}})$ ϵ_9 DP
$(\theta_{\epsilon_{1010}})$ ϵ_{10} RA-O
$(\theta_{\epsilon_{1111}})$ ϵ_{11} RA-E
$(\theta_{\epsilon_{1212}})$ ϵ_{12} VS
$(\theta_{\epsilon_{1313}})$ ϵ_{13} FG
1
λ_{21}
λ_{42}
λ_{63}
λ_{73}
λ_{94}
λ_{115}
λ_{136}
Vz
S
Cf
Cs
CFR
I
(ψ_{11})
(ψ_{22})
(ψ_{33})
(ψ_{44})
(ψ_{55})
(ψ_{66})
$-\beta_{27}$
$-\beta_{37}$
$-\beta_{47}$
$-\beta_{68}$
Gv (ψ_{77})
Gf (ψ_{88})
ψ_{87}

Figure 13. An example of a hierarchical model.

In the LISREL-terminology the measurement model specifies the relations between observed variables and latent variables, while the structural model specifies the relations between the latent variables. According to this terminology the model shown in Figure 13 contains both a measurement model and a structural model. However, hierarchical models of the structure of ability are extremely useful to provide a set of predictors, from which other latent variables are predicted. A hierarchical model may, therefore, be viewed as a measurement model in itself.

It is not immediately obvious how a hierarchical model may be used as a measurement model, since this requires that the disturbances of lower-order factors are used as independent variables. The disturbances may, however, be expressed as latent variables, which may then be used as predictors (or dependent variables).

An example of this kind of model specification is shown in Figure 14. In this model, which was used in section 5.1 to investigate the Raven test (cf Figure 9), hierarchical measurement models are used both for the independent variables and for the dependent variables. As may be seen from the Figure the disturbances of the first-order factors are turned into latent variables by being specified as orthogonal factors having a relation of unity with the first-order factor.

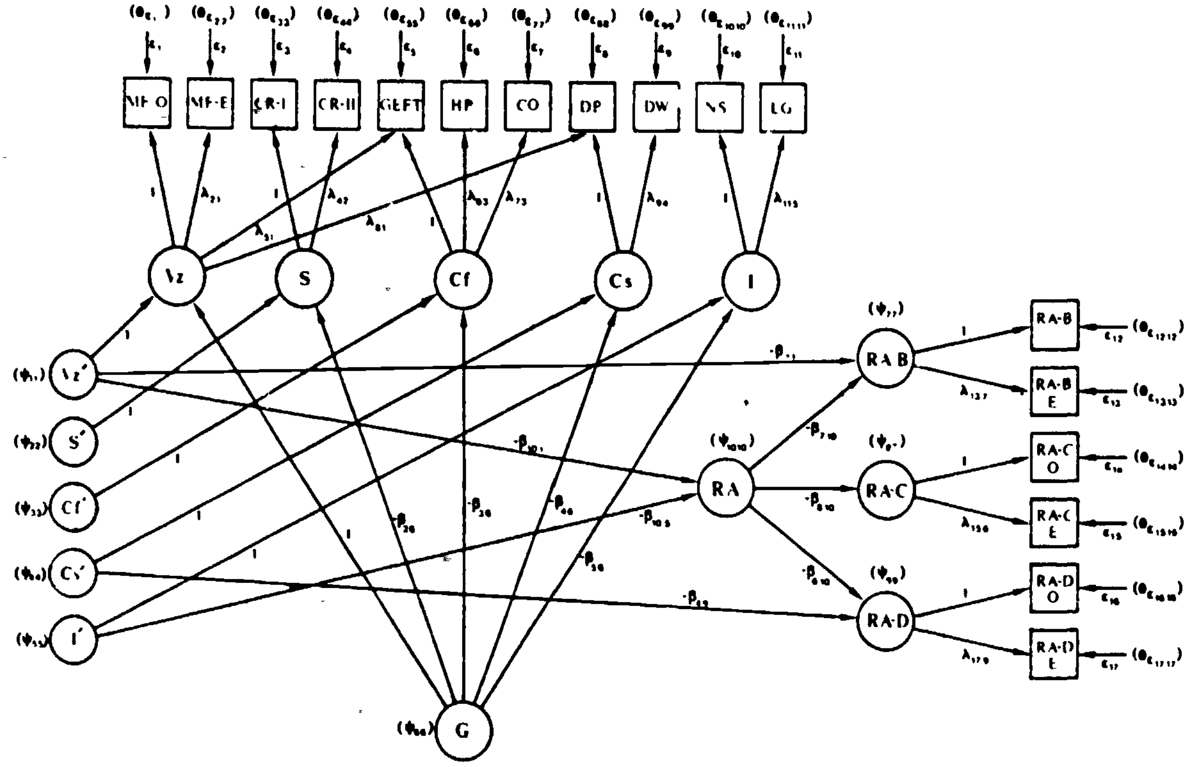


Figure 14. An example of higher-order measurement models.

APPENDIX 2: The Rasch model

The Rasch model is a latent trait model in which the probability of a correct answer to an item is expressed as a function of two parameters. One parameter describes the difficulty of the item (σ_i,i=1,...,k) and one parameter describes the ability of the tested subject (ξ_v,v=1,...,n). Denoting a correct answer to item i by subject v as $A_{vi}=1$ we have

$$P(A_{vi}=1) = \frac{\exp(\xi_v-\sigma_i)}{1+\exp(\xi_v-\sigma_i)}$$

There are three basic assumptions in the Rasch model: the assumption of unidimensionality, the assumption of local statistical indenendence, and the assumption of equal item discrimination. The assumption of unidimensionality is the most important, even though it has been argued that it is impossible or very difficult to uphold a distinction between these assumptions (Gustafsson, 1980b). The meaning of the assumption of unidimensionality is that there is only one latent trait affecting performance on all items in the test.

The probability of a correct answer to an item (i) can be expressed as a function of the ability variable (ξ). This function is called the item characteristic curve (ICC) and can be expressed in the following way

$$f_i(\xi) = \frac{\exp(\xi-\sigma_i)}{1+\exp(\xi-\sigma_i)}$$

In a similar way may for each person the probability of a correct answer be shown as a function of item difficulty, to form a person characteristic curve (PCC).

There are several different goodness-of-fit tests that may may be used to evaluate the fit of data to the model. The underlying rationale is expressed by Gustafsson (1980b):

> It does seem that the Rasch model can be violated in basically two ways: either a model is needed to describe the data which contains two or more parameters for each person, which would be a violation of the assumption of unidimensionality; or a model is needed which contains two or more parameters for each item, which would be a violation of the assumption of the form of the ICCs; or, of course, a combination of these. If the Rasch model holds true for a set of data the item parameters are invariant from one group of persons to another and the person parameters are invariant from one group of items to another (p. 209).

There are two types of test that can be used. One type is called ICC tests, since they are sensitive to violations of the assumption of equal ICC's. The other type is called PCC tests, since they are sensitive to differences in the PCC's.

The computer program PML (Gustafsson, 1979) has been used to estimate the parameters in the model and to test fit. This program offers three different goodness-of-fit tests and one graphical test of deviations from the model assumptions:

a. The Anderson conditional likelihood ratio test (Anderson, 1973). This test is a test of the hypothesis that item parameters are invariant over different groups of persons. When the grouping is done according to level of performance (high-low) on the total set of items the test is called the A-ICCSL test, since it is sensitive to differences in the slopes of the ICC's. When the grouping is done according to other criteria such as, for example, age or sex, it is called the A-ICC test. With this kind of grouping it is a test of unidimensionality.

b. The Martin-Löf chi-square test (Martin-Löf, 1973). This test too is sensitive to heterogenous slopes of the ICC's and is referred to as the ML-ICCSL test. It is calculated by forming a chi-square sum of the deviations between observed and predicted frequencies of correct answers for each score group. Even though the ML-ICCSL and the A-ICCSL tests are sensitive to the same type of deviations, there are differences between them when used on smaller samples.

c. The Martin-Löf test of homogeneity of two sets of items. This is a test of the hypothesis that two sets of items measure the same ability. With items grouped according to level of difficulty it is sensitive to differences of the PCC slopes and, accordingly, it is called the ML-PCCSL test. With items grouped according to other criteria, which should be theoretically derived, it is sensitive to deviations from unidimensionality and is called the ML-PCC test.

d. The graphical test of item fit is sensitive to variations in the slopes of the ICC's. It is simply a plot for each item of observed proportion of correct answers for each scoregroup against the corresponding predicted proportions.

APPENDIX 3: Descriptive data

Table 25. Correlations between the tests in the reference battery (N=981).

	MF-O	MF-E	CR-1	CR-2	GEFT	HP	CO	DW	DP	RA-O	RA-E	NS	LG
MF-O	1.00												
MF-E	.79	1.00											
CR-1	.33	.34	1.00										
CR-2	.40	.41	.74	1.00									
GEFT	.47	.49	.33	.36	1.00								
HP	.42	.43	.44	.45	.48	1.00							
CO	.43	.43	.45	.47	.50	.56	1.00						
DW	.18	.21	.14	.18	.25	.27	.23	1.00					
DP	.26	.24	.14	.18	.22	.20	.20	.38	1.00				
RA-O	.38	.39	.25	.31	.32	.34	.28	.18	.21	1.00			
RA-E	.40	.41	.27	.33	.35	.38	.31	.18	.21	.81	1.00		
NS	.35	.38	.32	.38	.40	.40	.40	.25	.13	.40	.43	1.00	
LG	.35	.34	.35	.38	.41	.41	.39	.22	.15	.40	.43	.52	1.00
ANS	.14	.12	.13	.14	.14	.17	.17	.15	.10	.12	.12	.22	.20
ALS	.14	.13	.19	.15	.17	.21	.19	.16	.08	.17	.17	.27	.25
Op-O	.32	.29	.20	.23	.31	.31	.29	.17	.13	.30	.30	.42	.39
Op-E	.26	.27	.25	.25	.31	.27	.26	.17	.13	.30	.30	.42	.38
Sw	.33	.33	.27	.31	.39	.38	.36	.27	.15	.40	.42	.56	.51
Ma	.38	.42	.34	.40	.43	.41	.41	.21	.11	.41	.43	.68	.49
Eng	.33	.29	.24	.26	.41	.37	.36	.22	.18	.36	.38	.49	.44

	ANS	ALS	Op-O	Op-E	Sw	Ma	Eng
ANS	1.00						
ALS	.49	1.00					
Op-O	.16	.18	1.00				
Op-E	.14	.19	.70	1.00			
Sw	.21	.32	.68	.65	1.00		
Ma	.18	.23	.50	.50	.66	1.00	
Eng	.20	.29	.56	.56	.78	.61	1.00

Table 26. Means and standard deviations on the tests in the reference battery for the groups with complete data (n=981) and incomplete data (n=243).

	Complete data		Incomplete data		
	$\bar{x}$	s.d.	$\bar{x}$	s.d.	
MF-O	9.51	3.10	8.93	3.46	
MF-E	9.19	3.25	8.90	3.52	
CR-1	52.84	21.72	50.77	23.23	
CR-2	44.53	18.25	41.68	19.89	
GEFT	3.93	2.46	3.69	2.57	
HP	71.08	23.73	67.75	24.91	
CO	22.67	8.12	21.44	8.42	
DW	11.89	3.36	11.03	3.56	
DP	12.79	3.28	12.23	3.09	
Ra-O	12.79	2.85	12.24	3.15	
Ra-E	11.65	3.27	10.81	3.48	
NS	8.04	3.78	6.93	3.52	
LG	11.18	3.56	10.10	3.69	
ANS	4.49	2.68	4.11	2.64	
ALS	4.57	2.20	4.33	2.04	
Op-O	11.36	3.20	10.71	3.21	
Op-E	10.72	2.98	10.01	3.13	
Sw-Ach	69.28	17.02	60.38	20.39	(n=118)
Ma-Ach	51.94	14.41	44.43	15.97	(n=49)
Eng-Ach	101.05	24.29	94.99	30.08	(n=77)

APPENDIX 4: List of abbreviations

Abbreviations of factors

Abbr	Factor
Cf	Flexibility of Closure
CFC	Cognition of Figural Classes
CFR	Cognition of Figural Relations
CMC	Cognition of Semantic Classes
CMR	Cognition of Semantic Relations
Cs	Speed of Closure
F	General Fluency
Fa	Associational Fluency
Fi	Ideational Fluency
Fw	Words Fluency
G	General Intelligence
Gc	Crystallized Intelligence
Gf	Fluid Intelligence
Gv	General Visualization
Gr	General Fluency
Gs	General Speediness
I	Induction
k:m	Spatial-practical-mechanical
Ma	Associative Memory
Mk	Mechanical Knowledge
Ms	Memory Span
N	Numerical facility
P	Perceptual Speed
R	General Reasoning
Rs	Syllogistic Reasoning
S	Spatial Orientation
v:ed	verbal-educational
Vz	Visualization

Abbreviations of tests

Abbr	Test
ALS	Auditory Letter Span
ANS	Auditory Number Span
CO	Copying
CR	Card Rotation
DP	Disguised Pictures
DW	Disguised Words
GEFT	Group Embedded Figures Test
HP	Hidden Patterns
LG	Letter Grouping
MF	Metal Folding
NS	Number Series
Op	Opposites
RA	Raven Progressive Matrices

REFERENCES

Anastasi, A. Differential psychology (third ed.). New York, MacMillan, 1958.

Andersen, E.B. A goodness of fit test for the Rasch model. Psychometrika, 1973, 38, 123-140.

Balke-Aurell, G. Förändringar i begåvningsinriktning i relation till utbildning och yrkesverksamhet. (Changes in ability structure in relation to education and occupation.) Licentiatavhandling från Pedagogiska institutionen, Göteborgs universitet, 1973.

Balke-Aurell, G. Changes in ability factors as related to educational and occupational experience. Göteborg: Acta Universitatis Gothoburgensis, in press.

Bentler, P. Multivariate analysis with latent variables: causal modeling. Annual Review of Psychology, 1980, 31, 419-456.

Binet, A., & Henri, V. La psychologie individuelle. L'Année Psychologique, 1895, 2, 411-463.

Bock, R.D. Word and image: sources of the verbal and spatial factors in mental test scores. Psychometrika, 1973, 38, 437-457.

Botzum, W.A. A factorial study of the reasoning and closure factors. Psychometrika, 1951, 16, 361-386.

Burt, C.L. The factors of the mind: An introduction to factor analysis in psychology. New York: MacMillan, 1941.

Burt, C.L. The structure of the mind: A review of the results of factor analysis. British Journal of Educational Psychology, 1949, 19, 100-111; 176-199.

Cattell, R.B. A culture-free intelligence test, I. Journal of Educational Psychology, 1940, 31, 161-179.

Cattell, R.B. Some theoretical issues in adult intelligence testing. Psychological Bulletin, 1941, 38, 592.

Cattell, R.B. The measurement of adult intelligence. Psychological Bulletin, 1943, 40, 153-193.

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

Chase, W.G., & Simon, H.A. The mind's eye in chess. In W.G. Chase (Ed.) Visual information processing. New York: Academic Press, 1973.

Cronbach, L.J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 297-334.

Cronbach, L.J. The two disciplines of scientific psychology. American Psychologist, 1957, 12, 671-684.

Cronbach, L.J. Essentials of psychological testing (2nd ed.). New York: Harper and Row, 1960.

Cronbach, L.J., & Snow, R.E. Aptitudes and instructional methods. A handbook for research on interactions. New York: Irvington, 1977.

Das, J.P., Kirby, J.R., & Jarman, R.F. Simultaneous and successive cognitive processes. New York: Academic Press, 1979.

Fairweather, H. Sex differences in cognition. Cognition, 1976, 4, 231-280.

French, J.W. The description of aptitude and achievement tests in terms of rotated factors. Psychometric Monographs, 1951, No. 5.

French, J.W., Ekstrom, R.B., & Price, L.A. Manual for kit of reference tests for cognitive factors. Educational Testing Service, Princeton, N.J., 1963.

Galton, F. Hereditary genius: an inquiry into its laws and consequences. London: MacMillan, 1869.

Galton, F. Inquiries into human faculty and its development. London: MacMillan, 1883.

Glaser, R. Individuals and learning. The new aptitudes. Educational Researcher, 1972, 1 (6), 5-13.

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

Guilford, J.P. Thurstone's primary mental abilities and structure-of-intellect abilities. Psychological Bulletin, 1972, 77, 129-143.

Guilford, J.P. Fluid and crystallized intelligence: Two fanciful concepts. Psychological Bulletin, 1980, 88, 406-412.

Guilford, J.P., & Hoepfner, R. The analysis of intelligence. New York: McGraw-Hill, 1971.

Gustafsson, J.-E. Interaktion mellan individ- och undervisningsvariabler. Introduktion och litteraturgenomgång. (Interaction between aptitudes- and instructional variables. Introduction and review of literature). Rapporter från Pedagogiska institutionen, Göteborgs universitet, nr 63, 1971.

Gustafsson, J.-E. Verbal and figural aptitudes in relation to instructional methods. Studies in aptitude-treatment interactions. Göteborg: Acta Universitatis Gothoburgensis, 1976.

Gustafsson, J.-E. The Rasch model for dichotomous items: Theory, applications and a computer program. Reports from the Institute of Education, University of Göteborg, no. 63, 1977.

Gustafsson, J.-E. PML: A computer program for conditional estimation and testing in the Rasch model for dichotomous items. Reports from the Institute of Education, University of Göteborg, 1979.

Gustafsson, J.-E. Testing hierarchical models of ability organization through covariance models. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980a.

Gustafsson, J.-E. Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 1980b, 33, 205-233.

Gustafsson, J.-E. Matching aptitudes and treatments: The ATI-paradigm in research on individualization of instruction. In N. Sövik, H.M. Eikeland & A. Lysne (Eds.) On individualized instruction. Oslo: Universitetsförlaget, 1981.

Gustafsson, J.-E. & Lindström, B. Analyzing ATI data by structural analysis of covariance matrices. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Harman, H. Modern factor analysis. Chicago: University of Chicago Press, 1967.

Härnqvist. K. Manual till DBA. (Manual for the DBA battery.) Stockholm: Skandinaviska Testförlaget AB, 1968.

Härnqvist, K. A note on the correlations between increments, cumulated attainment and a predictor. Reports from the Institute of Education, University of Göteborg, no. 62, 1977.

Härnqvist, K., & Stahle, G. An ecological analysis of test score changes over time. Reports from the Institute of Education, University of Göteborg, no. 64, 1977.

Harris, L.J. Neuropsychological factors in the development of spatial skills. In J. Eliot & N.J. Salkind (Eds.) Children's spatial development. Springfield, Illinois: Charles C Thomas, 1975.

Hendrickson, A.E., & White, P.O. Promax: A quich method for rotation to oblique simple structure. British Journal of Statistical Psychology, 1964, 17, 65-70.

Horn, J.L. Fluid and crystallized intelligence: a factor analytic study of the structure among primary abilities. Unpublished doctoral dissertation, University of Illinois, 1965.

Horn, J.L. Organization of abilities and the development of intelligence. Psychological Review, 1968, 79, 242-259.

Horn, J.L. State, trait and change dimensions of intelligence. British Journal of Educational Psychology, 1972, 42, 159-185.

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

Horn, J.L. Personality and ability theory. In R.B. Cattell & R.M. Dreger (Eds.) Handbook of personality theory. Washington: Hemisphere, 1977.

Horn, J.L., & Bramble, W.J. Second-order ability structure revealed in rights and wrongs scores. Journal of Educational Psychology, 1967, 58, 115-122.

Horn, J.L. & Cattell, R.B. Refinement and test of the theory of fluid and crystallized intelligence. Journal of Educational Psychology, 1966, 57, 253-270.

Humphreys, L.G. The organization of human abilities. American Psychologist, 1962, 17, 475-483.

Humphreys, L.G. Critique of "Theory of fluid and crystallized intelligence: A critical experiment." Journal of Educational Psychology, 1967, 58, 129-136.

Humphreys, L.G. The construct of general intelligence. Intelligence, 1979, 3, 105-120.

Jöreskog, K.G. A general approach to confirmatory maximum likelihood factor analysis. Psychometrika, 1969, 34, 183-202.

Jöreskog, K.G. A general method for analysis of covariance structures. Biometrika, 1970, 57, 239-251.

Jöreskog, K.G. Simultaneous factor analysis in several populations. Psychometrika, 1971, 36, 409-426.

Jöreskog, K.G. A general method for estimating a linear structural equation system. In A.S. Goldberger & O.D. Duncan (Eds.) Structural equation models in the social sciences. New York: Seminar Press, 1973.

Jöreskog, K.G. Analysing psychological data by structural analysis of covariance matrices. In R.C. Atkinson, D.H. Krantz & R.D. Suppes (Eds.) Contemporary developments in mathematical psychology, Volume II. San Francisco: W.H. Freeman & Co., 1974.

Jöreskog, K.G. Structural equation models in the social sciences: Specification, estimation and testing. In P.R. Krishnaiah (Ed.) Application of statistics. North Holland Publishing Co., 1977.

Jöreskog, K.G. Structural analysis of covariance and correlation matrices. Psychometrika, 1978, 43, 443-477.

Jöreskog, K.G., & Sörbom, D. Statistical models and methods for test-retest situations. In D.N.M. de Gruitjer, L.J.Th. van der Kamp & H.F. Crombag (Eds.) Advances in psychological and educational measurement. London, 1976.

Jöreskog, K.G., & Sörbom, D. Statistical models and methods for analysis of longitudinal data. In D.J. Aigner & A.S. Goldberger (Eds.) Latent variables in socioeconomic models. Amsterdam: North-Holland Publishing Co., 1977.

Jöreskog, K.G., & Sörbom, D. LISREL IV. A general computer program for estimation of linear structural equation systems by maximum likelihood methods. Universi-y of Uppsala, Department of Statistics, 1978.

Kelley, T.L. Crossroads in the mind of man. Stanford, Calif.: Stanford University Press, 1928.

Ljung, B.-O. The adolescent spurt in mental growth. Stockholm: Almqvist & Wiksell, 1965.

Lohman, D. The relationship between hypnotizability and speed of closure. Technical report no. 6, Aptitude Research Project, School of Education, Stanford University, 1977.

Lohman, D. Spatial ability - A review and reanalysis of the correlational literature. Technical Report no. 8, Aptitide Research Project, School of Education, Stanford University, 1979.

Maccoby, E.E. Sex differences in intellectual functioning. In E.E. Maccoby (Ed.) The development of sex differences. Stanford: Stanford University Press, 1966.

Maccoby, E.E., & Jacklin, C.N. The psychology of sex differences. Stanford: Stanford University Press, 1974.

Marshalek, B. The complexity dimension in the radex and hierarchical models of intelligence. Paper presented at American Psychological Association Convention, San Francisco.

Martin-Löf, P. Statistiska modeller. Anteckningar från seminarier läsåret 1969-70 utarbetade av Rolf Sundberg. 2:a uppl. (Statistical models. Notes from seminars 1969-70 by Rolf Sundberg. 2nd ed.). Institutet för försäkringsmatematik och matematik vid Stockholms universitet, 1973.

McNemar, Q. Lost: our intelligence? Why? American Psychologist, 1964, 19, 871-882.

Nebes, R.D. Hemispheric specialization in commisurotimized man. Psychological Bulletin, 1974, 81, 1-14.

Olson, D.R. On the relations between spatial and linguistic processes. In J. Eliot & N.J. Salkind (Eds.) Children's spatial development. Springfield, Illinois: Charles C Thomas, 1975.

Pawlik, K. Concepts and calculations in human cognitive abilities. In R.B. Cattell (Ed.) Handbook of multivariate experimental psychology. Chicago: Rand McNally, 1966.

Raven, J.K.C. Progessive Matrices. Sets A, B, C, D & E. London: H.K. Lewis & Co., 1938.

Raven, J.K.C. Guide to the Standard Progressive Matrices. London: H.K. Lewis & Co., 1960.

Resnick, L.B. (Ed.) The nature of intelligence. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1976.

Simon, H.A. Identifying basic abilities underlying intelligent performance of complex tasks. In L.B. Resnick (Ed.) The nature of intelligence. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1976.

Snow, R.E. Research on aptitudes: A progress report. In L.S. Schulman (Ed.) Review of Research in Education, Vol. 4. Itasca, Illinois: Peacock, 1977.

Snow, R.E. Theory and method for research on aptitude processes. Intelligence, 1978, 2, 225-278.

Snow, R.E. Aptitude processes. In R.E. Snow, P.A. Federico & W. Montague (Eds.) Aptitude, learning and instruction: Cognitive process analysis. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1980.

Snow, R.E. Toward a theory of aptitude for learning I. Fluid and crystallized abilities and their correlates. In M. Friedman, J.P Das & N. O'Connor (Eds.) Intelligence and learning. New York: Plenum Press, in press.

Spearman, C. The proof and measurement of association between two things. American Journal of Psychology, 1904a, 15, 72-101.

Spearman, C. General intelligence objectively determined and measured. American Journal of Psychology, 1904b, 15, 201-293.

Spearman, C. The abilities of man. London: MacMillan, 1927.

Sternberg, R.J. Factorial theories of intelligence are all right almost. Educational Researcher, 1980, 9 (8), 6-13.

Svensson, A. Sociala och regionala faktorers samband med över- och underprestation i skolarbetet. (The relation of social and regional factors to over- and under-achievement.) Rapporter från Pedagogiska institutionen, Göteborgs universitet, nr. 13, 1964.

Svensson, A. Relative achievement. School performance in relation to intelligence, sex and home environment. Stockholm: Almqvist & Wiksell, 1971.

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

Sörbom, D. Detection of correlated errors in longitudinal data. British Journal of Mathematical and Statistical Psychology, 1975, 28, 138-151.

Sörbom, D. A statistical model for the measurement of change in true scores. In D.N.M de Gruitjer, L.J.Th. van der Kamp & H.F. Crombag (Eds.) Advances in psychological and educational measurement. New York: Wiley, 1976.

Sörbom, D. An alternative to the methodology for analysis of covariance. Psychometrika, 1978, 43, 381-396.

Sörbom, D. Personal communication in 1979.

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

Thurstone, L.L. An experimental study of simple structure. Psychometrika, 1940, 5, 153-168.

Thurstone, L.L. A factorial study of perception. Chicago: University of Chicago Press, 1944.

Thurstone, L.L. Multiple factor analysis. Chicaco: University of Chicago Press, 1947.

Thurstone, L.L., & Thurstone, T.G. Factorial studies of intelligence. Psychometric Monographs, No. 2, 1941.

Undheim, J.O. Ability structure in 10-11 year-old children and the theory of fluid and crystallized intelligence. Journal of Educational Psychology, 1976, 68, 411-423.

Undheim, J.O. Broad ability factors in 12- to 13 year-old children, the theory of fluid and crystallized intelligence, and the differentiation hypothesis. Journal of Educational Psychology, 1978, 70, 433-443.

Undheim, J.O A neo-Spearman model to replace Cattell's theory of fluid and crystallized intelligence. University of Trondheim, Norway, 1979.

Undheim, J.O. Toward a restoration of general intelligence. University of Trondheim, Norway, 1980.

Undheim, J.O. Toward a restoration of general intelligence. A selection of papers on factor-analytic dimensions of intelligence. Doctoral dissertation, University of Trondheim, 1981.

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

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

Vernon, P.E. Ability factors and environmental influences. American Psychologist, 1965, 20, 723-733.

Vernon, P.E. Intelligence and cultural environment. London: Methuen, 1969.

Werts, C.E., Jöreskog, K.G., & Linn, R.L. (1973) Identification and estimation in path analysis with unmeasured variables. *Educational and Psychological Measurement*, 1973, *78*, 1469-1484.

Wiley, D.E. The identification problems for structural equation models with unmeasured variables. In A.S. Goldberger & O.D. Duncan (Eds.) *Structural equation models in the social sciences*. New York: Seminar Press, 1973.

Wissler, C. The correlation of mental and physical traits. *Psychological Monographs*, 1901, *3*, No. 16.

Witkin, H.A. Individual differences in ease of perception of Embedded Figures. *Journal of Personality*, 1950, *19*, 1-15.

Witkin, H.A., & Moore, C.A., Goodenough, D.R. & Cox, P.W. Field dependent and field independent cognitive styles and their educational implications. *Review of Educational Research*, 1977, *47*, 1-64.

Witkin, H.A., Oltman, P.K., Raskin, E. & Karp, S.A. *A manual for the Embedded Figures test*. Palo Alto: Consulting Psychologists Press, 1971.

Wittrock, M.C. The cognitive movement in instruction. *Educational Psychologist*, 1978, *13*, 15-29.

Reports from THE DEPARTMENT OF EDUCATION UNIVERSITY OF GÖTEBORG

1980:01 Johansson, Britt: Need of knowledge in nursing and demand for knowledge in nursing education. July 1980

1980:02 Dahlgren, Lars-Owe och Franke-Wikberg, Sigbrit: Social Structure of Society through the Eyes of University Students. October 1980

1981:01 Svensson, Lennart: The concept of study skill(s). Paper presented at the Sixth International Conference on Improving University Teaching. Lausanne, Switzerland, July 9-12, 1980. March 1981

1981:02 Gustafsson, Jan-Eric: An introduction to Rasch's measurement model. March 1981

1981:03 Alexandersson, Claes: Amadeo Giorgis empirical phenomenology. March 1981

1981:04 Åsberg, Rodney: Hos to understand cognitive differences from a cross-cultural perspective? August 1981

1981:05 Reuterberg, Sven-Eric och Svensson, Allan: Financial aid and class bias in higher education. October 1981

1981:06 Gustafsson, Jan-Eric, Lindström, Berner & Eva Björck-Åkesson: A general model for the organization of cognitive abilities. December 1981