GÖTEBORG STUDIES

IN EDUCATIONAL SCIENCES 124

Monica Rosén

Gender Differences in Patterns of Knowledge

ACTA UNIVERSITATIS GOTHOBURGENSIS

Gender Differences in Patterns of Knowledge

GÖTEBORG STUDIES

IN EDUCATIONAL SCIENCES 124

Monica Rosén

Gender Differences in Patterns of Knowledge

ACTA UNIVERSITATIS GOTHOBURGENSIS

ISBN 91-7346-333-7
ISSN 0436-1121

Distribution: ACTA UNIVERSITATIS GOTHOBURGENSIS
Box 222
SE-405 30 Göteborg, Sweden

ABSTRACT

Title: Gender differences in patterns of knowledge.
Language: English
Keywords: Gender differences, knowledge complexity, cognitive skills, feminist thoughts, gender system, hierarchical models, multivariate methods, latent variables, structural equation modelling.
ISBN: 91-7346-333-7

The overall purpose of this thesis is to study gender differences in patterns of knowledge. Knowledge is given a broad definition to emphasise a socio-cultural perspective and to enable the building of bridges between research traditions. The dissertation comprises three separate studies, which have been previously published, and an integrative essay. In the latter, the research approach and results are described and elaborated upon from different theoretical, methodological, and feminist perspectives. Moreover, controversies and paradoxes in the history of educational measurement and research on gender differences are discussed. In the final section of the essay the usefulness of quantitative research approaches for the understanding of gender differences, is discussed in the light of feminist critique.

The core assumption for the studies is that "knowledge" in whatever form it appears, is always complex, and that observable variability may be analysed in terms of structural patterns – patterns of knowledge.

The first two empirical studies utilise performance scores from 13 different cognitive ability tests and from standardised achievement tests in mathematics, Swedish and English. The data was collected in 1980 and comprised all 12-year old students in grade 6 in two communities. The third study utilises performance scores from Document reading tasks, selected from the 1990/1991 IEA Reading Literacy study. This analyses comprised 9-year-olds and 14-year-olds, in representative samples from 25 and 22 countries respectively. In order to reveal latent patterns, a multivariate hierarchical approach was adopted for all three studies, with the aid of structural equation modelling.

The first study revealed a similar latent structure of ability dimensions for boys and girls. However, despite almost equal observed performance, substantial gender differences were found in the latent dimensions of cognitive abilities. The girls showed higher levels on general analytical and verbal-educational dimensions, whilst the boys showed higher levels on a general spatial dimension, and on several narrow dimensions related to verbal, numerical and spatial content. Variability differences were found on spatial dimensions only, where the spread was wider in the male group.

In the second study, the results from the first study were further investigated in an analysis of the impact of missing data. In a multivariate analysis where the test scores from boys and girls lacking data on the three standardised tests were included, the pattern of gender differences on latent dimensions changed. The girls advantage on the analytic dimension increased, while their advantage on the verbal-educational dimension decreased slightly. The boys advantage on narrow numerical and verbal dimensions decreased, while their advantage on spatial dimensions increased.

There were two major results in the third study. First, performance on Document reading tasks are determined both by general Document reading proficiency and by the specific content in the tasks. Second, gender differences were found in all these dimensions, but the pattern of gender differences varied between countries, implying strong cultural influences. Again, the patterns of gender differences on latent dimensions were not deducible from observed scores.

In the essay, a socio-cultural explanation is proposed, according to which, the differences found in the empirical studies have a material ground in both the vertical and the horizontal dimensions of the gender system.

Acknowledgements

Now, as I approach the final tense hours of my dissertation, I look back over my time as a doctoral student and realise the fulfilment I have obtained through being part of a stimulating workplace, with wonderful and caring work colleagues.

Both of my mentors, Inga Wernersson and Jan-Eric Gustafsson have, through all these years always been there for me. I am not only grateful for the knowledge, skills and insight, but also the immense inspiration and new challenges they have given me. They have shown a genuine interest in my writing, tremendous faith in my capacity and positive feeling towards me as a person. This support has been invaluable.

In the autumn of 1992 I received my doctoral studentship and without this it would have been extremely difficult to have concluded this dissertation. The same year, through the MAGIK-project (Models for Analysis of Group Differences in Cognitive Abilities), I had the opportunity to make my debut in the conference world through AERA in San Francisco, where I was able to present some of the analyses which were later published as the first empirical study of this dissertation. Two years later, in conjunction with UCLA school of Education, USA, I had the opportunity, for three months, to improve my multivariate analysis methods under the supervision of Professor Bengt Muthén. It was there that the analyses behind my second article, the missing data study, was undertaken. It was an enlightening time, and I am extremely grateful for the friendship and help I received from Professor Muthén and his PhD students. With the completion of these two studies I was able to take a Licentiate degree before Christmas 1995. These studies were financed by the Swedish Council for Research in the Humanities and Social Science (HSFR).

My third empirical study evolved from the MALI -project (Multivariate Analyses of Literacy) offering the possibility to re-analyse data from the 1990/91 IEA Reading Literacy study, and I would therefore like to offer a very special thank you to Professor Ingvar Lundberg, the supervisor of the MALI-project. It was an enjoyable project with an enormous amount of data, an immense amount of modelling and great inspiration gained from the project leaders. From the international comparative approach, I have found an additional dimension for my gender perspective. The comparative study has been partly financed by the Bank of Sweden Tercentenary Foundation (Riksbankens Jubileumsfond) and partly by HSFR.

The integrative essay has been written during the past year and many people have helped with valuable points of view and encouraging voices. I owe a debt of gratitude to Dr. Mac Murray for his work, at short notice, before the final seminar with my first draft. I would also like to offer a thankyou to Anette Andersson, Eva Gannerud, Ann-Katrin Jakobsson and Anna Lindbom for friendly critical reading and constructive suggestions. Further good advice and appreciated views on parts of the contents have been received from Kjell Härnqvist, Sven-Eric Reuterberg and Allan Svensson.

Several sources of inspiration and insight and have come from the "gender perspective on pedagogy group" (KKP) and "the quantitative working seminar" (KVAR), where I have been given opportunities to raise issues and present analyses throughout my work.

The English written language has been a challenge, therefore I would like to put forward a special thank you to Peter Driscoll who accomplished a great deal of work in a short period of time in reading and giving suggestions for the final manuscript.

With respect to the professional editorial help I received from Inga-Lill Berntsson, Agneta Österlund and Doris Gustafson without whose help the dissertation would hardly have been ready in time, I am eternally grateful.

Finally, to my family and closest friends, who at times have wondered what I have been trying to achieve, especially over the past few months. In spite of this they have shown me the greatest understanding and have always been there for me when I have needed them. Thank you to my loved ones.

Mölndal in September 1998

Monica Rosén

Table of Content

Gender differences in patterns of knowledge: An integrative essay

Monica Rosén
Department of Education and Educational Research
Göteborg University

Part I

Introduction

This dissertation comprises three empirical large-scale studies, which all investigate gender differences in patterns of knowledge. In this essay, the empirical studies are linked to each other theoretically and methodologically. The purpose has been to place the studies in a societal and a research context, and particularly so from a feminist perspective. I have devoted this final essay some length for several reasons. One is to make the studies more accessible to researchers not familiar with advanced statistics. Another is to apply feminist perspectives and considerations on educational measurement research. A third is to illustrate the necessity of research on gender differences in patterns of knowledge.

In the first chapter some of the major problems and questions I have dealt with during the course of my research are presented. The context and the research problem is described. The main purpose is to discuss the necessity of research on gender differences in patterns of knowledge.

In the second chapter, some of the methodological problems of investigating gender differences are described. I also present the rational for my choice of method, along with a description of the theory and methodology used for the empirical studies. The results of my empirical studies are presented in the third chapter. I have here allowed myself a more extended discussion of the results and the research praxis in relation to both the history and to contemporary research on gender differences in aspects of knowledge and skills. In the fourth chapter gender differences in patterns of knowledge are reflected upon from a socio-cultural perspective through the lenses of the gender system. I end this essay by discussing the feminist critique against quantitative methods.

First of all, however, two expressions need clarification, namely gender and patterns of knowledge.

The distinction between “sex” and “gender” is a frequent topic for debates within feminist research and epistemology. A common use of the term “sex” is to restrict it to referring to biological distinctions between males and females, while reserving the term “gender” to refer to the psychological features or attributes associated with these categories (e.g., Deaux, 1985). This is similar to my understanding of the matter. The use of gender is also more accurate for the connection to “the gender system” identified by feminist researchers (e.g., Hirdman, 1987), since it marks the cultural and structural dimension. My written language is more inconsistent, though, because it has been difficult to decide when sex should be used rather than gender. Thus, I have favoured the use of gender. I have also frequently used “males” and “females” regardless of age level, which in my empirical studies ranges between 9 and 15.

The main focus of my dissertation is on patterns of knowledge in relation to gender. The data used originated as measures of performance on cognitive tasks. This performance is a manifestation of knowledge. Knowledge, in whatever form it presents itself, is never unidimensional. Furthermore, knowledge is constructed and developed in a socio-cultural context. I use "patterns of knowledge" to emphasise this complexity. In my studies, I use different terms like abilities, skills, capacities etc. All of these are constructs, referring to the same phenomenon, i.e. knowledge. This expression illustrates a conception of the matter, which also enables bridging between research traditions that seldom come together.

However, concepts of knowledge, like "intellectual ability" and "rationality", are not value-free or neutral constructs in our society, and neither are any notions of gender. The societal value judgements of these, along with a connection to social history, does make the subject controversial. Lather (1991) has stressed that science and politics are never entirely separable, her message is: "nothing is innocent and everything is dangerous."

Gender Differences in Patterns of Knowledge: Research Context and the Problem

In this section I will present and discuss the research context for the three articles included in the dissertation. Two aspects, in particular, are central. The first is to identify the relevance and need to investigate gender differences in patterns of knowledge, which has been the overall purpose of my three articles. The other is to discuss some of the many conflicts between feminist thoughts and actions and the psychometric research tradition to which my empirical studies belong.

The context

One of the most problematic issues in investigating gender differences is the ever-present hierarchical relation between females and males - the gender system as feminists have defined it (e.g., Hirdman, 1987; 1988; Harding, 1986; Scott, 1988). On the structural level females is the subordinate group, which makes it impossible to disregard the political side of the matter. This will become more obvious in my attempt to elaborate on some of the feminist lines of thought that are central to my research interests.

Investigations of gender differences in proficiencies and skills have a long history, but there is still need for further analyses. In a review of gender differences in school in the Nordic countries Wernersson (1989) concludes that most studies are concerned with gender differences in classroom interaction. This problem is of obvious pedagogical interest since one goal for education is to provide equal opportunities for males and females (Lpo, 1994):

> *The school has an important task to bring about and anchor in the pupils the values that our society rests upon. The inviolability of human life, individual freedom and integrity, the equal value of all humans, equality between women and men and solidarity with the weak and vulnerable are those values the school shall form and bring about.* (Lpo, 1994, p.5)

Gender aspects of cognitive performance is one important part of this goal. There is always a need for information regarding reasons why differences emerge and develop, and what the consequences may be in a longer perspective.

For example, at levels above compulsory school, educational choice is strongly gender differentiated (Reuterberg & Svensson, 1998; SCB, 1997a). The previous vertical division of education seems to diminish, or at least be postponed until the entry of working life (SCB, 1995; SCB, 1997b). The horizontal gender division in Sweden is clearly demonstrated by the choice of programs in upper secondary school. Females represent 70% or more of the students on programmes directed towards e. g., 'Health Care', 'Child Recreation', and 'Social Science', while programmes directed towards e. g., 'Industry', 'Construction', and 'Electrical Engineering' have more than 90 % male students (SCB, 1997a). Only a few educational programmes have a more equal distribution of males and females e. g., 'Media', 'Restaurants and Catering', 'Business Administration' and 'Natural Science'.

At the university level, the vertical dimension is no longer as obvious as it was just a few of years ago. Males were then over-represented on the more prestigious educational alternatives (e. g., Medicine, Law, Architecture, and Graduate engineer). A rather dramatic shift has taken place in the past few years. In 1996/ 97 a total of 50% or more of the new students, on all programmes but engineering, are women (SCB, 1998).

However, in postgraduate education, males still form the larger group, and the proportion of males increases by graduate level (Licentiate and Doctorate). Women are still a minority at the top of the academic hierarchy (SCB, 1997a). The horizontal division at the undergraduate level remains intact. Women dominate the field of teaching and health-related science with 77 and 89 per cent respectively, while men dominate the fields of technology/natural science with 73 per cent (SCB, 1997a).

One question is if the prestige associated with some programmes will diminish as the proportion of females increase. There are several indications that this may be the case. An example is medical education, where women now are equally represented as males. There are reports of a decreasing status of the medical profession. Sub-fields of medicine are becoming genderized and ranked hierarchically, i.e. differentiated both in terms of specialisation and gender, where the female fields have a lower status (Einarsdottir, 1997). Another example is the increasing number of female school principals, previously a male occupation (Olofsson, 1996). Salaries are decreasing, and the power that previously was attached to the position has moved to other areas, despite the fact that school principals of today have a better education (Ullman, 1997). Education is one important strategy for women to reach the same status as males (e.g., Florin & Johansson, 1993), so reports like these are troublesome

Knowledge is power, and equal levels of competence should remove any legitimate argument for female subordination. The maintenance of female subordination may be understood by the two principles: the rule of distinctive separation of the two sexes, and the rule of the "male norm". The latter principle is also referred to as the "hegemonic masculinity" principle (Connell, 1987), which states that higher value is automatically assigned to things masculine (Hirdman; 1987, 1988).

Studies of the development of knowledge and skills during the school years have so far provided very little information that helps to explain the pattern of horizontal or vertical division between males and females in secondary education.

Patterns of gender differences enjoy a great interest in public media as well as in peoples' belief system. The privileged position of men *on the structural level* partly explains why many old beliefs about female and male "nature" appears so frequently in both media and in private conversations. Many of the beliefs reflected address notions of gender differences in cognitive abilities, proficiencies and achievements. Research has an important role in enriching our understanding of gender differences in cognitive abilities and their relation to performance, and also in setting bounds on speculation in these areas.

There are two major reasons for my interest of patterns of knowledge and gender. The first is for the societal and educational reasons mentioned previously. The second is the lack of feminist work within this field.

There are several reasons why feminists are lacking in mainstream educational research, and particularly so in the field of educational measurement. It was early acknowledged that this field had numerous misinterpretations and prejudices against women (Woolley, 1910 cited in Shields 1975; Hollingworth, 1914), and it was and still is a well-established male research area. According to Hallberg (1992) the common basis for the feminist critique against traditional science, is the presumption that the male/masculine has an important impact on both form and content of research. This presumption comprises science as an institution, theories, the philosophy of science and the definition of research problems. Science is regarded as one of many social institutions where "the male" is hegemonic. Traditional science is defined by males from male perspectives.

Another reason for female/feminist avoidance/abandoning of science is that traditional methods have not proved useful or efficient enough for the research questions and perspectives urged by women and feminist researchers (e. g., Reinharz, 1992; Davies & Esseweld, 1989; Lather, 1991; Hallberg, 1992). Yet another motive may be identified in female/feminists researchers biographical writings, where their personal experiences from traditional research reveals a number of conditions and incidents of oppression in their daily working praxis (SOU, 1995:110).

In conclusion it can be argued that the lack of females/feminist researchers is unfortunate for two reasons.

- First, traditional educational research needs female/feminist researchers since androcentrism is more or less unaware, and needs to be concretely identified and balanced.
- Second, from a pedagogical viewpoint, this type of research is often quite influential on the educational system, which in turn has an impact on the gender system.

Gender differences are often given biological explanations, which sometimes refer to previously abandoned theories (as for example "man- the hunter and woman-the gatherer"). It seems particularly important to contrast such ideas with well-founded results and illustrations of how socially constructed the reality is. The nature/nurture question seems, however, always to be present when gender differences are in focus.

The problem and its history

As mentioned before, the quantitative measurement tradition and the research on human cognitive abilities or intelligence has been strongly criticised by feminist researchers. This is in part due to the history of the subject. From the start, some hundred years ago, there existed a more or less explicit assumption of female intellectual inferiority (e. g., Shields, 1975; Rossiter, 1982; Dijkstra, 1986; Fausto-Sterling, 1985), and several studies aimed to prove this assumption (e. g., Siegwald, 1944). Prejudice and misinterpretation of observed gender differences have been pointed out repeatedly during all these years (e. g., Hollingworth, 1914; Maccoby and Jacklin, 1974, Gould, 1981; Tavris, 1992). There is perhaps no wonder that many feminist researchers have come to the conclusion that theses types of studies only produce sexist results, which implies that the female inferiority assumption is a part of both the methods and the theory (Fausto-Sterling, 1992; Reinharz, 1992; Mizra, 1998).

It may, however, seem paradoxical but it is the same criticised research tradition, which did early, and against common belief, produce scientific evidence which showed females and males to be equally intellectually capable. It is undeniable that several distinguished male scholars expressed doubts about females' intellectual ability level, and particularly so when their data tended to show greater variability in male groups. It is also true that many of them changed their position, as methodology was refined and empirical evidence emerged, and as the critique from feminists and others was heard (e. g., Noddings, 1992; Hyde & Linn, 1986).

It is a fact that most of the early tests of intellectual abilities were developed by male researchers, and often in collaboration with "field-workers" (e.g, doctors and teachers). The time when "modern" cognitive test development started coincided with a societal need for more "objective" criteria and effective methods, rather than mere "subjective" judgements, when ranking or diagnosing people. The "testing age" began with Alfred Binet in France 1905, Charles Spearman in England 1904, and Lewis Terman in USA around 1916 (Cattell, 1987; Block & Dworkin, 1976).

In France, Parisian school authorities asked Binet and his colleague Simon to clarify the diagnosis of irremediable forms of backwardness in schoolchildren, which resulted in the so-called Simon-Binet test. The test aimed to distinguish between true mental defect and mere lack of school progress.

Spearman in London was more interested in the fundamental question of the definition of intelligence. More specifically, he asked whether we should think of intelligence as a single power or as a crowd of faculties, as seemed to be implied by Binet's multifocal view. Spearman is also the originator of factor analysis, a method developed for answering this question.

In the USA, around 1910, Terman at the Stanford University, took over Binets test and revised it. Block and Dworkin (1976) describe the revision as follows:

> *By editing, rearranging, and supplementing the original Binet tests, he [Terman] finally worked out a series of tests for each age which the average child of that age in about one hundred Californian children could pass.* (Lippman, in Block & Dworkin, 1976, p. 6*)*

An adaptation of the Stanford-Binet test was then put together in a period of a few weeks, for the army in World War I. Thousands of (male) recruits were tested, and the interpretations from analyses of these masses of data started the IQ-controversy. The tests became instruments for classifying people, rather than a measure of intelligence. What was considered experimental research in Europe quickly became a large-scale industry in the USA. The American use does not, however, disqualify research questions regarding human cognitive competencies and their measures; on the contrary it urges for more in depth understanding.

From the feminist critical review of early research in the field it seems that the beliefs about womens' intellectual capacities, if any, were limited to expectations that females would not perform as well as males on cognitive tests. The most common belief in all Western societies at the end of the 19th century was that females were not suited for intellectual work (e. g., Noddings, 1992). One may argue that the tests and the methods were "gender blind", because if females were thought of at all, during the course of development of ability tests, it was seldom explicitly reported. To my knowledge, typical female tasks were never considered as models for cognitive abilities or test development. It is thus possible (but not necessarily the case), that both the conception of abilities and their measures would have looked different if more females had been involved and/or listened to in those early days. Noddings (1992) gives several examples of intellectual capacities valued by women, which do not appear in current standard batteries (e. g., listening, oral and written text interpretation, and interpersonal reasoning). It is thus plausible that current scholarly definitions of the human intellect are too narrow.

Many of the early cognitive tests have been revised and are still used in both research and elsewhere, and thus in need of deeper understanding. The samples used today for development and validation of tests are usually mixed in terms of males and females selected from more comparable groups than in the early days, when boys were given better educational opportunities. Furthermore, female researchers are usually involved in the process of instrument development and evaluation. Although the researchers of the area of today probably have quite different conceptions of both gender and cognitive tests, the historical burden cannot be entirely disregarded. However, although the "hegemonic masculinity" characterised the development of both cognitive tests and their related theories, it is not clear what consequences this may have had.

The feminist scholarly interest

Above, the feminist standpoint regarding gender differences in relation to cognitive performance and its connotations that this is an uninteresting and unnecessary research domain has been discussed. As Chipman (1988) states, gender differences in cognitive abilities is "*a far too sexy topic*". Her conclusion is based on the fact that gender differences are so small, or rather, that the variations within the female population is rather similar to the variations within the male population. A similar conclusion has been put forward by Jacklin (1989), one of the authors of *The psychology of sex differences,* which is one of the most important books on gender difference. Her argument is similar, namely that the interest should vanish since gender seems to be of no importance in relation to intellectual abilities.

At the same time, however, there is a feminist discourse of gender differences in thinking, reasoning and acting styles (see Tavris, 1992), which in a way refers to the same phenomena. Here the issue is expressed as "women's ways of knowing" (Belenky, Clinchy, Blythe, Goldberger & Tarule, 1986), understanding (Tannen, 1990), reasoning (Gilligan, 1982), which implies that women's minds work different from men's. I believe that as long as these types of beliefs or notions are put forward by researchers or in the public opinion, the research domain is warranted.

However, given the small differences found, I agree that focus should not be on the question which is the better sex on the vertical scale, which usually is the subject in the public debate. Nevertheless, the vertical aspect should not be completely neglected either, particularly when gender differences in knowledge and skills seem to agree poorly with gender differences in power and status.

What is more interesting is the dynamics of both the within- and the between- group differences, the possible implications of them, and their relations to other qualities (e. g., self-esteem, interest, motivation, opportunities, educational and occupational choices, quality of life) and societal (e. g., selection to higher education and occupations) actions. It is also important to remember that although differences should not be regarded as eternal, it is not self-evident that similarities are either.

Finally, newly developed theories and methodologies offer alternative ways to analyse and understand cognitive achievement data, and this automatically calls for new investigations and reanalyses of group differences.

Common quantitative analysis procedure

As noted among critical feminist researchers, gender differences are often reported in traditional studies of knowledge and skills but not much elaborated in terms of explanations and theoretical discussions (e. g., Deaux, 1985; Wernersson, 1989). Since all notions of gender difference also always have a political or ideological dimension, it seems extra important that the purpose of investigating gender differences is made explicit, and that the meaning and interpretation of the result is expressed by the researcher. The risk of misinterpretation and misuse is always present in *all* types of research, but particularly so if results are left without comment.

Nevertheless, it belongs to common research practice to check for gender differences in statistical analyses, even though gender is not the main focus of interest. One important reason for this is, of course, to satisfy a curiosity and see if the results are the same for females and males. An analysis of gender differences may also be a part of the validity and reliability analysis. If gender differences (or no differences) are found when not expected this may be an indicator of phenomena that calls for further analysis. However, as Eagly (1997) points out, whether to report the result of the gender comparison or not is a decision made by the researcher. If there are not any differences, and the researcher finds null results uninteresting, for example, the researcher may chose to exclude the analysis of sex differences in the report.

One may, however, ask if the results may be readily interpreted if gender differences are described without further discussion. As a minimum the reasons for investigating gender differences should be made explicit. On the other hand, a large part of the research on differences in cognitive dimensions may be considered as a search and description of the unknown or theoretically undeveloped, so interpretations, explanations and theoretical discourses remains to evolve. This fact accentuates the importance of feminist involvement in this particular research praxis.

The problem of single variable research

In most studies of cognitive performance, gender differences have been investigated with respect to one single variable at a time. The implicit assumption in this approach is that performance on the single variable is determined by only one underlying dimension, for example that performance on a mathematics test is governed by mathematical ability, and consequently interpretations may be made in terms of such an ability. There is rarely any analysis of how different subject areas relate to one another, and how such patterns are related to gender. As Snow and Lohman (1989) have argued:

> *...sign-trait interpretations of test scores and their intercorrelations are superficial summaries at best. At worst they have misled scientist, and the public, into thinking of fundamental, fixed entities, measured as* amounts (Snow & Lohman, 1989, p. 317)

This traditional *univariate* way of investigating gender differences is problematic for several reasons. One is because it gives an unfair picture of group differences, since it singles out a particular proficiency dimension from a larger context of many proficiencies or abilities without even discussing it in such terms. Perhaps this is what Chipman (1988) has in mind when she criticises traditional research for reporting gender differences in mathematical and spatial ability areas without considering the influence of other variables such as general cognitive ability. In her own words:

> *In the research on women and mathematics, it turned out that the obsession with understanding sex differences in mathematics performance and course enrollment (themselves greatly exaggerated in popular and much professional thinking) detracted even from the illumination of those phenomena. Researchers focused on variables that seemed relevant to the difference neglecting others that are more important in understanding the underlying phenomena of individual performance and enrollment choice such as general cognitive ability* (Chipman, 1988, p. 48).

Chipman also points at the need for knowledge regarding what processes are involved in successful mathematical problem solving. Another reason concerns validity since the univariate approach disguises the difference between test performance and underlying ability both in statistical analyses and in the researcher's interpretations. How performances on different cognitive tasks are related to one another and how such patterns may be related to gender thus became the main concern in my three empirical studies. One of Willingham and Cole's (1997) conclusions in their extensive study of gender and fair assessment was that

> *Understanding the underlying skill level may be critical to understanding the nature of gender differences and effective ways to address fairness issues in testing*. (Willingham & Cole, 1997, p. 170).

My research interests thus seem to coincide with a contemporary interest in the matter.

Summary

In this chapter, the context, the social history and the many problems, paradoxes and needs associated with large-scale analysis of gender differences in patterns of knowledge have been considered, and particularly so from a feminist viewpoint.

- Aware of the critical voices, the necessity of research on gender differences in patterns of knowledge has also been argued. The societal need, peoples' belief system, the power of large-scale studies, the pedagogical need to understand educational performance and measures, and, as I argue, the feminist need, has framed the importance of the matter to me, as it has to others in the field. I have also stressed that one of the major problems in the field is the limited number of feminist researchers.
- As is often forgotten in the heated feminist debate, the same tradition that is criticised, is also the very same tradition which repeatedly has shown that any notion of female intellectual inferiority is groundless, and the same tradition needs to continue to do so. It is also easy to overlook that the same type of methods and instruments that now are being criticised have also served to improve gender equality.
- Finally, I have pointed at the problem of traditional single-variable examination procedures, a problem, which has guided me through all three articles. As has been acknowledged before, knowledge is always complex.

In the next chapter, the problem and possibilities of investigating gender differences in patterns of knowledge will be further elaborated, but this time in terms of methodology and associated theoretical assumptions.

Theory and Method

In this chapter theoretical and methodological issues are considered. First problems of investigating group differences in performance are discussed. Then a hierarchical model of abilities, which provides the framework for the empirical studies, is described. Finally the method of analysis is presented.

Gender differences in educational measures

A general problem with studies of group differences is the polarisation between groups that automatically occurs when differences are presented. This is particularly true for the study of gender differences where females are contrasted with males. Thus, whatever males are, do, accomplish or think is not what females are, do, accomplish or think. Although distinctions between the groups are necessary for our understanding, they often tend to get exaggerated and make us forget similarities. For researchers interested in group differences this is a well-known and intricate problem. There is still no good solution for how to handle similarity and differences simultaneously. As frequently has been pointed out, the gender difference in cognitive performance is generally very small (e.g., Hyde & Linn, 1988). Of all the factors that create individual differences, gender usually accounts for less than 2 percent.

Gender differences are often expressed in terms of mean and variance differences. Group averages are indicators at the group level and should thus not be translated to the individual level. The tendency to apply group differences to each and everyone within a group is not only incorrect but also a pedagogical problem. In fact, there may in reality exist no single person, who in every respect fits the average female or the average male reported on in research.

Another general problem in investigating socially determined group differences such as gender differences in knowledge and skills, is the risk of interpreting results as eternal truth given by nature. Behaviour and proficiencies arise in a social context of interactions that are shaped of time and culture, and an important question for research is to identify sources or conditions for understanding both individual and group differences. The persistence of some gender differences over time and/or nations, is often interpreted as a law of nature, but as research also has shown, social change takes time.

From an educational point of view, the less a problem is understood or the more complex the problem is, the harder it is to act upon gender differences. There is no reason to assume that old findings are eternally valid; on the contrary, there is reason to assume that differences between groups like gender vary as the society develops in various ways.

Related to this is the problem of the universal impression that reported gender differences often give. Sometimes this has to do with how the researcher has expressed the difference found, but more often how results are treated in contexts outside the research report. If gender differences are perceived as universal, they are almost always interpreted as a result of nature (or evolution). When and if consistent patterns of gender differences are found, it may well be a sign of an unchanging society. This is also why it is necessary to investigate differences between groups such as gender.

One way to investigate the nature/nurture question is by comparative studies, for example by comparing gender differences in patterns of knowledge over countries. A problem is, of course, that even if similar patterns are found they may have different explanations in each country. On the other hand, if there is a dissimilar pattern across countries, then biological and evolutionary theories and explanations are seriously challenged.

There are, of course, many other reasons for comparative studies than the nature/nurture question, and gender equity is one of them. In recent years, many societies have become increasingly concerned with the twin imperatives of equity and efficiency. Education represents one of the largest societal investments. Evidence that not all groups within society are equally able to receive and derive benefit from the investment, inevitably raises questions related to the equitable distribution of resources and benefits. Comparative studies may thus reveal patterns that call for further research. Furthermore, an international perspective on national patterns of gender differences may enrich our understanding as well as it might encourage new questions and hypotheses. Last but not least, comparative studies provide improved opportunities to empirically investigate various hypotheses of socio-cultural influences on gender differences.

The problem of comparability

As many researchers have observed (e. g., Maccoby & Jacklin, 1974; Hyde, 1981; Hyde, 1990; Hyde & Linn, 1986; Hyde & Linn, 1988; Hyde, Fennema & Lamon, 1990; Linn & Petersen, 1985; Reinhartz, 1992) a substantial part of the studies that report gender differences has included highly selected and/or small samples. Some of the samples suffered from restriction of range from the very beginning, which is no problem as long as the groups that are to be compared are selected equally. Others were selected as convenience samples, which often makes it impossible to investigate whether the groups are comparable or not. In most cases, however, there is no comment, report or analysis of missing data, which makes the results questionable. Hyde and Linn (1986) in their textbook on meta-analysis techniques point at this type of methodological difficulties. Analysis is likely to be confusing because there are holes in the framework of samples and measures, and they interact in ways that often cannot be controlled. The comparability question is one of the most fundamental in studies of group differences since it concerns the reliability and validity of conclusions. Chipman (1988) is quite critical towards meta analysis where the researcher collects all relevant studies and computes a standardised measure of size of the effect of gender in each sample. Chipman points at the possibility that all or most of the studies included in a meta-analysis may be biased or non-representative in the same way because of research convenience, a problem that she feels has been given far too little attention when gender differences are reported.

The problem of the American dominance

The dominance of USA in this field of research creates problems that are similar to the universality problem, that is the problem of over-generalisation. American reports are persuasive by mere quantity and it is quite easy to assume that the patterns of gender differences and the theories and explanations of this culture are universally valid, thereby overlooking possible cultural influences. This may be particularly problematic for the research and understanding of group differences. Although extensive reviews and reanalyses like Maccoby and Jacklin's "*The psychology of sex differences*" (Maccoby, 1966; Maccoby & Jacklin, 1974) and Willingham and Cole's "*Gender and fair assessment*" (1997) are extremely valuable for researchers all over the world, it is easy to forget that all the studies they review are selected, and most of the selected studies are American.

Maccoby and Jacklin (1974) evaluated a large body of work on sex differences (the term gender was not yet introduced). In their book they repudiated a number of common claims for sex differences, for example that girls are more "social", "suggestible", "passive" and have lower "self-esteem" than boys, that boys are more "analytic", that girls are more affected by heredity and boys by environment, that girls lack achievement motivation, that girls are auditory and boys are visual. They found four types of differences to be fairly well established, namely that girls have greater verbal ability than boys, that boys excel in visual- spatial and in mathematical ability, and that boys are more aggressive both physically and verbally. They also left a number of the common claims on sex differences to be determined in future research. Their conclusions have in later research been criticised for not recognising the disparities among studies and the relatively small size of many of the differences cited (Wilder & Powell, 1989).

The ETS gender study (Willingham & Cole, 1997) is the most recent, and one of the most extensive studies ever done on gender in educational settings. It describes patterns of gender differences from the fourth grade to graduate school and comprises nationally representative samples as well as self-selected samples of high-achieving students. Over 400 tests and other measures were analysed. Large-scale analyses of this kind are impressive, and they tend to serve as a norm, reference and validation of other smaller studies, and tend to carry weight on studies conducted in other countries too. Studies that show results of gender differences not in line with their conclusion thus risk to be disregarded, especially within the USA, but also outside. In the ETS gender study articles from all over the industrialised world were selected in the review of research. This makes it even more tempting to generalise their results and conclusions on a universal level, although most of their major conclusions were made on the analysis of American data. The table below summarises some of their results.

Table 1: Gender differences in grade 12 in the USA (Willingham & Cole, 1997)

Differences in Means		**Differences in Spread of Scores**	
Females	**Males**	**Females**	**Males**
Verbal-Writing	Mechanical/Electronic		Civics
Verbal-Language use	Geopolitical		History
Perceptual Speed	Natural Science		Science
Short-Term Memory	Spatial Skills		Math
Study Skills	Math-Concepts		Reading
Verbal-Reading,			Writing
Math-Computation			
Abstract Reasoning			
Verbal-Vocabulary Reasoning			
Social Science			

= Very Small	= Medium-Large	= Very Large

In terms of cognitive performance they found the following differences for the 12th graders: Females showed on the average better results on a wide array of verbal productive tests as well as on measures of math computation (arithmetic calculations, e. g., multiplication, division, fractions and percentage), perceptual speed, short-term memory, study skills, abstract reasoning and social science. Males showed on the average better results on math concepts (higher level math skills, e. g., concepts of number, definitions and procedures, quantitative reasoning, applications, problems solving, knowledge specific to algebra and geometry) mechanical/electronic, geopolitics, natural science and spatial tests. Females performed on the average better in most subjects, while males showed a higher average on traditional science subjects and on spatial tests. The differences were ranked in terms of magnitude and are reproduced by the grey-scale in the table; differences marked dark grey were considered very large, difference marked medium grey were considered medium to large, and differences considered very small are marked light grey.

In their analysis of changes over time, the main pattern was that gender differences increased from grade 4 to 12. They concluded that at the younger age-levels, the differences were non-existent or minor. Larger differences did not occur until later and then at different times for different subjects.

They also reported that the spread of scores, i.e. the variance, was greater in the male group for writing, reading, math, science, history and civics. The largest variability differences were found in civic, history and science, while the difference in reading, writing and math was considered small. Again, the trend analysis indicated that the differences grew with age. They commented the results as follows in their executive summary:

> *Gender differences are not easily explained by single variables such as course-taking patterns or types of tests. They not only occur before course-taking patterns begin to differ and across a wide variety of tests and other measures, but they are also reflected in different interests and out-of-school activities, suggesting a complex story of how gender differences emerge* (The ETS Gender study, Executive Summary, Cole, 1997, p. 16)

It is important to keep in mind that their results, methods, assumptions and interpretations are as open to challenge as any other (smaller) study in any other country. And most important, gender differences may be different in different countries, if cultural and contextual influences are.

In the following section an alternative multivariate research approach will be described, which enables investigations of complex data, since it may reveal the patterns of knowledge that explain test performance. However, first a model, closely associated with the methodology will be described.

A hierarchical theory of human cognitive abilities

The question whether human cognitive abilities are unitary or multifaceted has been occupying educational psychologists all over the world ever since Binet and Spearman did their work in the beginning of the 20th century. The theoretical model that today has gained most empirical support (Carroll, 1993) is a hierarchical model with three levels or strata. The various ability constructs and notions of relations between general, broad and narrow cognitive dimensions, rely on theoretical ideas from the past (Gustafsson, 1984; Gustafsson & Undheim, 1996; Carroll, 1993). The hierarchical model combines Spearmans theory of general intelligence as a unitary dimension, Thustone's findings of multiple abilities and Cattell and Horn's (Cattell, 1963; Horn & Cattell, 1966) ideas that human cognitive abilities can be hierarchically ordered and their distinction between fluid intelligence and crystallised intelligence (Gustafsson, 1984, 1988).

Today's scientific view on intelligence has its roots in the beginning of the century when psychological measurement started. Before that time, the notion of intelligence was founded in personal beliefs and philosophical thoughts among distinguished men (Cattell, 1987). Cleary (1992) points out that gender differences in aptitudes and achievement were noted long before there were any test scores to compare, referring to both Plato and Aristotle, and of course to the male advantage. Hollingworth's request, in 1914 (cited in Walsh, 1997) bear witness that the introduction of systematic observations would be of great importance for the so-called "woman question". Hollingworth asked that a psychology of women should be written:

> *based on truth, not opinion; on precise, not on anecdotal evidence; on accurate data, rather than remnants of magic* (Hollingworth, 1914, p. 49).

Spearman's questions whether the human intellect should be thought of as "a single power" or "a crowd of faculties" is considered fundamental (Spearman, 1904). He also developed the first factor analytic method to investigate this question. Building on the correlational technique developed by Francis Galton at the end of the 19th century, Spearman found that all measures of cognitive performance were positively correlated, and that the correlation was highest among complex and abstract tasks. His interpretation of this pattern was that all tasks share a common dimension, *g* (general intelligence), and that each task also require an ability specific to that task (Cattell, 1987)

Thurstone (1938) extended Spearman's unidimensional model to encompass multiple factors. With a newly developed factor analytic technique (multiple factor analysis) Thurstone was able to identify about a dozen "primary" factors in large-scale empirical studies. The number of "Primary Mental Abilities – PMA's" was later considerably extended by Thurstone and his followers. There was, however, a growing disenchantment with multiple factor analysis and the PMA system, which had to do with its limited utility for describing the structure of ability, and for its tendency to yield as many factors as there are types of test items (e.g., Humphreys, 1962). Furthermore, differential aptitude batteries did not seem to have differential predictive power for achievement in different subject matter areas, which questioned the value of primary abilities in practical applications (Gustafsson, 1992a). Broader abilities were thus needed for both theoretical and practical reasons.

One way to bring order among PMA's is to analyse the correlations between factors, and thereby identify so called second-order factors. This approach yields a hierarchical organisation, which includes both broad and narrow ability dimensions. Horn and Cattell, (1966) applied such techniques to construct a hierarchical model with two broad factors, fluid intelligence and crystallised intelligence. They also identified some further broad factors (e.g., General Visualisation, *Gv*, and General Fluency, *Gr*)

Carroll (1993) has reanalysed most studies of the structure of abilities, and has extended the Cattell and Horn model into a model with three levels. Using confirmatory factor analysis Gustafsson (1984) also arrived at a hierarchical model with three levels. This model is described in greater detail below (see also Figure 1).

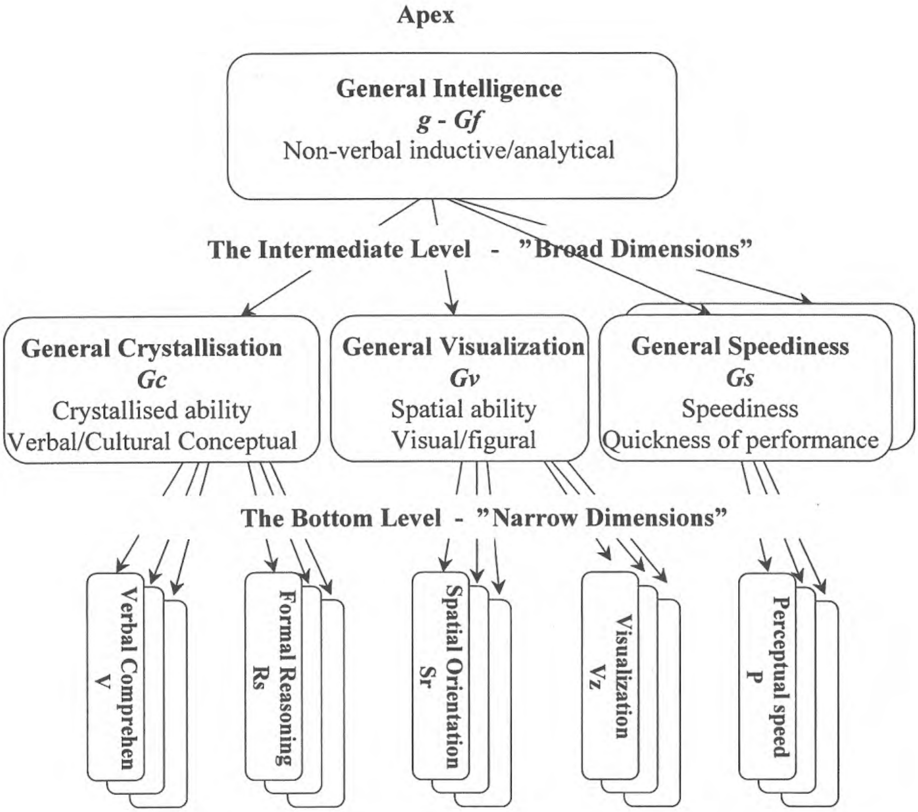


Figure 1: The hierarchical organisation of human cognitive abilities

General intellectual ability

At the apex there is *g*, general intelligence. Current interpretations regard *g* as a combination of Spearman's general intelligence concept, and Fluid intelligence (*Gf*) from Cattell and Horns Gf-Gc theory (Cattell, 1943, 1971, 1987, Horn & Cattell, 1966). Recent findings from correlational and experimental research (Gustafsson, 1984, 1992b, 1997: Kyllonen & Christal, 1990; Carlstedt, 1997) provide evidence in support of equating *Gf* and *g*. Gustafsson (1992b, 1997) describes Spearman's (1923, 1927) theory of *g*, which involved both a quantitative and a qualitative aspect. The qualitative aspect is expressed in terms of three principles; "eduction of relations" (rule inference), "eduction of correlates" (rule application) and "apprehension of experience". The first two principles aim to capture basic aspects of reasoning, while the third corresponds to what is now called meta-cognition. The quantitative aspect of *g* was formulated in terms of "mental energy", which should be understood as expressing individual differences in limitations on the ability to keep more than a limited number of items in mental focus at the same time (Gustafsson, 1997).

The definition of *Gf* in Cattell and Horns Gf-Gc theory share much of Spearman's definition of *g*, although they themselves describe *Gf* and *Gc* as a twin- form of Spearmans *g*. However, *Gf* is thought to precede *Gc*, "*Crystallised intelligence is a product over time of earlier fluid actions*" (Cattell, 1987, p. 94), indicating that *Gc* is dependent on *Gf*. According to the theory, *Gf* is thought to reflect biological and neurological factors, and factors such as incidental learning. *Gf* is thus thought to be unconnected with cultural skills "...[Gf] *rises at its own fate and falls despite cultural stimulus*" (Cattell, 1987, p. 97). *Gf* is thought to be best reflected in so-called "culture-fair" tests, i.e. in tasks reduced of cultural qualities, and in tasks that are novel to the person. Examples of tasks measuring *Gf* are analogies, classifications, matrices, series and topology-matrices.

As has already been mentioned several studies indicate that *g* and *Gf* are empirically and theoretically inseparable (Gustafsson, 1984, 1988, Undheim, 1981; Undheim & Gustafsson, 1987). Gustafsson (1997) concludes that this finding supports putting *Gf* at the apex of the hierarchy, which thus emphasises reasoning as the central component of intelligence. He also offer a rational for the existence and influence of *g* in any cognitive task in the following way:

> *"...there is a factor of general intelligence, which is equivalent with fluid ability, because every task, in one stage or another, involves a least a minimum of rule–inference. For example, vocabulary tests have been found to be highly loaded on the general factor, which is not because the process of retrieval an response require much of analytical ability, but presumably*

because the process of acquiring the meaning of words to a large extent is an inductive, rule-inference, process of learning from context" (Gustafsson, 1992b, p. 19)

Other research has indicated that the *g-Gf* dimension may be a reflection of "working memory" (Kyllonen & Christal, 1990). Thus, the principles of Spearman's theory have lately gained much empirical support, and the definition of *g* as an analytical non-verbal reasoning dimension now seems firmly established.

Broad cognitive ability dimensions

On the intermediate level a number of broad ability dimensions have been identified of which the most important one in educational contexts seems to be *Gc*, Crystallised intelligence. The term "crystallised" is meant to imply the "*freezing shape of what was once fluid ability*" (Cattell, 1987, p. 115). *Gc* like *Gf* also is thought to reflect the capacity for abstraction, concept formation, perception and eduction of relations. The difference is that *Gc* is associated with systematic influence of acculturation, and is central in tasks of a verbal-conceptual nature (Gustafsson & Undheim, 1996).

General visualisation, *Gv*, is another broad dimension spanning over a range of tasks with spatial content, a dimension which according to Cattell (1987) reflects good visualisation resources. Another broad dimension, *Gs*, General speediness, is thought to reflect speed and accuracy in cognitive performance, and *Gr*, General fluency, is thought to reflect retrieval from memory storage (Cattell, 1987). Later research has proposed a few additional broad dimensions (Carroll, 1993; Gustafsson & Undheim, 1996)

Narrow dimensions

At the lowest level of the hierarchical model a large number of narrow, specialised ability dimensions identified in the Thurstonian multiple factor tradition is identified. Examples of primary abilities are Verbal comprehension (*V*), Numerical ability (*N*), Cognition of figural relations (*CFR*), Visualisation (*Vz*), Spatial orientation (*Sr*) and Flexibility of closure (*CFR*). Narrow dimensions are thought to be determined by practice and experience as well as interest and motivation.

Choice of method

Theories of the structure of human intellect have throughout history been developed in rather close relation to factor analytic techniques. One may say that each new model in history relies on methodological advancements. The current three-level structure of cognitive dimensions has been identified with a refined form of exploratory factor analysis (Carroll, 1993), although a use of structural equation modelling (described below) yields more stable and precise results, and opens for further questions and analyses (Gustafsson, 1997). In my three studies the latter type of factor analysis and models was applied, and the methodology will be further commented on in relation to the results in the following sections.

There are several reasons for my choice of multivariate quantitative research methods for studying gender differences in patterns of knowledge:

- My ambition was to study complex patterns of multiple variables in terms of variation, structures and relations simultaneously, and there is no other method that can handle such a large amount of information.
- The great variation within each of the female and male groups overlaps to a large degree. Therefore a study of gender differences seemed best performed in the light of individual differences, in order to give a fair picture and avoid unduly polarisation. The need for large samples was obvious.
- Furthermore, large-scale studies have many other advantages: they are usually powerful, they tend to give a reliable impression if performed correctly, and they tend to influence researchers and decision makers as well as the public if the results are made known.

These are not unimportant reasons if one wants to provide information that may reveal structural patterns of differences between groups like gender, and if such information can be useful in the work of change towards gender equality. However, my choice of methods was primarily based on the research question.

Thus, the perspective and the approach I have chosen in my empirical investigations, has its roots in the latest advances within the psychometric tradition for investigating and understanding individual and group differences in cognitive functioning (Gustafsson, 1992b; Gustafson & Undheim, 1996). The power of structural equation modelling, for the study of cognitive performance has been demonstrated in studies that has adopted a multivariate approach with latent variables organised in accordance to the hierarchical structure described above (e. g., Gustafsson & Balke, 1993; Gustafsson, 1994; Härnqvist, 1997).

The same method has also been demonstrated to be useful and relevant for the study of group differences (Gustafsson, 1992a). The method assumes and investigates latent structures, and although the starting point for the analysis is sameness, this tool not only enables investigations of the question *if* these latent structures are the same, but also enables identification of *how* these structures may vary between the two groups.

In all three studies I have chosen to contrast this multivariate technique with traditional univariate methods, for two reasons. The first was to demonstrate the power of the method, which is not yet widely spread. The second reason was that the contrast in itself contributes additional information compared to that which either of the two types of analysis would have contributed.

As my research shows, when gender differences are put in relation to individual differences they are usually small. But sometimes this is deceptive, since small is easily confused with unimportant.

Some strong feminist voices have pleaded for a general rejection of quantitative methods as research tools (Reinharz, 1979; Fox Keller, 1985; Harding, 1986, Hallberg, 1992). As Reinharz (1992) asserts

> *One root of this criticism is hostility to statistics that are seen as part of the patriarchal culture's monolithic definition of 'hard facts'*. (Reinharz, 1992, p. 87).

Although others have pleaded against methodological essentialisms (e.g., Martin, 1994; Jayaratne & Stewart, 1995; Elisasson, 1987, 1994; Rosén & Wernersson, 1996; Wernersson & Ve, 1997) the limited number of feminist researchers in the field of educational measurement and psychometrics, bear witness of the persuasive power of arguments against quantitative methods.

Some of the critique that feminists and others have levelled against traditional research certainly is warranted. I am, however, not convinced that such a standpoint benefits feminist interests. I will, however, scrutinise the critique in greater detail in the last chapter of this essay

Methodological advances within performance research

Confirmatory factor analysis or, more generally, structural equation modelling (SEM), has its roots in the psychometric approach to the study of individual and group differences. Ever since its beginning, the psychometric approach has been relying on scores on quantitative scales. With the tools of factor analysis, these scales may be analysed so that the dimensions of knowledge and skills that form the structure of individual differences may be identified.

The explorative factor analytic approach has been criticised for failing to provide deeper theoretical insights into the nature of intelligence, since it fails to specify the processes by which problems are solved (e.g., Resnick, 1976; Sternberg, 1977). However, instead of abandoning the psychometric approach, which seemed logical and rational at the time, there is now a well-grounded confidence in the new possibilities that structural equation modelling offer.

In the following, I will try to describe the methodology without getting into all the technicalities. But first I will give a description of some of the central terms.

Manifest vs. Latent

In recent research on individual differences in knowledge and skills a distinction is made between the "*manifest*" and the "*latent*", both theoretically and methodologically. In theory, terms such as knowledge and skills refer to abstractions that are not directly observable, or *latent*. The manifest part refers to observable and observed behaviour, like achievement on tests, which may be viewed as indirect measures of latent dimensions.

The methodological or statistical meaning of latent and manifest variables is at a superficial level equivalent with the theoretical distinction. The observed or manifest variable is the one that "meets the eye", while the latent variable is disguised. Examples of manifest variables are, in the context of knowledge and skills, the *observed performance* scores on various types of achievement tests. Observed performance scores may be total test scores, sub-test scores or performance on particular tasks or items. In traditional univariate research, manifest scores are used as indicators of some more general and/or abstract cognitive skill, such as when performance on a vocabulary test is interpreted in terms of verbal ability.

In contrast, the latent variable is in a multivariate context part of a mathematical model, in which the latent variable accounts for the correlations that exist between performance on various cognitive tasks. The parameters of such models may be estimated with statistical methods, and the fit of the model may be tested. The model parameters (i.e. the latent pattern) may differ in various ways between groups, which may be investigated statistically. Such latent variable models may be built, analysed and evaluated with the aid of SEM (e. g., Jöreskog & Sörbom, 1993a, 1993b; Gustafsson & Stahl, 1997).

Modelling

Under any observed performance hides a latent pattern, a structure with relations to the manifest variables. When correlations between tasks are sorted out and transformed into latent variables, a latent variable pattern emerges. This "sorting procedure" is called modelling. The latent variables are then assumed, on logical grounds, to be closer to the theoretical constructs than is any observed manifest score. It may be observed that this "sorting" procedure is quite similar to the procedures for identifying theoretical constructs within qualitative research.

In SEM the researcher thus proposes a particular model with a particular number of latent variables (= factors), and also the pattern of relations between observed (manifest) and latent variables. The model is then estimated from data, which usually is the covariance matrix for all the variables. The process towards the final model is usually a stepwise procedure, because the originally hypothesised model often requires some modification. To evaluate the fit there are statistical tests which indicate whether the model fits the data or not, and so-called modification indices which provide suggestions for improving the model. It should, however, be noted that meaningful models cannot be constructed on the basis of modification indices alone. The process of model fitting, whether guided by modification indices or not, must be given a theoretical rational. The overall χ^2 goodness-of-fit test provides an evaluation of the fit of the complete model, and involves a comparison between observed covariances and the covariances implied by the model. If this measure is large in relation to the degrees of freedom, the model must be rejected. However, the chi-square measure is sensitive to large sample sizes, so other global fit indices have been developed. A broadly accepted measure is the "Root Mean Square Error of Approximation" (RMSEA) index, which provides a measure of model fit independent of the sample and model size. More detailed information about measures of the model fit is offered elsewhere (e. g., Gustafsson & Stahl, 1997; Jöreskog & Sörbom, 1993a, 1993b; Bollen & Long, 1993)

Rich information is gained from a structural equation model. In addition to observed scores, means and covariances may be estimated for each latent variable. Furthermore, the relations between latent and observed variables, the structure, is described by coefficients called factor loadings. These are regression coefficients, which, when standardised, are the correlations between latent and the observed variables. Similarly, a model may include relations among latent variables in a structural model. If groups are compared in a multiple group model the information from the analysis may be massive if there are differences with respect to many parameters.

Factor analysis in principle enables division of the observed variance into two parts, one that is due to the common factors and one that is unique to the observed variable. The unique variance may, in turn, be divided into two parts, one called "specificity", which is the reliable part of the unique variance, while the other part represents random sources of influence (measurement errors). All measures are more or less "polluted" with measurement errors, which are sorted out from the analysis.

Research utilising SEM and these measurement considerations, has demonstrated that observed measures, such as scores on vocabulary tests, are usually not indicative of only one underlying latent dimension, such as verbal ability but several (e. g., Gustafsson, 1984; Gustafsson & Balke, 1993). For example, scores on vocabulary tasks both measure a specific verbal knowledge, and more general cognitive skills.

These advances and findings within psychometric research indicate that when differences between females and males (or other groups) are observed on manifest scores, it may be due to differences in many different latent dimensions and to measurement errors. Furthermore, a similar observed score does not necessarily imply similarity between groups in underlying dimensions or latent patterns.

In my studies, for example, it is demonstrated, that the specific factors can cause the performance to be higher or lower than would be expected from the common factor, which is due to knowledge and skills that are important for a narrow domain.

Summing up, in this chapter I have presented some problems, the theory and methodology chosen for my three empirical studies. The principles of the method and the technique are quite accessible, but it takes some experience to conduct this type of analysis. This is obvious when results from this type of analysis are published. A substantial part of the articles are often concerned with all the technicalities, while a minor part addresses the substance. Hopefully, this will change, as the methodology becomes more established and used.

An investigation of gender differences in the latent patterns of knowledge thus coincided with my research ambition. Furthermore, as has been previously described, it seems also to meet a need that researchers within and outside the field have called for. The main focus for my studies included in this dissertation (part II, part III and part IV) are the patterns of knowledge that underlie observed performance scores, and in what way such latent pattern differ between males and females. In the following chapter, I will present and further elaborate on the findings from those studies.

The Studies

In the following sections the more specific purposes of each of the three empirical studies will be described, together with a somewhat extended discussion of results and issues related to theory, practice and contemporary research. General problems and purposes, theory and methodology have been described in preceding chapters. However, methodological points and clarifications are offered during the presentation of results.

Theories of gender and gender differences will be partially dealt with during the course of discussion, although some feminist socio-cultural reflections are saved to the final chapter.

A common goal for the three studies presented here have been to investigate observed gender differences in cognitive performance in terms of differences in underlying constructs of abilities and proficiencies. As the previous chapter aimed to clarify, the multivariate approach, using structural equation modelling (SEM), enables investigation of the relations between theoretical constructs of cognitive abilities, performance on operative cognitive tasks and their interrelations. Furthermore, many of the typical problems of measurement can be handled with the aid of SEM.

Patterns of knowledge in the two first studies, consists of an investigation of gender differences in various aspects of intellectual abilities. The difference between them is that the results in the first are analysed in more depth and challenged by the second in an analysis of missing data. Thus, the second study addresses the comparability problem and the problem of selectivity discussed in the previous chapter.

The third study aims to analyse gender differences in so-called "Document reading" between countries. Document reading is assumed to be an important part of reading skills, where the patterns of knowledge from these skills represent a mixture of broad and narrow abilities as well as "task-specific" knowledge in the theoretical model of intelligence. The study comprises an investigation of the influence of underlying dimensions, their relation to gender and how gender differences in patterns of knowledge vary between countries. Their connection to the hierarchical model of cognitive abilities is limited to a theoretical discussion. The influence of culture is demonstrated, which should be regarded as a contribution to the eternal question of gender differences as determined by culture or biology.

The three empirical studies are throughout this essay referred to as follows and fully presented in Part II, Part III and Part IV:

1. "The first study" or "The ability study" as an alias for the article Gender differences in structure, means and variances of hierarchically ordered ability dimensions (Rosén, 1995)

2. "The second study" or "The missing data study" as an alias for the article *Gender differences in hierarchically ordered ability dimensions: The impact of missing data* (Rosén, 1998)

3. "The third study" or "The comparative study" as an alias for the article *Gender differences in reading performance on documents across countries* (Rosén, in press)

Data background to the ability study and the missing data study

Both the first and the second study comprise an analysis of gender differences in cognitive abilities. The data analysed came from a project with the primary aim of analysing and describing individual differences in learning strategies. Central for the project was the question of the structure of human abilities and above all the question whether intelligence is unitary or multifaceted. For this purpose the designers of the project had selected a number of different cognitive tests that seemed to be the most promising variables previously developed in traditional differential psychological research. These tests were also intended as reference variables when evaluating new cognitive constructs. Another important aim for the project was to collect and document a reference material of tests and subjects to be used in further research of individual differences in learning as well as group differences. The project and the data collected is described in greater detail elsewhere (Gustafsson, Lindström & Björk-Åkesson, 1981) and for the present purpose I will only give a brief description of the data I used.

The sample consisted of 1224 students in grade 6 from 1980. It was more or less all the students in two communities on the Swedish West Coast. In the autumn of 1980, when most of the students in the sample had reached the age of 12, members of the research project administered a test battery of 13 ability tests to the students. The tests in the battery were selected primarily to capture the broad cognitive dimensions that theoretically are thought to be the most important ones for describing the individual variation in cognitive performance. The 13 ability test used all refer to three broad cognitive ability dimensions, General Intelligence/Fluid intelligence (*g-Gf*), General visualisation (*Gv*) and Crystallised intelligence (*Gc*).

In their regular activities in grade 6, schools were also recommended by the state to administer three standardised achievement tests (Swedish, English and mathematics), and which also represent measures of *Gc*. The three standardised achievement tests consisted of 4-7 sub-tests, and for the analysis in both the first and the second study, the sub-scores from these different parts were used. The total number of test-scores included in the analysis was 29. Altogether 981 students had completed all three standardised tests together with all ability tests, and those 981 students became my research subjects in the first study. The analysis in the missing data study comprised all 1224 students.

The ability study

The first study had three specific aims, of which one was to contrast the traditional univariate approach with a multivariate one. "Traditional" refers to the practice of comparing means and variances on one single observed test, sub-test or task score at that time. The multivariate approach used, considers several scores, and treats them as more or less interrelated in that they share latent sources of variance. With the aid of SEM, these latent sources can be sorted out as mentioned earlier, and interpreted as more valid measures of the theoretical constructs.

The second aim was to investigate the assumption of gender similarity in the hierarchical structure of cognitive abilities. The theory described above was thus translated with SEM into its empirical correspondence of a hierarchical latent variable model. With this approach the observed variance for all the variables was "sorted" into a number of broad and narrow latent dimensions for each observable task, along with the measurement errors. In the modelling of the data the theoretical model was thus fitted and adjusted to the data. Gender was included in the analysis after an acceptable model was arrived at. Gender differences were then investigated in terms of means, variances and relations between latent constructs and observed variables.

The third aim was to investigate if potential mean differences in latent dimensions were influenced by various model assumptions. The multivariate approach allows for investigation of questions, such as if and how the latent mean differences between the groups are affected if the variances, or the measurement errors, or the relationship between the observed and the latent variables are allowed to differ between females and males.

The first and most striking result was the contrast between the observed pattern of gender differences, e. g., the differences in test performance and the differences that occurred on the latent dimensions, which gave a completely different picture of the pattern of gender differences observed on the surface. In the following I will present and discuss the results study by study.

Contrasting the results of a traditional univariate approach with the results of the multivariate latent approach, serves two purposes. It distinguishes between *performance* on tests and the *latent cognitive dimensions* of the test that cause differences in performance. This means that common interpretations of test performance can be problematized not only theoretically but also empirically. It also demonstrates that both approaches are needed, since the relations between the observed and the latent contribute to a more complex understanding. Thus, for each study I will start with the results of the univariate analysis, move on to the multivariate analysis and then discuss the relation between the two. Enclosed are also a somewhat extended discussion about the meaning and implications of the results, compared to what is presented in the published articles.

In the final chapter I will further elaborate a socio-cultural explanation in the light of the gender system, for the gender differences in patterns of knowledge found. My reflections will hopefully serve as an invitation for further discussions and research.

Gender differences in observed performance

Mean differences in the observed performance (manifest scores), were, on the ability tests, apparent in 4 out of 13 tests, on 6 out of 7 sub-tests in Swedish, on 3 out of 4 sub-tests in English, and on 1 out of 5 sub-tests in mathematics, i.e. 14 out of the 29 different test scores. In most cases the difference was to the female advantage. Differences in variability in these observed measures were apparent in 5 of the ability test scores and in 2 of the sub-test scores in Swedish and English respectively. A greater variability was found in the male group in each of these 7 test scores.

While describing the observed differences more specifically I will also highlight that which a traditional single-variable approach would have indicated. Although connections to previous and contemporary research are made in each article, some references are made here to put the results into a larger context.

Gender differences on tests of Fluid intelligence

In a traditional analysis it would have been difficult to draw any strong conclusions on the basis of the results from the 13 ability tests. Three of them are considered to be typical tests of *Gf*, that is "Number Series", "Letter Grouping" and "Ravens Progressive Matrices". A minor male advantage on the first, and a somewhat larger female advantage was observed on the two latter. The inconsistent pattern makes any conclusion about an advantage in either direction doubtful. This is particularly so if *Gf* is interpreted to be "General intelligence" in the IQ-score meaning described earlier. Gender differences have not been expected on IQ tests.

When Binet and Simon (1905) constructed the first psychologically useful measure of intelligence, their belief was that there were no gender differences in general intelligence (Hyde, 1990), and their use of intelligence test did not reveal any either. However, as Hyde acknowledges, another interpretation is that the test constructors were successful in producing tests that confirmed their prior belief. It may thus be that the patterns of there not being any gender differences from the test, gives a misleading picture of reality. If similarity is a criterion of test validity, it certainly makes the question harder to explore. Hyde also comments that Binet and Simon's beliefs regarding females were rather unusual, given the popular views of the time (Shields, 1975).

According to Hyde (1990) it was Thurstone's conceptualisation of intelligence as being composed of multiple abilities (rather than one as Spearman suggested), that laid the foundation for research on gender differences in verbal, mathematical, and spatial abilities separately. It thus seems that the tradition of investigating gender differences on one single dimension at a time, without connecting different measures to each other, has its roots from this time.

Although tests of induction were developed early, it was not until later they were referred to as tests of Fluid intelligence (*Gf*), when Cattell and Horn's (Cattell, 1971; Horn & Cattell, 1966) hierarchical theory of the intellect was accepted as being more sensible and useful than the Thurstonian one. Test of *Gf* are, among others, the so-called culture-fair, e. g., the Ravens Progressive Matrices test. Explicit assumptions regarding gender differences in *Gf* or in performance from tests of *Gf* are not available, and early research on *Gf*-type tests did not reveal any consistent pattern of gender differences. In Maccoby and Jacklin's (1974) review of research on gender differences in cognitive abilities and cognitive styles, there are not any studies, which refer to Fluid intelligence explicitly. However, in their review there are three studies of Ravens Progressive Matrices test (I: age 7, 8, 11, sample size: 72; II. age 8-0, sample size 12 & 350; III. age 13 sample size 240), of which I and III did not show any gender difference while II showed a male advantage. Other "resoning" tests showed a mixed pattern of gender differences, but the most common conclusion was "no differences".

The idea of a dimension like "General intelligence" has now reappeared, at the apex in the hierarchical model described in the previous chapter. The methodological implications of the hierarchical theory of intelligence with three levels are, however, not yet widely adopted.

To conceptualise "general intelligence" as a composite of several dimensions, as represented in IQ-tests, is still common in contemporary research. Tests of IQ are also commonly used in public and popular media (mostly in the USA) as a reference to signs of general intelligence. However, to view general intelligence as a composite of a number of narrow ability dimensions is something quite different from the *g* (General intelligence) in the hierarchical theory. Two conceptions and applications of the same constructs undoubtedly cause confusion, and it will probably take a while before a terminology that differentiate between these two views is agreed upon among scholars and elsewhere.

The developers of IQ-like tests do however share the opinion with Binet and Terman, that there are not any gender differences in general intellectual ability, and the tests are thus constructed accordingly (Matarazzo, 1972, McNemar, 1942). The idea that there could be any "real" gender differences in *g* or *Gf* is not socially acceptable, and thus quite controversial to investigate. A female or male advantage in some tests of *Gf* would probably be interpreted as either accidental, caused by chance or by specifics in the test, or as a sign of maturity level.

Gender differences on Spatial tests

In my study a female advantage was found in 'Hidden Patterns' and 'Disguised Pictures', both spatial tests. Taking into account that none of the other spatial tests revealed any gender difference, a reasonable conclusion may have been that the consistent male advantage in this domain finally seems to have vanished, at least in Sweden. Questions regarding gender differences in spatiality has in the research community caused many animated debates, (e.g., American Psychologist, September, 1986). As Linn and Petersen (1985) report, spatiality is not a unitary construct, and the "consistent" pattern of male advantage reported through history (Anastasi, 1958; Maccoby & Jacklin, 1974) has not only decreased (Feingold, 1988; Emanuelsson & Svensson, 1990) but seems restricted to particular aspects of spatiality, as for example "spatial perception" and "mental rotation". On "spatial visualisation," Linn and Petersen conclude in their extensive meta-analysis of previous research, that the differences found were neither significant, nor meaningful.

There have been voices pleading that the differences found in spatial abilities are given disproportional attention with reference to their "smallness" (e. g., Chipman, 1988), while others have elaborated on what practical consequences small differences may have (Burnett, 1986). Others again argue that the differences are substantial (e.g., Sanders, Soares & D'Aquila, 1986). When females on the average perform better on spatial tests, the comments are generally few in the research literature. Some scholars have challenged the notion of gender differences in the spatial domain by elaborating on whether they really exist (Caplan, MacPherson., & Tobin, 1985; Halpern, 1986)

Gender differences on Standardised Achievement tests

The female advantage on the standardised tests in Swedish and English would confirm and validate the fact that females gain higher school grades. This would also confirm the rather consistent pattern of female advantage in verbal abilities reported over the years (e.g., Anastasi, 1958; Maccoby & Jacklin, 1974, Hyde & Linn, 1988, Willingham & Cole, 1997). However, these researchers also emphasise that the "consistency" of the patterns is dependent on how verbal abilities are defined, indicating that there are fields within the verbal domain where there are no gender differences to be found, or fields where sometimes the reverse pattern is shown. The issues for research on gender differences are now directed towards which types of verbal ability show these differences, and the developmental timing of the appearance or disappearance of the differences (Hyde & Linn, 1988). Male advantage on tests of verbal abilities seem only to occur when the male and female samples are selected from different groups, as in the case of the verbal scale on the American Scholastic Aptitude Test in 1985.

The almost gender-neutral result on the standardised test in mathematics from my first study, seems to support the conclusion that there no longer exists a male advantage in mathematical abilities. However, most (American) studies of gender differences in mathematics performance indicate a male advantage, which increases with age. This is particularly true on tasks that require more than computational skills (Fennema & Petersson, 1985, Willingham & Cole, 1997). Mathematical skills are presumed to be of particular importance for career development, which explains why mathematics and its relation to gender is another area that has caused many controversies among scholars and practitioners (Fox, Brody & Tobin, 1980; Friedman, 1989). Since the type of mathematical tasks on which males outperform females are tasks that are described as requiring "high cognitive complexity" (Fennema & Petersen, 1985), it is no wonder that the same idea of female intellectual inferiority, as was the common belief not long ago, reoccurs.

Reports of gender differences in mathematical skills have been criticised for utilising unreliable samples, male-biased performance measures and both incorrect and sexist interpretations. It seems that most of the differences reported are very small or non-existent if sampling and previously taken mathematics courses are considered (Peterson & Fennema, 1985). In their extensive review and meta- analysis of 100 studies, Hyde, Fennema and Lamon (1990) came to the conclusion that gender differences are small, and actually favoured females in young samples of the general population. However, differences favouring males emerged and increased with age and increasingly selected samples, and were largest in highly selected samples, and samples of highly precocious persons.

Global statements of consistent and/or robust male advantage in mathematical abilities made previously (Maccoby & Jacklin, 1974; Halpern, 1986) were thus misleading, but the problem of female under-representation in mathematical areas later on in life is not artificial. The critique has then turned the research question from "why *can't*" females do as well as males in mathematics?" to "why *don't*" females do as well as males in mathematics?"

(Willis, 1990). This turn of the research question implies that females abandon the mathematical field for other reasons than lack of ability. If in Sweden there was a female disadvantage in mathematics among 12-year-olds in 1980, this was not deducible from the observed score on the Standardised Achievement test in mathematics in my study. Females are still "underrepresented" in most educational programs where mathematics is one of the key subjects.

Variability differences between females and males in observed test scores

"The greater male variability hypothesis" has its origin in historical observation of a wider array of behaviour among males compared to females, accompanied with more sophisticated notions of anatomy and physiology. These observations were translated into a dual model of strength/weakness which was determined by nature (Hollingworth, 1914; Shields, 1975; Dijkstra, 1986; Noddings, 1992;), and used as an explanation for existing inequitable social circumstances. The functionalistic view of nature, introduced by Darwin, has its parallel in the understanding of sex differences. According to Shields (1975) Charles Darwin fitted the observed variability difference into his functionalistic evolutionary theory in *The Decent of Man*. It was his cousin Francis Galton who first claimed empirical scientific ground for the conclusion that women tend in all their capacities to be inferior of men (Galton,1907, referred in Shields, 1975). One of the many illuminating examples from Shields illustrates the logic of that time:

> *That men should have greater cerebral variability and therefore more originality, while women have greater stability and therefore more 'common sense', are facts both consistent with the general theory of sex and verifiable in common experience* (Geddes & Thompson, 1890, cited in Shields, 1975, p. 743).

Males were all of a sudden the more variable sex intellectually, enriched with many more virtues than the female sex, however female traits were also acknowledged in that "the female" was seen as complementary to "the male":

> *The presentation of evolutionary theory anchored in yin yang concepts of functions represents the most positive evaluation of the female sex offered by the 19th century science. Whatever woman's shortcomings, they were necessary to complete her nature, which itself was necessary to complete man's*" (Shields, 1975, p 746).

Females, who were seen as the opposite of males by default, were seen more restricted or even invariable intellectually:

> "*Women is a rule, typical; man, individual. The former has average, the latter exceptional features...there is incomparably less variation between women than men. If you know one, you know them all, with but few exceptions*" (Dijkstra, 1986, p. 129, citing Nordau, 1885).

Variability differences in test scores were, at this time, linked to these notions. Distinguished scholars, such as Edward. L. Thorndike, Havelock Ellis, James McKeen Cattell and Lewis Terman thought they had found the reason why females were missing among geniuses, and were under-represented in the upper half of the vertical division of labour and society. This was, and is, of course, a provocative hypothesis, and it has been heavily criticised from feminist scholars on both methodological and theoretical grounds (e.g., Hollingworth, 1914; Noddings, 1992). In light of the critique and reanalysis of data, several of these distinguished male scholars later changed their position;

> "*The woman who is a potential poet, novelist, lawyer, physician, or scientist usually gives up any professional ambition she may have had and devotes herself to home, husband and children. The exclusive devotion to domestic pursuit robs the art and sciences of a large fraction of the genius that might otherwise be dedicated to them. My data strongly suggest that this loss must be debited to motivational cause and the limited opportunity rather than lack of ability*" (Terman, 1975, cited in Hyde & Linn, 1986, p. 223)

The greater male variability hypothesis has gained mixed empirical support in modern research (Feingold, 1994; Willingham & Cole, 1997). When gender differences in variability occur in research, they need to be explained, however.

Common findings in contemporary American research indicate greater male variability in measures of science, history and civics, and in mathematical and spatial abilities, but no difference is found in verbal ability (Feingold, 1992; Maccoby & Jacklin, 1974, Willingham & Cole, 1997). In a Swedish study, a consistent pattern of somewhat greater variability has been found among males on one type of spatial test "metal folding" in grade 6, but not on the mathematical or the verbal tests (Emanuelsson & Svensson, 1989).

In the observed distribution of scores in my study, greater male variability was found from the tests of Fluid intelligence, in one short-term memory test, in most sub-tests in English, and in one sub-test is Swedish. When gender differences are found in the tails of the distribution, they only concern a very limited number of persons. Furthermore, these persons are the most deviants from the average within the group. However, the observed differences in the lower end of the distribution confirm reports of male students as having more difficulties in school subjects (e.g., Wernersson, 1989). A common hypothesis is that this has to do with maturity level and/or negative student-teacher interaction (Entwisle, Alexander & Olson, 1997).

The latent structure

The multivariate modelling part was the next step of the analysis described in the ability study. Before commenting on the pattern of gender differences in latent constructs, a few remarks will be made on the hierarchical model that was fitted to the data. In order to account for the individual differences in all of the 29 tests scores, 13 latent dimensions were introduced.

The statistical model was built in a top-bottom manner, which means that the first latent variable entered into the model was *g*, capturing the common variance across tasks. The broad and narrow dimensions were then imposed successively, and were formed by the residual variance, i.e. the remaining variance, after *g* and the other factors had been taken into account. A common way to mark that a latent variable accounts for residual variance is by entering a prime (′) after the variable label, like *Gc′*.

The model included three broad dimensions (*Gv'*, *Gc'*, and *Gs'*) all influencing a number of tests each, and there were nine rather narrow dimensions of specific knowledge or skills influencing particular types of tests. The model shows that all the observed performance measures were accounted for by more than one underlying dimension.

A simplified illustration of the empirical model is presented in the figure below. Squares represent manifest variables while circles represent latent variables. The arrows indicate the direction of the relations. The prime denotes residual variables, and latent variable labels are presented in italics.

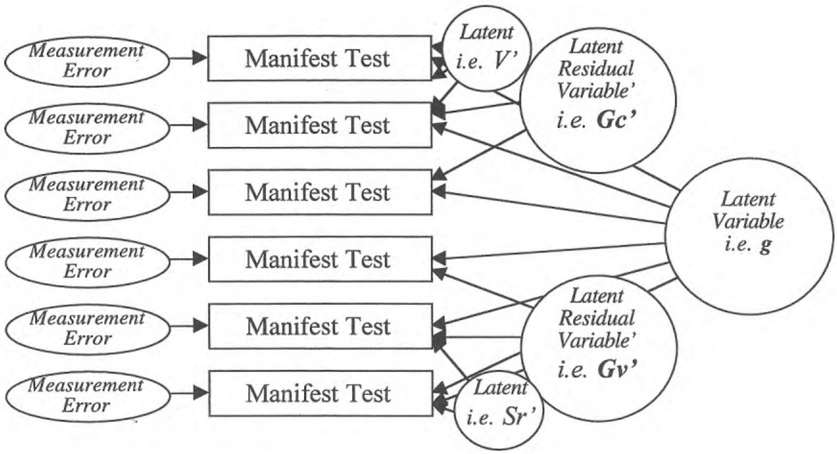
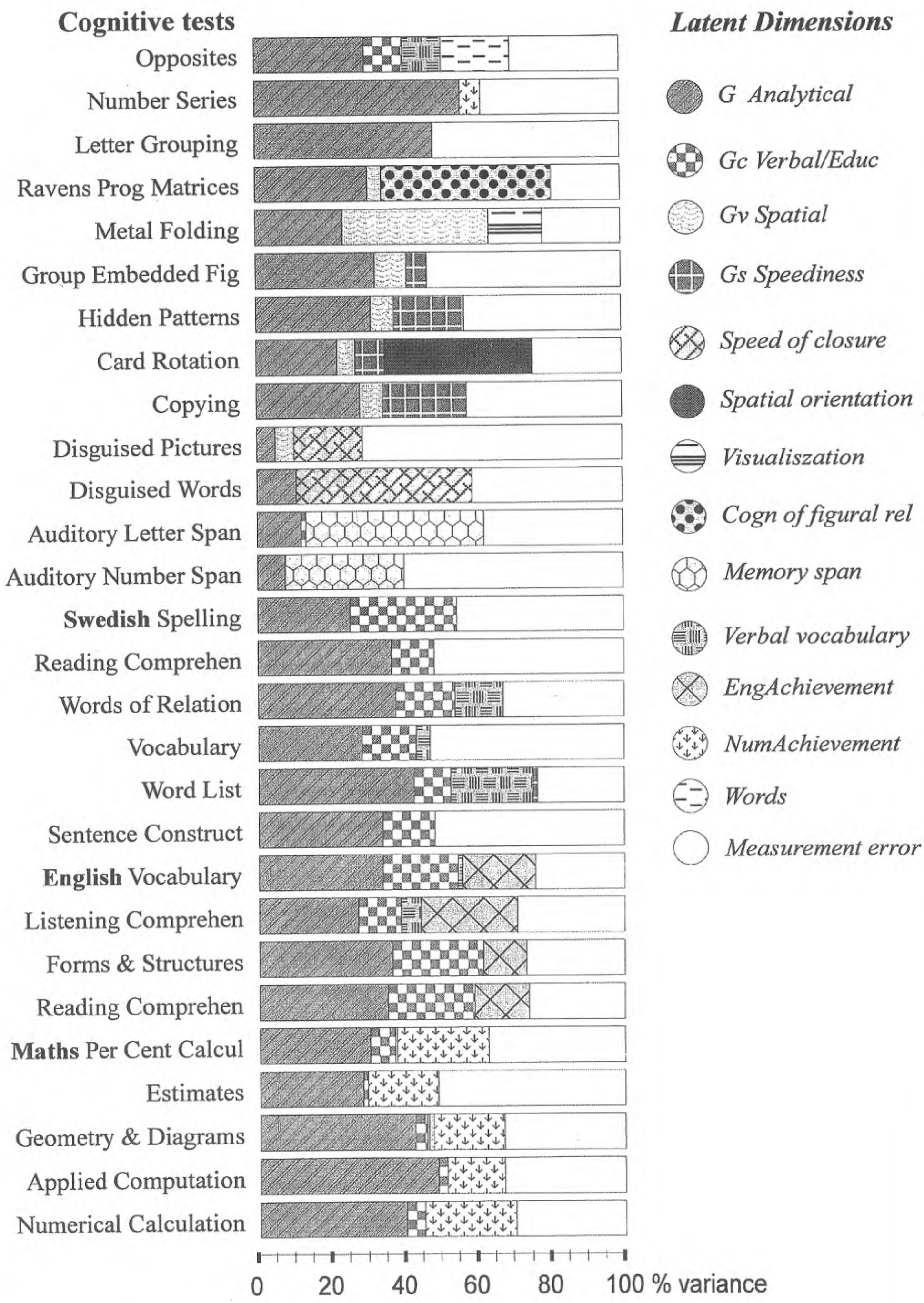


Figure 2: A simplified nested factor model.

The model is a so-called nested-factor model. A nested-factor model expresses the multiple influence of several latent dimensions on each single observed variable (Gustafsson & Balke, 1993). First there is the general factor, influencing *all* the observed scores. Then follow the broad factors, influencing a wide array of tasks that shares some characteristics, but not all. Finally, there is a number of narrow factors influencing a limited set of task that share some specific characteristic

The amount of variance accounted for by each ability dimension in each manifest test score is illustrated in the graph below. Each bar represents the total observed variance (100%) in each test. The proportions of variance that are explained by the different latent ability dimensions are represented by the different pattern within each bar. For example, the observed variance in the spatial Metal Folding test was determined to be 24% by G, 38% by Gv' and to be 15% by Vz', while about 23% of the variance was due to measurement errors. Another example is the sub-test Estimates in the Standardised Mathematics test, were 28 % of the variance was accounted for by G, 1 % by Gc and 19 % by *NumAch'* factor, and 52 % was due to measurement errors.


Cognitive tests
Opposites
Number Series
Letter Grouping
Ravens Prog Matrices
Metal Folding
Group Embedded Fig
Hidden Patterns
Card Rotation
Copying
Disguised Pictures
Disguised Words
Auditory Letter Span
Auditory Number Span
Swedish Spelling
Reading Comprehen
Words of Relation
Vocabulary
Word List
Sentence Construct
English Vocabulary
Listening Comprehen
Forms & Structures
Reading Comprehen
Maths Per Cent Calcul
Estimates
Geometry & Diagrams
Applied Computation
Numerical Calculation
0
20
40
60
80
100 % variance
Latent Dimensions
G Analytical
Gc Verbal/Educ
Gv Spatial
Gs Speediness
Speed of closure
Spatial orientation
Visualiszation
Cogn of figural rel
Memory span
Verbal vocabulary
EngAchievement
NumAchievement
Words
Measurement error

It is noteworthy that *g* is about equally influential in each sub-test of the standardised achievement test, with on the average 35% of the variance. *Gc'* on the other hand, which is assumed to be the typical school achievement factor and strongly influential on school performance, accounts for between 11 and 30% in the language tests, but only for between 1 – 6 % of the variance in the mathematics sub-tests. This pattern may, however, be due to the fact that the content in the mathematics tests was basically numerical, i.e. the verbal content was in many of the tasks quite limited. On the other hand, mathematics training in school takes place in a verbal cultural context, so one might have expected more influence from *Gc* as proposed by the Gf-Gc theory. Perhaps *Gc* becomes more influential as the students grow older and better trained in mathematics.

Gender differences in the latent variable model

In latent variable models gender differences can be investigated with respect to several aspects. First of all there are a number of questions that concern model fit. Does the model fit the data from each group equally well in terms of fit indices? Are the patterns of relations between observed and latent constructs the same in the groups? Is the strength of the relations equal between groups? If the model is completely different from one group to the other, a comparison between groups in terms of means and variances in latent dimensions would be meaningless. If there are only minor differences in the model between the groups, then it may be investigated if and how these aspects affect the pattern of differences in latent means.

The patterns of relation between latent constructs and observed test performance were similar for boys and girls, although the strength of a few relations differed somewhat. The differences were, however, so few and small, that the same pattern has to be validated in more studies before one can draw any conclusions. These differences did not, however, affect the means on any of the latent dimensions in any substantial way. This study thus supports the hypothesis that the structure of human intellect is the same for males and females

Gender differences in latent ability dimensions

The pattern of mean differences between boys and girls in latent constructs was strikingly different from the pattern in observed scores. Girls had a substantially higher mean on General intelligence (*g*) and on the broad Crystallised intelligence *(Gc')* dimension. Boys on the other hand showed a somewhat higher mean on the broad spatial dimension General Visualisation *(Gv')*, and on some of the narrow spatial dimensions as well (Speed of Closure (*Cs'*) and Spatial orientation (*Sr'*)). Boys also had remarkably higher means on the narrow dimensions of Numerical Achievement (*Num Ach'*) and Verbal-vocabulary (*V'*). Gender accounted for some 30% of the variance in the *NumAch'* factor and some 20% in the *V'* factor.

The rather large difference in *g* to the female advantage was quite unexpected, given the above-mentioned assumption of equality in this respect. When gender differences in this broad intellectual ability have been reported, it has only been on the basis of performances from IQ-scores and similar composite scores (Feingold, 1992). Such types of tests have almost always shown a male advantage whenever gender differences have been found (Willingham & Cole, 1997). The male advantage has, however, only been found in selected samples, which makes comparison between females and males in this respect more complicated to interpret. This is also a circumstance that easily is lost in the public debate, and thus supports beliefs of male superiority in this respect.

General intelligence in the modern theoretical sense is not easily captured by any single test, but instead is always involved in all test performances as demonstrated here. Lack of valid measures is therefore one of the reasons why such findings have not been reported previously. The empirical support for *Gf* being inseparable from *g* described earlier has helped in defining the construct of *g*, why more studies of gender differences in this respect can be expected in the future. Investigations of group differences in *g* thus require tests of *Gf* to differentiate between General intelligence and any general dimension. As Gustafsson (1994) has pointed out, it is impossible to differentiate between *g* and *Gc* in IQ- tests unless some of the tasks are measuring *Gf* exclusively, and the variance decomposed in a latent variable mode, similar to the one in this study.

The fact that history lacks findings like this also makes the result somewhat doubtful. There is, however, one parallel study that shows similar results. Härnqvist (1997) too found a female advantage on *g*. One possible explanation for the female advantage has to do with the age-group that is investigated (12-13 year- olds) in the present. This advantage may be a reflection of differences in the process of development, and if the onset of puberty also involves a dynamic phase in terms of intellectual development, then females at this age will be ahead. Ljung (1965) suggested that the earlier onset of puberty among girls may be accompanied by an earlier spurt in mental growth for females than for males. However, Härnqvist (1997) investigated gender differences in *g* and other ability dimensions in a representative Swedish sample of intact school classes in grades 4 through 9, and found an increasing advantage on *g* for girls. One may ask at what age a mental spurt for boys is to be expected, and if such a spurt really would remove the difference. Härnqvist's analysis comprised the ages 10-16, and the difference did not decrease, but the level on *g* decreased for both sexes in grade 9. The mental spurt hypothesis has not gained much empirical support, but one reason for that may be the lack of longitudinal measures over a wide age range. Another problem is the gender differentiation that increases by age, which makes the two groups difficult to compare.

As has earlier been mentioned, gender differences are not expected on so called IQ-tests. At some point in the history of test development it was decided that intelligence is equally distributed between the sexes at each age level. Reasons for this rather radical turn is probably that the differences found did not make any sense or were politically uncomfortable since the test were used for ranking and selection. The rationale for adjustments of the tests was that there should not be any gender differences in tests of general intelligence. Gender differences in tasks were then cleaned out by either removing those items or tasks that showed a difference, or matching it with an item or task that showed a similar difference in the opposite direction (Matarazzo, 1972; McNemar, 1942, Halpern, 1992). To hide or remove gender differences in this technical way does makes it difficult to investigate if there are any differences between groups. The matching strategy still permits investigation of gender differences, although not with univariate methods. The exclusion strategy makes analysis of gender differences meaningless. It should again be noticed that IQ-tests are not a valid measure of the type of General intelligence referred to in my study. Gustafsson (1994) has found that although *g* is highly involved in IQ scores, such tests also involve a fair amount of *Gc*.

A related reason why gender differences in *g* have not been investigated, or reported, is also the multivariate hierarchical approach, which is not yet widely spread. This may be due to the fact that the model has just recently been empirically validated by an extensive reanalysis of most previous test data (Carroll, 1993). Another reason may be that these types of analysis are quite complex and have been rather tedious to perform. This is all changing now, as the usefulness is demonstrated and as more user-friendly software becomes available.

The female advantage on the broad Crystallised intelligence dimension (*Gc'*) was more expected both from previous research and from the data reported here. Females have consistently been reported to have been awarded higher average school marks (Willingham & Cole, 1997; Emanuelsson & Fischbein, 1986) which together with the average higher performance on most of the verbal school achievement test makes this advantage reasonable. In my study the observed female advantage on most sub-tests of the standardised achievement tests in Swedish and English is a reflection of this *Gc* advantage. In the verbal tests, *Gc* and *g* together account for the main part of the variance. The female advantage is thus due to their relative advantage on both these broad latent dimensions. However, in the more narrow dimensions of the school achievement tests, males were on the average substantially stronger (*NumAch'*, *EngAch'* and *V'*).

The observed pattern did not indicate any difference in mathematics, and thus supported the feminist notions that female inferiority in this respect is nothing but a "mathematical myth" (Fox, Brody & Tobin, 1976). Although boys and girls performed equally well on the mathematics test in this study, the difference in the more narrow underlying numerical construct needs some explanations.

Males and females may use different approaches in solving the tasks. It seems that male performance to a higher degree is explained by their proficiencies on narrow dimensions, while female performance to a higher degree is explained by their proficiencies on broad dimensions. In a qualitative study of how Swedish pupils at the upper level of compulsory school meets chemistry, physics, and technology in the classroom interaction (Staberg, 1992), different learning styles were noted. It was concluded that boys have a practical approach to science, whereas girls have a theoretical approach. Girls' learning style was in the study characterised by *work* and boys' by *play*. Perhaps the male advantage on this narrow dimension is a mark of female subordination in the quantitative area.

The feminist literature has repeatedly been pointing out that the quantitative area is a well-established male domain where females are more or less expected to be - and treated as if they are - less successful regardless of performance, (Fox et. al., 1976; Chipman, Brush & Wilson, 1985). Staberg found that boys' interest and learning styles were favoured, "*..often because they manage to masculinise the content and to influence the teaching methods*" (Staberg, 1992, p. 180).

A less conflict-oriented explanation may be that the differences in the underlying pattern are an expression of differences in interests and focus. Another of Staberg's conclusions was that girls prefer knowledge connected with their own and others' lifes, while boys are interested in things and in making things. This does not mean, of course, that any of the suggested explanations rules the other out. The question is what consequences these differences underlying mathematics performance may have. Although mathematical tasks in grade 6 may be solved with different approaches, more advanced math may perhaps require more specialised skills. It is thus of great interest to investigate how these differences change with development, if and how they affect student perceptions of self, self- confidence, interests, and future choices in terms of education and occupation.

If the quantitative domain is established as a "male domain", the verbal domain is an equally established "female domain", so the huge male advantage on the narrow Verbal dimension (*V'*) was even more surprising. *V'* is interpreted as a narrow verbal factor involved in tasks that require some specific vocabulary knowledge. In the present test battery *V'* was involved in six of the ten language sub tests and in the Opposites ability test. The proportion of variance accounted for by the factor varied between 1-22%. The female advantage in observed performance on these tests is due to their advantage on the broad dimensions of *g* and *Gc* rather than to their level on the more narrow verbal-vocabulary dimension. The most reasonable explanation of the male advantage on *V'*, is that the tests were developed under the awareness of historically consistent female advantage within this domain, why the choice of words in these tests more or less unknowingly has been selected to compensate for this expected difference (Härnqvist, 1997).

The male advantage on the narrow English Achievement dimension (*Eng Ach')*, although not as large, is equally unexpected. The factor accounted for 12- 26 % of the variance in the English tests, while *g* and *Gc* accounted for the main part'. On all the observed scores girls performed significantly better than boys, which again has to do with their advantage on *g* and *Gc'*. The male advantage on the *EngAch'* factor may perhaps be attributed to the content of the tests, i.e. a similar hypothesis to the advantage on *V'*. Another hypothesis to account for the male advantage in these narrow verbal dimensions may be the test situation, which can be interpreted as a competitive situation. There have been some reports which indicate a larger male preference for competitions in the classroom setting (Staberg, 1992; Wernersson, 1989, 1992a; Willingham & Cole, 1997).

On the broad spatial dimension *Gv'*, and on the more narrow spatial dimensions of "spatial orientation" (*Sr'*) and "speed of closure" (*Cs'*) a higher male average was found. Again, in the observed pattern of performance the differences were not seen, a pattern that is consistent with contemporary feminist research. The historical and consistent male advantage in the spatial domain thus reappears in the latent pattern. Gender similarity may thus have been reached in solving spatial tasks, but this is obviously not the same as gender similarity in spatial abilities.

In this study a substantial male advantage on several latent spatial dimensions was found, and some hypothetical explanation for this advantage is suggested further below. Again one must understand that the lack of difference in the observed pattern is due to the female advantage in *g*, which accounts for substantial parts of the variance in these tests. Different approaches for solving the tasks are again hypothesised, since it is obvious that spatial tasks can be solved with the aid of other abilities than spatial.

These findings of mean differences in various ability dimensions between the groups, invites to reflect upon the dynamics between various dimensions in the hierarchical model, why this will be done before discussing the results of the variability differences found in latent dimensions

Theoretical implications I - some reflections

The pattern of gender differences in the underlying dimensions relative observed performances raises questions regarding the dynamics between *g*, the broad dimensions of *Gc* and *Gv* and the more narrow dimensions. Female performances are, on the average, better than males on most of the cognitive tests, which is due to their relatively higher level on broad ability dimensions.

In the Cattell and Horn Gf-Gc theory these broad dimensions are assigned relatively greater importance than both other broad dimensions, and the more narrow ones in the hierarchical model, and they also account for the larger part of the variance in each task.

However, looking at the male group's profile regarding spatial dimensions, the pattern of a higher male average on both the broad *Gv* dimension and on several of the narrow dimensions suggests that these may be of equal importance for solving cognitive tasks.

If one assumes that *g* or *Gf* is a prerequisite for developing more narrow dimensions as the Gf-Gc theory suggest, one interpretation of the pattern is that males overachieve while females are held back in developing proficiencies in narrow dimensions given their general intellectual capabilities. However, it is also plausible that various proficiencies develop more independently of *g* and Gc. Perhaps the development of more narrow dimensions to a larger degree are influenced by interest and motivation, than by "higher-order" thinking skills, at least at this age.

Females equal performance on mathematics tests, and males superiority on the *Num Asch'* dimension, indicate that the two groups may focus on different objectives during their lessons in mathematics. Females seem to nurture their general intellectual abilities, while males seem to nurture their numerical achievement skills. Staberg's (1992) study of how males and females confront the science subjects in school provides some support for this hypothesis, where females were found to seek more "connected knowledge" and expressed an urge to "really understand", while the males seemed to treat the science subjects in a more "playful" way.

From the theory of *g* and Gf-Gc one may argue that females ought to perform better in mathematics relative to their level in *g* and/or males ought to improve their general intellectual ability relative to their performance in math. From this pattern there is at least two questions for future research: (1) What is the dynamic between *g* and narrow math abilities; and (2) is it true that females and males partly nurture different types of abilities while solving mathematical tasks? The first question requires longitudinal data, while the other probably requires other research approaches to be explained. However, if females underachieve relative to their general capacities, then there is a support for the feminist arguments, that females are not provided equal opportunities within male defined subject domains.

Looking at *V'* and *EngAch'*, narrow factors that are involved in the language tests, the pattern of gender differences is even more spectacular, males outperforming females on the latent verbal dimensions while females are outperforming males in observed performance. If the relation between general and specific ability dimensions is so hypothesised, the females should perform even better on these tests. From classroom research we know that males get more attention, different types of attention and more of the teachers' time, i.e. they tend to dominate the classroom (Wernersson, 1989; Staberg, 1992). However, little empirical evidence exists which indicates that this inequality pays off in any cognitive sense. The results from this multivariate analysis may, however, be interpreted in such terms. As mentioned before, the male advantage on these dimensions may be explained by the choice of content in the test.

The findings in this study do not provide any information concerning the relation between broad and narrow ability dimensions. One way to investigate the development of different cognitive profiles may be to investigate groups which are high and low respectively on broad dimensions, and analyse how they are related to the levels on other abilities, and preferably on longitudinal data (see Sandqvist, 1995; Andersson, in press).

Variability differences between females and males in latent abilities

The pattern of variability differences was in this study also remarkably different from that in manifest scores. Differences in variability were found in 3 of the 13 latent dimensions: *CFR'*, *Ms'* and *Gv'*. Boys' variability was about 45 % larger than girls' in the two narrow dimensions and approximately twice as large in the broad spatial dimension. The Ravens Progressive Matrices test also showed a variability difference, but not nearly as large in the latent construct of *CFR'*. The Raven Progressive Matrices test is the only indicator of the narrow dimension *CFR* but also an indicator of Fluid intelligence, which makes it difficult to interpret the manifest variability difference. The latent analysis, however, indicates that the observed variability difference is due to variability differences in spatial dimensions rather than in *g*. The magnitude of the difference was not obvious from the observed pattern. The variability difference on the observed Auditory Number Span test matches the variability difference in the latent Ms' variable although the latter is larger. For the Auditory Letter Span test, however, there is not any trace of a similar difference, in spite of the fact it is supposed to reflect the same narrow dimension. Thus, it is difficult to generalise from these two observed variables.

For *Gv'*, there is nothing in the comparisons of the observed scores that indicates such a large gender difference in variability. Greater variability within a group compared to another indicates longer "tails" in both ends of the distribution of scores, although the debate about gender differences usually is restricted to the difference in the upper tail. There is now a growing concern about the boys in the lower end of the distribution, although the debate has been couched in terms of unfair treatment of all boys in school (Entwisle, Alexander & Olson, 1997)

As mentioned earlier gender differences in variability has a long history surrounded by controversies of different kinds. Greater variability among males has often been suggested to have a predetermined biological basis, although always without direct empirical support. Today, this historical belief is translated into what seems to be an eternal research question. Feingold (1994) argues that cross- cultural comparisons are of critical importance for conclusions about causality between biology and variability. Consistent patterns of equal differences over nations are considered to support biological theories, while a pattern of inconsistent differences across nations would support cultural theories. Feingold's meta- analysis of test performance data from many countries indicates that the only thing consistent is the irregularity of variability differences between males and females. Given numerous methodological considerations, Feingold concludes that gender differences in variability are likely to be due to cultural factors or to the interaction between cultural and biological factors rather than to purely biological factors.

In my study greater male variability was found in about 1/3 of the observed scores, most of them on the verbal school achievements test, and the difference in most of the cases were not very large. On the latent dimensions the variability differences were restricted to the spatial domain.

Hypotheses on spatiality

In terms of explanations of gender differences in variability in spatial abilities, research has so far not been very successful. However, I will suggest some different hypotheses to account for the pattern found.

Variability differences were only found with respect to the spatial abilities. Males have an average advantage on these dimensions, so it may be that males get better opportunities to develop these abilities. However, the within-group differences are large, which indicates that this is true for a subset of the male group only. The hypothesis suggested is thus, that "spatial activities" are more common in male settings than in female ones, but that these male settings do not include all males. There is reason to believe that these settings are constructed outside the school curriculum. This is supported by the substantially higher male average on latent spatial dimensions, while there is a consistent female advantage in almost all academic school subjects, as reflected in the school marks (Emanuelsson & Fichbein, 1986).

Furthermore, in a previous study of the same data analysed here (Gustafsson & Balke, 1993), an analysis of the ability dimensions' predictive power with respect to school achievement showed only very weak relations between a spatial dimension and school marks. The only school marks for which *Gv* showed any predictive power were technology, which is a typical male area of interest, handicraft and drawing.

Other research shows that boys more than girls enjoy competitive games (Staberg, 1992), and "single-minded" devotion to a particular task (Eccles, 1987), which also may be of some importance in accounting for this pattern. The hypothesis thus is that proficiencies in spatial areas are developed through typical male activities outside of school, activities which are restricted to particular groups of males.

Another explanation is that the variability difference is a reflection of maturity differences. At this age level boys are on the average less mature than girls, and it is reasonable to assume that this type of visual skills develops in the course of physical maturation. There are voices in the literature, which stress that many of the problems young boys have in the early school years are due to their lesser maturation, and which may have severe effects on their perception of the self and on the schoolwork later on in life (Entwisle, Alexander & Olson, 1997).

However, this pattern of both gender differences in means as well as variability differences may well change as more spatial tasks are implemented in schoolwork. An example of such possible change is the growing use of computers in the school environment.

The usual assumption in large-scale studies is that the missing data patterns are random across groups. It is true that groups with missing data often are weaker achievers than those with complete data, and particularly so when the data is obtained from a "high-stake" test situations as the standardised achievement tests are. Fortunately enough, the database used in the present study includes information enough to allow further analysis of the selectivity hypothesis. Thus, in the second study, the impact of missing data on the model and on the gender mean differences in latent dimensions was investigated.

The missing data study

The first study revealed a pattern of differences quite unexpected from the test performances as well as from previous research. Aware of the common critique from feminists and others that a substantial part of the research that has been reporting gender differences in cognitive performance, has been biased in terms of selective samples and/or in the choice of operative measures, the same suspiciousness should be directed towards my study. It thus became a concern to find out whether the results in my first study to any degree could be attributed to unintended selection mechanisms. Since the first study was based on the part of the sample that had complete data on all the tests only, one potential source of bias was the exclusion of individuals who were lacking all or some of the three standardised achievement tests.

The study relies on a methodology for investigating or controlling for, missing data in the same multivariate fashion as was used in my first study, which has been developed by Muthén, Kaplan & Hollis (1987). It thus became my project to challenge the results of my first study by applying this new methodology on the same set of data. However, as was demonstrated in the first study, the contrast between traditional univariate analysis and the multivariate approach in itself provided valuable insights and it also demonstrated the power of the multivariate approach. It thus became a separate aim to contrast the two approaches again.

As emphasised in my second study, missing data is not merely a statistical problem: the phenomenon can also be addressed from a more substantial point of view, and thus contribute to the understanding of gender differences and related issues. The data that some students in my sample were missing were, one, two or all three standardised achievement tests (in Swedish, English and mathematics). These tests are indicators of school achievement, designed as an aid for the teacher to calibrate marks, and thus a direct "high-stake" test for the students, and indirectly for the teacher. School achievement is typically influenced by factors at different levels, such as the individual level, the classroom level, the school level and so on. Since none of the sampled schools in my study were missing, possible sources of missingness were missing individuals and missing classes. It is a well known fact that less successful individuals are more likely to avoid test situations than others, but as long as the "bias" is the same in all groups, it does not cause any problems in interpretations of group differences. In my first study, as in most other studies, the missing data distribution was assumed to be equal over the two groups, and thus treated as random. If this assumption is incorrect, the results of the first study may be more or less invalid. Both the pattern of gender differences and the hierarchical model itself may be affected by a non-random missing data pattern.

There are several reasons why individuals and classes may lack data. Valid reasons such as "normal" illness, usually results in random attrition in collected data. However, there are a number of less innocent reasons why some individuals do not show up or respond to standardised achievement tests, such as test anxiety, lack of motivation, lack of sense of meaningfulness, parents' expectations, teachers' and/or class mates' expectations and so on. All of them may contribute to systematic errors in the data. There are also a number of reports of gender specific expectations as well as theories of gender stereotyped behaviour, e. g., low self esteem or low performance confidence in particular school subjects, learned helplessness, test-anxiety, and resistance strategies (Staberg, 1992, Wilkinson & Marrett, 1985; Wilder & Powell, 1989; Öhrn, 1990). Such behaviour may, in turn, result in a missing data pattern that is different for boys and girls, and thus affects the analysis and the interpretation of group differences.

When complete classes are missing, other mechanisms may be involved. About half of the missing individuals belonged to missing classes. The tests are designed to help teachers assign comparable grades. One reason not to participate may thus be fear of exposing the achievement level of an individual class. Another reason can be ideological reasons of the teachers' or the school principal. Parents may also be influential in such a decision if they, for one reason or another, would perceive the tests to be unfair for their children. Missing classes does not necessarily result in biased samples from a general or a gender point of view.

There may, however, also be gender-related mechanisms of attribution at the class-level. The frequently reported observations of male dominance and females' lesser visibility in the classroom is one basis for that. It may thus be that in some classes the male performance more than the female performance serves as a basis for teachers' decision whether to participate or not.

All the above examples show that there are theoretical as well as substantial grounds to investigate the potential impact of missing data on the pattern of group differences. The missing data generating mechanism may be interacting with gender, and several underlying mechanisms at different levels may be working simultaneously. It is thus not unreasonable to suspect that any kind of missing data may be biased with respect to gender. The implications are, however, not easily predictable, so missing data analysis seems important.

Impact of missing data on manifest performance

The first examination of data concerned the different types of attrition, and the proportion of boys and girls in each category. About half of the attrition came from missing classes, and the other half from missing individuals. About 55 % of the students in the five missing classes, and the same proportion among missing individuals, were females. The number of missing females in the whole sample was 121, and the number of missing males was 122. The observed means and standard deviations on the 13 ability tests in the reduced group with complete data compared to the complete sample were more or less identical, so the assumption that data is missing at random in the first study seemed well grounded. However, when both attrition and gender and gender-attrition interaction were investigated, with respect to the observed performance scores, the result showed a significantly lower performance level among the attriters ("attriters" are the 243 student that did not have the complete data, "completers" are the 941 students with the complete data) on all but four spatial tests. This pattern is, as mentioned earlier, quite common in the social sciences. However, the gender- attrition interaction analysis indicated a deviating pattern of gender differences among attriters compared to the gender differences found among the completers on three of the 13 ability tests. The gender differences in means within the attriters group were, on the three tests of fluid intelligence, larger and to the female advantage. The mean differences on the rest of the test were similar to those of the first study.

With respect to observed variability, differences were found both within and between each sex when attriters were compared with the completers. However, the pattern of differences was hard to interpret in substantive terms since some of it was assumed to be random effects. The overall pattern indicated that the male attriters were less successful on most of the tests compared to the female attriters. This also implies that the observed variability difference may have been underestimated, since the less successful males were missing. This observed missing data pattern could have influenced the pattern of differences in underlying constructs.

In the next stage of this study, class level attrition and individual level attrition were compared against the completers in terms of means and variances, first for the total group and then for females and males respectively. The overall conclusion was that class level attrition did not substantially contribute to any bias, while the attrition on the individual level did. Female attriters did not deviate from female completers, indicating a random missing data pattern for females. However, male attriters deviated quite substantially from male completers, indicating that the overall male performance in the first study was slightly over-estimated compared to the female performance.

Impact of missing data on latent patterns and gender differences

In the multivariate missing data analysis the group with missing data is included in the analysis so that the maximum likelihood estimates are based on all available information. This procedure results in more correct and reliable estimates than does the analysis where cases with missing data are excluded. Again, the multivariate analysis showed a pattern that was not predictable from the observed pattern in the univariate analysis.

The first question was whether the hierarchical model was affected in any way. It was found to be affected, although not in any substantial way. There were no indications of a pattern of relations between observed and latent constructs different from that obtained in the preceding study. There were, however, some differences in the strength of some of the relations. The influence of *g* increased somewhat on almost every test, as did the influence of *Gv'* on all the spatial tests. The opposite pattern was found for *Gc'* and most of the narrow dimensions, whose influence decreased on their target tests. The variance in the latent variables *g* and *Gv'* increased slightly, while the variance in the rest of the factors remained stable or decreased inconsiderably. The overall deviations from the original model in the first study were systematic but small, and did not alter the interpretation of the relations between manifest scores and latent dimensions.

Gender differences in latent constructs were affected in the following ways. The female average advantage in *g* increased, as did the average male advantage on the spatial dimensions *Gv'*, *Sr'* and *Cs'*. Gender now accounts for 1 % more of the variance in those dimensions. The male average advantage on the narrow *V'*, *NumAch'* and *EngAch'* dimensions decreases similarly as the female average advantage on *Gc'*. The amount of variance explained by gender drops hardly at all in *Eng Ach'*, to 6 % in *Gc'*, to 28 % in *Num Ach'*, and to 17 % in *V'*, indicating that the magnitude of gender difference is not dramatically altered in any latent dimension. The influence of the missing data analysis on gender differences in latent dimensions in the first study is summarised in the table below:

Table 2. Summary of results from the missing data study

Latent Ability dimension	**Gender difference in the first study**	**Impact of missing data on the gender difference**
g - General Intelligence	Female advantage	Increased advantage
Gc' - Crystallised Intelligence	Female advantage	Decreased advantage
Gv' - General Visualisation	Male advantage	Increased advantag
Gs' General Speediness	No gender difference	No gender difference
Sr' Spatial orientation	Male advantage	Increased advantage
Cs' Speed of closure	Male advantage	Increased advantage
Vz' Visualisation	No gender difference	No gender difference
CFR' Cognition of figural relations	No gender difference	No gender difference
Ms' Memory Span	No gender difference	No gender difference
V'- Verbal-Vocabulary	Male advantage	Decreased advantage
EngAch' English achievement	Male advantage	Decreased advantage
Word' - Word	No gender difference	No gender difference
NumAch' Numerical Achievem.	Male advantage	Decreased advantage

My own expectations before the analysis were that gender differences would decrease and perhaps disappear altogether in some of the latent dimensions. Again the results were something of a surprise, Thus, the pattern of mean differences in observed scores changed differently than the pattern of mean differences on the latent variables, when cases with missing data were accounted for. In the univariate regression analysis, when attrition and gender-attrition interaction were taken into account, the pattern of gender differences changed in favour of females on almost all observable performance scores. This made sense, since the female attriters performed equally well on all of the ability tests as did the female completers, while the male attriters performed far worse than did the male completers. The missing females thus seemed to match the missing at random assumption, while the missing males did not. From the changes in the observed pattern one may have suspected that the female advantage on *Gc'* would increase, while the male average level on most latent dimensions would remain the same, and that the level on the achievement factors (*V'*, *EngAch'* and *Num Ach'*) would decrease. The pattern of differences on the latent dimensions changed however differently.

The female advantage on *Gc'* decreased while their average advantage on *g* increased. Male average advantage on the more achievement oriented factors *V'*, *NumAch'* and *EngAch'* decreased a little as expected, while their average on the broad *Gv'* factor and the narrow spatial dimensions of *Sr'* and *Cs'* quite surprisingly increased. In a way this pattern supports the idea that spatial abilities are developed outside of the school context, and particularly so in male activities. If lower achieving boys resist school more than other students, and particularly so at "high-stake events" such as the Standardised Achievement test-situation, it may be perceived to be, and also contribute to a higher average on spatial dimensions when included in the analysis, it could be an indication that they develop these skills elsewhere.

Despite the fact that those missing males who were included in the analysis in general performed at a lower level on most tests than the male completers, the gender difference found in the three spatial dimensions increased. This again raises the question of the relation between *g* and spatial abilities. However, the dynamics of the relation between latent ability dimensions can only be understood in analyses of longitudinal data. New methods have been developed to investigate growth in latent variable models (Muthén, 1997) which are particularly valuable when longitudinal data from at least three time-points are available.

The general findings from the missing data study were that the pattern of difference changed, despite the almost invisible deviation of mean differences between boys and girls in the small group with missing data, compared to the larger group with complete data. This result implies that missing data analysis may be of importance to understand gender differences and to identify and control for differential selection mechanisms, even in cases when the attrition seems trivial.

Theoretical and methodological implications II – some reflections

This result raises questions about what impact a gender-biased attrition may have had in other studies of gender differences in cognitive performance. The most surprising result is the increase of the male average on spatial tasks, despite the lower performance in the male attriters group. Another rather peculiar result was the decrease of the female average advantage on *Gc'* despite the fact that female attriters did not deviate in terms of observed means and standard deviations from female completers. The increase of the female average advantage on *g* is more understandable, since the inclusion of less highly performing males lowered the average level for the male group. It seems to depend on two co-operative phenomena. One is the increased influence of *g* on all tests when the information from the attriters' performance is included in the analysis, and the other is the representation of different cognitive levels and profiles in the male and female attriter group.

The impact of missing data on variability differences was not investigated in this study. Such an analysis would require a multivariate multiple group analysis of males vs. females with complete or incomplete data respectively. At the time the analysis was performed, the technique was rather complicated and time consuming, and the analysis was thus not performed. Today, however, more user- friendly software for undertaking these types of analysis is available, which makes such analyses more easy to perform. The more general research question of whether gender differences in variability on observed measures completely or partly is due to selection mechanisms is, however, valid and is an urgent issue for future research.

Leaving the effects of missing data and returning to potential causes for differential missing data patterns for females and males in this particular data set, there are a few issues to be discussed. In 1980 when the data was collected, every grade 6 class in Sweden was encouraged to administer the three standardised achievement tests in order to help the teachers to calibrate grades that were comparable over classes and schools. The classes were, however, not forced to take the test, although most of them did. Performance on these types of tests can also be viewed as a measure of the teachers' degree of success, therefore low achievement in a class could be one of the reasons why a teacher decides not to give the test. Another possibility is that it is classes with many low achieving boys that are missing, which would support the notion of missing classes as a sign of the influence of the male norm in the classroom. None of these hypotheses received support in the present study, although boys in the missing classes performed relatively lower on two of the ability tests. It was rather the missing individuals than the missing classes that contributed to the bias.

Since these types of tests, i.e. standardised achievement tests, tend to be very important, primarily for the students since they form one of the bases for their grades, there are at least two plausible hypotheses for why low achieving boys, and not to the same extent low achieving girls, are lacking this type of information. Perhaps weak boys are more or less explicitly encouraged to stay away when the test is administered. Perhaps such signals are directed more to these boys than to similarly performing girls, because of the boys' relatively higher visibility in the classroom. It is well known that boys, and particularly low achieving boys, are more disturbing in the classroom than females. Since this may affect the class' result negatively, the students may be aware of it, and the teacher may feel the possibility to give high grades to high achievers to be unduly restricted. The hypothesis of the male performance as a norm for the class level is thus applicable here too.

These findings support the idea of differential selection mechanism for females and males, and indirectly the male norm hypothesis. Perhaps the pattern that low achieving males drop out while low achieving girls do not, applies to most of the high stake tests. The selection mechanism to the Swedish Scholastic Aptitude test for example, indicates that those who chose to take the test are high achieving boys while the high achieving girls chose not to, to the same extent (Reuterberg, 1998a). Further research is needed to investigate this phenomenon in order to understand and perhaps alter the observed behaviour.

This study demonstrates that analysis of missing data is not only needed for a more reliable analysis, it may also provide valuable information for the understanding of gender differences in performance as well as in different performance situations. Missing data analysis may thus also be considered as an analysis of contexts, which in turn, are often omitted in large-scale studies.

In relation to the variability discussion, the usefulness of comparative studies was stressed, to investigate notions of biologically determined differences between females and males. Consistent patterns of gender differences over time and countries has made both researchers and others to suggest biological explanations, while other argues for socio-cultural explanations. In the third study this issue was investigated in a comparative analysis of gender differences in patterns of knowledge on a particular type of reading tasks.

The comparative study

The results from the multivariate analysis approach adopted in my two previous studies indicate that reanalysis of gender differences may contribute deeper insights regarding gender differences in performance. When the IEA Reading Literacy data became available for further analysis, it captured my interest for several reasons. The primary one was the opportunity to address the question of social and cultural influences, and investigate how the multivariate approach adopted in previous studies would contribute to our knowledge of gender and gender differences. Furthermore, it allowed me to address some of the concerns listed in the background description of my studies.

Multivariate analysis of comparative data may provide relevant information to the nature/nurture debate. It is a fairly common belief that the patterns of cognitive functioning are more or less universal, and when gender differences are found it is easy to maintain such assumptions. It has been argued that cross-cultural studies are needed to *"buttres biological (or evolutionary) explanations for gender differences*" (Feingold, 1994, p. 82). SEM makes it possible to separate common latent sources on the one hand, and measurement error and specific task variance on the other. Such analyses may uncover gender differences in underlying sources across countries.

Furthermore, gender differences have been found to change with development, and the most common pattern is that they increase (e. g., Willingham & Cole, 1997). Generally there is little evidence of when they begin and why and how they develop. There is, however, some evidence that they change from one time to another (Emanuelsson & Svensson, 1990: Feingold, 1988). The IEA reading literacy data consisted of cross-sectional data from two age groups, which would allow for some insights on the subject.

Throughout history test validity has been, and still is, a major problem for the identification of "true" differences in cognitive functioning between females and males as well as other groups. Test bias is an eternal problem in educational measurement, and sometimes reported gender differences have been demonstrated to be constructed by the researcher, or researchers have been accused of doing so. The choice of content in the tasks may give one sex an unfair advantage over the other, thus creating differences which have nothing to do with the targeted proficiency.

Various kinds of cognitive tests have sometimes been shown to "unfairly" benefit or misjudge one group over the other (e.g Block & Dworkin, 1976; Gould, 1981; Tavris; 1992; Mirza, 1998). The problem is often linked to the researchers' interpretation of the pattern. When for example Benbow and Stanley (1980) stated that superior male mathematical ability is the best explanation for sex differences (which they claimed to be biologically determined) on SAT-M in a sample of highly gifted seventh- and eight-graders, it had an unfortunate impact on the societal beliefs of gender differences, and on attitudes towards mathematics. Their conclusions formed headlines in public media like: "*Do males have a math gene?*" *and "The Gender Factor in Math. A new study says that males may be naturally abler than females*" (referred in Eccles & Jacobs, 1987, p. 351).

The scholarly critique was massive. Some argued that the SAT-M was not justifiable as a measure of mathematical aptitude (Schafer, 1981). Eccles and Jacobs (1987) in their critique concluded that Benbow and Stanley's conclusion was premature at best, but very influential on students and parents attitudes in the subject of gender and math. Later, Benbow & Stanley (1987) changed their position and concluded that the reasons for the sex differences they found on the SAT-M test were unclear.

As has been demonstrated in the first two studies, observations on a single variable cannot easily be translated into judgements about level on particular ability dimensions. In fact, observed scores are, according to the latent variable model, always due to more than one underlying dimension. The approach used in the third study may on the surface look different from the one in the two previous studies, but it is in fact quite similar. With SEM the variance on different tasks can be divided into two major latent sources, latent variables and measurement errors, and in the previous studies this was extended to a hierarchical model, with latent dimensions of different degrees of generality.

A similar approach may be used to analyse scores on different tasks that all aim towards the same proficiency (e. g., Document-reading proficiency). Thus, even a simple one-factor model may be regarded as a hierarchical model where the residual variance is not modelled. The measurement error of the manifest variables may thus be interpreted as a mixture of specific task variance and random error. Specific task variance may be due to the test-takers previous experience of similar tasks, i.e. rather concrete knowledge, and not on more general and abstract proficiencies. If there are enough tasks and there are residual correlations after the effect of the general factor has been taken into account, these may then be entered as more narrow dimensions and may perhaps be interpretable in terms of test characteristics that are shared by some but not all tasks.

Gender specific patterns over and above the intended construct are also of great interest for understanding differences between females and males. Since task scores have been shown to be rather rich of information regarding underlying dimensions other than the targeted, there may be additional signs of gender related differences, that may or may not be shared between nations. Tasks that use examples from the animal world may perhaps suit girls better, while a task that uses examples from the technical world are perhaps better suited to boys. What type of content that suits one sex better than the other may also vary across countries.

The relation between content and gender can be connected to theoretical notions of "the male world and the female world" as an expression of different experiences, which has been suggested for example by Ve (1982, 1991), Gilligan (1982) Tannen (1990), Wernersson (1992a, 1992b) and Staberg (1992) as signs of sex-stereotyping. By analysing gender differences in the specific task variance, it will be clear if the content in some tasks appeals more to one gender. Analysing the variance that is usually referred to as test-bias may thus contribute to the identification and understanding of gender related activities, gendered cultures and/or sex-stereotyping.

The IEA Reading Literacy data

The IEA reading literacy data used in the third study was collected 1990/1991 by the International Evaluation of Educational Achievement (IEA) and comprised two age-groups (9-year olds and 14-year olds) in some 30 countries. For the present study I selected 25 and 22 countries respectively. "Reading literacy" is defined by three categories of reading material, which are assumed to require different types of skills or knowledge. Narratives and Expository texts are known from previous studies and definition of reading literacy while Documents is a rather newly defined category (Elley, 1994, see also Barr, Kamil, Mosenthal and Pearson, 1992) in this line of research. The constructs are defined as follows:

Narrative prose: refers to continuos text materials in which the writer's aim is to tell a story whether fact or fiction.

Expository prose: refers to text materials designed to describe or explain something, factual texts.

Documents: refers to structured tabular texts such as graphs, tables, charts, forms, labels, lists, and sets of instructions where the reading requirement typically involves following directions and locating information, rather than following a continuos text. Documents often require the reader to handle spatial layouts, and furthermore, often involve the requirement of processing numbers.

As previously mentioned, many studies have reported gender differences, which indicate male advantage in spatial and numerical domains, although not consistently so. In reading proficiency however, a rather consistent female advantage has been established. Reading is thus as established a female domain as mathematics is a male domain. The inclusion of Documents may therefore change the pattern of gender differences in reading performance, at least on the surface.

The expanded definition of reading literacy may be viewed as an attempt to even out the presumed female advantage in reading abilities, that seems so persistent. It may also be viewed as a more realistic definition of "reading literacy" since much information in modern society has the features of Documents.

Document tasks thus have features that often have been found to suit the male population, since the content is often spatial or numerical (Elley, 1994, Wagemaker, 1996). From my two previous studies it was learned that such tasks are also determined by other ability dimensions than spatial and numerical ones. Furthermore, the content in traditional test-items that uses stimulus material such as maps, tables and charts, often seems to reflect male interests more than females. However, in the original IEA study, Document reading was assumed to require special skills different from those required on Expository and Narrative texts. The tasks were also chosen with particular care, so that gender differences would not emerge due to the content (Wagemaker, 1996). Items were thus selected with an eye towards gender differences, and "male tasks" were balanced with "female tasks".

It is thus impossible in a univariate analysis to determine whether observed gender differences, if any, are due to general document reading skills, or to the content features of the task. Similar gender differences across countries in a univariate analysis may in a multivariate analysis reveal that in some countries the difference is due to the content of specific tasks and in other countries to differences in Document reading proficiencies. Such multivariate analyses may also contribute important information relevant to the nature/nurture debate. These lines of reasoning form the background to my third study, in which the Document tasks in the IEA reading literacy data were analysed in greater depth with multivariate methods.

The specific aim of the third study was to describe and analyse gender differences on Document tasks, and to investigate if and how the pattern of differences varies over countries. Another aim was to compare the multivariate approach with a univariate approach.

The Document construct was represented by a number of passages that differed in content and form. Each passage was followed by three to seven questions, which were added up into passage scores. These passage scores were then utilised in the analysis. In the univariate analysis these scores were examined both as sub-scores and added together as a single index of Document reading proficiency. In the original IEA-study the total Documents raw score was transformed into so-called Rasch scores (Elley, 1994). These are assumed to be more reliable than are the manifest raw scores. However, they are univariate measures too, which do not allow for multidimensionality.

In the multivariate analysis a model was fitted to the passage scores. First a general dimension was entered into the model, which influences performance on all the passages (General document reading, *Gendoc*). The remaining variance in each passage, formed latent passage specific dimensions. The passage Maria's timetable, for example, was thus decomposed into two latent sources, *Gendoc* and *Maria'*. The model below illustrates the main model for the 9-year-olds.

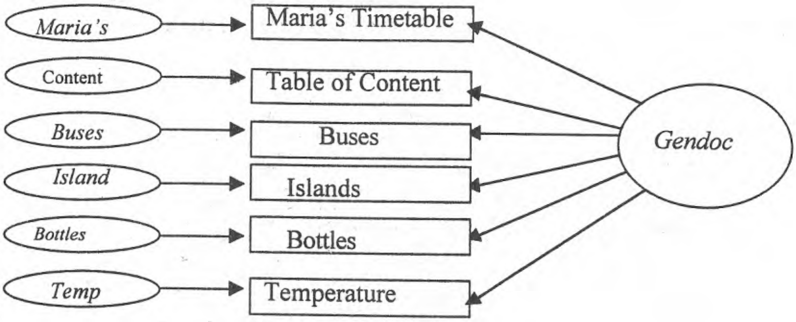


Figure 4: Latent variable model for Pop A, 9-year-olds.

For the younger group the model included one general document reading factor, along with six narrow latent passage-specific dimensions. Gender differences were investigated in each of these latent dimensions for each country. The results indicated a mixed pattern of gender differences across countries. In both the manifest Document score and the *Gendoc* dimension, the most common result was no gender difference. When a gender difference was found, it was usually to the female advantage. However, comparing the univariate pattern with the multivariate, it was not always the same countries that showed differences, the differences were not always in the same direction, and they were not always equal in strength. The difference in results between the approaches are illustrated in the table below.

The multivariate analysis also revealed a pattern of gender differences on the specific task dimensions, which indicates that some of the tasks give higher results for females, others for males, while others again are gender neutral. This pattern illustrates how important the content of a task may be for how well a group succeeds. Although a completely consistent pattern of gender differences was not found on any passage specific dimension, the overall impression was that there were two "female" contents, two "male" contents, and two gender neutral contents. The more "girl-ish" content required the student to "locate lessons on a school time table" and "to identify authors and page numbers in a table of content", while the more "boy-ish" content required the student to "locate points on a simple map" and "to locate time and trends on a temperature chart over 4 days".

The main model for the 14-year-olds is presented in the figure below. As can be seen, nine Document passages were included in the analysis.

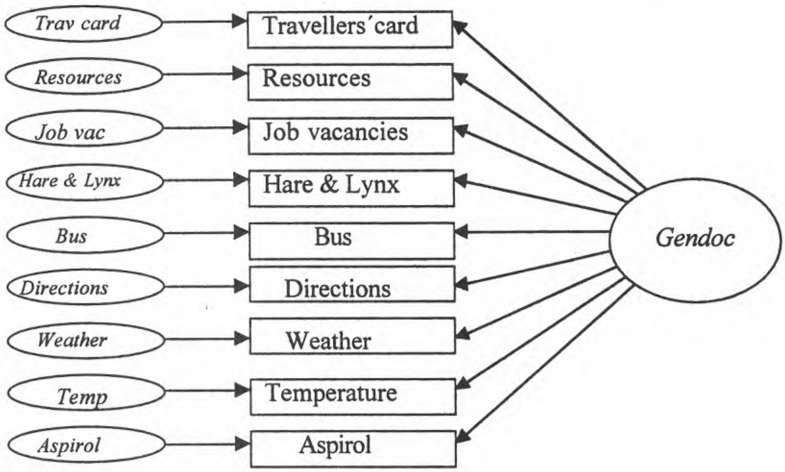


Figure 5: Latent variable model for Pop A, 14-year-olds.

For the older group the model building phase resulted in a model with one general document factor, *Gendoc*, and nine narrow latent passage-specific dimensions. However, in several countries the data indicated that the model required an additional factor in order to obtain an acceptable fit. This factor was related to all the passage scores of either the first or the second test booklet. This indicates that it may have something to do with the test administration, because the two booklets were administered on two separate occasions.

The univariate examination yielded a much more mixed pattern of differences in the total Document score, about a dozen of the 22 countries showing a gender difference. In about half a dozen countries a higher male average was found, and almost as many countries showed a higher female average.Differences were found somewhat more frequently with Rasch scores than with raw scores, and male advantage was more common when Rasch scores were relied upon.

A comparison of results can be studied in the table below for both age-groups.

Table 3 Gender differenes in Document Reading Skills across countries. A comparison of results of different analysis approaches.

Country	9-year-olds			14-year-olds		
	Latent Doct	**Rasch Doc**	**Manifest Doc**	*Latent Doc*	**Rasch Doc**	**Manifest Doc**
Belgium F	M*			F*		
Canada BC				F*	F*	F*
Denmark	F*		F*	F*		
Finland						
France				**Country not included**		
Germany East				F*		
Germany West				F*	M*	
Greece				F*	M*	M*
Hong Kong	F				M*	M*
Hungary				F*	F*	F*
Iceland	F*		F*		M	F*
Ireland						F*
Italy				M*	M*	
The Netherlands				F*	F*	F*
New Zealand	F*	F*	F*	F*	M*	M*
Norway	F*		F*	M*		
Portugal				M*	M*	M*
Singapore	F*	F*	F*	F*	M*	
Spain	M*	M*			M*	M*
Sweden				F*	F*	F*
Switzerland	M*				M*	
Trini. & Tob.	F*	F*	F*	**Country not included**		
USA				F*		
Venezuela		F*	F*	**Country not included**		
Slovenia	F*		F*	F*	M*	

* = non-random difference at the 5%-level

[ ] = No gender difference [F*] = Female advantage [M*] = Male advantage

Again, it was not always the same countries that showed differences, they were not always in the same direction, and they were not always equal in strength. As in the two previous studies, the observed pattern can be better understood when underlying dimensions are investigated, and measurement error sorted out.

The analysis of gender differences in the specific task dimensions indicates, that the observed performance differences to a large degree were due to differences in these dimensions. Again, a consistent pattern of gender differences was not found in any passage, although two of the passages tended to show female advantage and three passages tended to show male advantage. The more "feminine" content required the student to " fill in a traveller's card for a person, using already provided information" and "to answer questions about use, dosage and composition of Aspirol tablets". The more "masculine" content required the student "to locate time and trends on a temperature chart over 4 days" and "to use a key to locate resources in a detailed map" and "to interpret a bar graph showing population frequencies of hares and lynxes over 10 years". The rest of the passages showed few differences and mostly in both directions. On the latent booklet-factor, gender differences were found in eight countries, all but two to the male advantage.

The first issue to discuss is gender differences in "Document reading". The multivariate analysis revealed a pattern of gender differences across countries that was mostly to the female advantage. This may seem odd since this type of tasks is thought to require different skills than narrative and expository prose, due to its spatial and numerical content. However, the instructions to most of these tasks are in written language, so general reading ability may play an important part in solving these tasks as well. Therefore, to differentiate between Document reading and other types of reading, both types of tasks are required in the analysis.

In the hierarchical model of cognitive abilities reading is identified as a first-stratum ability below *Gc* (Carroll, 1993). Reading is, however, also described as a product of cultural evolution and relying on cultural transmission for its continued existence (Lundberg, 1991; Lundberg & Hoien, 1991), and it has an interactive and constructive nature (Lundberg, 1991). From this description of the skill it is reasonable to assume that reading performance is partially influenced by *g*, to a larger extent by *Gc'*, and by several narrow dimensions depending on the type of reading task. Document reading should also be influenced by spatial and numerical dimensions. However, to analyse Document reading tasks in terms of ability dimensions, indicators from a broad array of different ability dimensions are needed.

Gender difference on the Booklet factor' were found in many countries and to the male advantage. The tentative interpretation of the factor makes it difficult to draw any conclusions. However, if it is a reflection of the test situation, one possibility is that the factor is a reflection of a competitive climate, which has been suggested to appeal more to males than to females

(Staberg, 1992; Wernersson, 1989). Why this would be important for one booklet and not the other is however not easy to explain.

The passages with "masculine" content tend to represent traditional male interests, such as natural science domain, while the "feminine" content tend to reflect common female interests and responsibilities in both age groups. Gendered pattern like this may be interpreted as signs of sex-stereotyping. Females are supposed to take care of the daily practical and caring activites, while males are supposed to engage in more scientific issues, which also are reflected in their interests (Ve, 1982).

If the content in these passages were put on a traditional value scale, the male field would probably be considered as more valuable. An obvious sign of this are the efforts to persuade females to enter the traditional male domains of natural science, mathematics and technique. Another sign is the much larger emphasis that the current curriculum puts on these areas, compared to earlier versions (Ds, 1994). These are also the domains that are considered productive, while the female domains are considered reproductive or consuming (SOU, 1998:6).

The analysis of the two age groups imply that differences are not equally common in the younger populations as in the older, a pattern which matches observations made in numerous other studies. This phenomenon is, however, quite general for any kind of gender differences. The older we grow the more pronounced gender differences become, in preferences, in expectancies, in opportunities, in salary, in positions, in power and so forth (SOU 1998:6)

The nature/nurture controversy

Halpern (1992) states that the nature/nurture controversy has been debated for over 2000 years without resolution because it is essentially unanswerable. Yet, it has served as a framework for much of the research of sex/gender differences. One of the major sources to the controversy is socio-biology, i.e. the systematic study of the biological basis of all social behaviour. The discipline attempts to use evolutionary principles to explain the behaviour of humans and other animals. According to socio-biology more or less all behaviour is genetically determined. Womens' subordinate position as well as the gendered division of labour may thus be regarded as natural according to this theory. This position has, of course, been heavily criticised by feminists and others, by both empirical evidence and logical arguments (Bleier, 1984).

Another source of this controversy may be traced to the heredity question within behavioural genetics research. One of the major questions for this tradition is to investigate the relative influence of environment and heredity. Since differences in behaviour between the two sexes have been observed, the same question has been formulated for gender differences. Thus, questions like "are the cognitive differences due to factors that are inherent in the biology of maleness and femaleness, or due to differential sex-related experiences and expectations?" and "does nature or nurture play the greater part in the differences of the study?" are basic research questions (Halpern, 1992).

However, the way they are answered may have severe implications for different groups, as it has had for example on racial groups (Block & Dworkin, 1976; Gould, 1981; Mirza, 1998). As Halpern (1992) acknowledges, like all loaded questions, the answers sometimes backfire. "*Results could be, and have been, used to justify discriminations and/or affirmative actions based on sex*" (p. 5). This has led feminist researchers to regard "nature" as a political strategy of those committed to maintain the status quo of female subordination.

For individual differences very few modern psychologists would deny an interaction of biological, psychological and social influences on cognitive abilities, but for differences between groups, such as males and females, the influence of biology is strongly challenged (Shields, 1975, Bleier, 1984).

Cross-cultural studies of gender differences are thus of critical importance for conclusions about causality. Biological theories would predict that gender differences would be consistent over a wide range of cultures and countries, whereas socio-cultural theories would predict variations in gender differences over diverse cultures and nations. The multivariate analysis of gender differences in all latent dimensions of Document reading tasks is far from consistent across countries. Both age groups showed gender differences in both directions on the *Gendoc* dimension. About half of the countries among the younger students and about one third of the countries among the older students did not show any gender difference on the *Gendoc* dimension. Further, the analysis revealed a gendered pattern of influences from the content of the tasks on the observed performance. The empirical pattern suggests that socio-cultural explanations are considerably more reasonable than any biological or evolutionary theory.

With this lengthy discussion of the results in my three studies in relation to the social history and contemporary research I believe I have demonstrated the importance and the complexity of the subject and the need of continued research on gender differences in patterns of knowledge in various ways.

Reflections on Results and Methodology

This final chapter has two sections. In the first, I will reflect upon the results from my first study from a socio-cultural perspective, where feminist notions of the gender system are used as an analytical tool to interpret the gender differences found in my empirical studies. It will also be shown that my studies may contribute to the understanding of the gender system.

All my studies rely on large-scale data and statistical methods, and since most feminists seem to have abandoned such research conduct, I feel there is a need to scrutinise the feminist critique towards quantitative methods in some more detail.

Connections to socio-cultural observations

The empirical models in my studies are in essence system models, which enables connections to other system models such as the gender system. Hirdman (1993) defines the term "system" as a dynamic structure, a network of processes, phenomena, perceptions and expectations, which through its interrelations mediates patterns and regularities. By applying a "birds-eye-view" on complex observations, as cognitive performance, patterns of knowledge emerge. Applying the two logics of the gender system, i.e. the distinctive separation of the two sexes and the hegemonic masculinity, may add to the understanding of gender differences.

The overall pattern in the first two studies indicates that boys and girls at the age level studied seem to have developed somewhat differently with respect to cognitive dimensions, which in turn may be interpreted as a result of different approaches towards the learning environment. Despite more or less similar observed performance girls seem to a higher degree to nurture their general cognitive capacities, while boys seem to a higher degree to nurture specific capacities tied to particular subject domains. Boys thus seem to specialise, while girls seem to develop capacities useful for managing broader fields of cognitive problems. Other feminist researcher (Ve, 1982; Staberg, 1992; Eccles, 1987) have reported similar types of observations. The hegemonic masculinity principle emphasises the power relation, which indicate female behaviour always is valued in the light of male behaviour. Thus, specialised cognitive abilities are higher valued than broad abilities.

A common finding in the research body of gender differences in the classroom interaction is that boys tend to dominate the classroom scene verbally, physically and socially (Spender, 1982; Kelly, 1986; Wernersson, 1977; Wilkinson & Marrett; 1985, Öhrn, 1990; Staberg, 1992). Some of the general results are summarised by Wernersson (1992a):

- Boys talk more, ask more questions and "show off" in different ways, and thus demand more attention.
- The teachers attention is more directed towards boys. They get more help, more questions and more elaborated comments on their different initiatives. They also get more reprimands for disturbing behaviour.
- The content of lessons and teaching methods is chosen more often in accordance with the interest and needs of boys
- Girls, on the other hand, behave more as "responsible students", which the teachers uses to keep track in the lesson and sometimes to discipline the boys.

One may thus conclude that the gender system is apparent in the classroom setting. However, as Wernersson (1992b) argues the effects of the male dominance are unclear, particularly since girls tend to get higher grades and perform better on most school achievement tasks. Common hypotheses are that the male dominance causes females to develop less effective social skills and a negative self-concept. Their better grades has often been negatively explained by "pleasing the teacher" syndrome, rather than by superior knowledge (Wernersson, 1989; 1992b).

However, male dominance may have a cognitive price, if males due to the fact that they receive more educational resources, are able to develop specialised skills, whereas girls in their subordinate position does not get that opportunity. Their different cognitive profile may also interact with the fact that boys are more visible than are girls in the classroom. It could be that females broad and even cognitive profile make them perceived as (too) "all-round" and broad in their interests; they either lack exotic specialised knowledge or it becomes inseparable from their high general level, and thus becomes harder to recognise. If broad and narrow ability dimension can develop more or less independently from each other, then, perhaps the male price for their higher degree of visibility and unique competence is an uneven cognitive profile and weak spots.

According to Hirdman (1993) the gender system also creates different rationales for action, but such ideas belong at another analytical level. Others have argued that such differences may be reflections of positive and active choices (Ve, 1982, 1991; Wernersson, 1992, Eccles, 1987).

Ve (1982, 1991) has argued that males and females partly develop different rationalities, and that their basis for doing so is interest and values, which in turn is based on the division of labour. Females more than males are socialised into taking the wellbeing of others into account, a habit which they bring with them into the learning environment. Males more than females are socialised to consider their individual interest, i.e. the principle that most paid production relies on.

Eccles has argued that an expectancy/value model for achievement choice is required to understand gender differences. The pattern she refers to is males' "single-minded" devotion to future occupational roles. She offers some evidence that both females and males value the cost for various choices and act upon that. When asked what concerned them most about their graduate training, females discussed the personal and family costs while males were more likely to be concerned with their academic reputation and status among their professors and peers. Ve reports similar results which she argues is a reflection of females more responsible rationality and males more technical-limited rationality (Ve, 1982).

Haavind (1985) acknowledges that today's ideal is that females shall enter the public sphere and males shall participate equally in the private sphere, at the same time as the caring work shall be upgraded on the value scale. She further acknowledges that there no longer exist formal hindrances for females to enter the public sphere, and most women are now engaged in paid work. This has made females more independent of males, which among other things is reflected in the increasing number of divorces (initiated predominantly by females) and single-parent (mothers) families. Women's caring responsibility for children has, however, decreased only marginally.

The message to young girls is thus that they need to develop broad competencies in order to meet all the demands from both the ideal organisation of gender- relations, and the demands from the public and private sphere. Females may thus neither afford, nor value "single-minded" devotion to any narrow area.

Among feminist researchers today there is wide-spread idea that males and females develop somewhat different rationalities (Gilligan, 1982; Ve, 1982, 1991; Staberg, 1992; Martin, 1991). Rationality, as Martin (1991) acknowledges, is a highly developed capacity for abstract analytic thought, critical thinking, self-government and independence, a trait that historically was attached to the male part of human beings. The gender difference is that females seem to have included care, concern and connection as a part of their rational behaviour.

The definition of rational thinking is very similar to the definition of *g* and *Gc*, dimensions, where females in my study showed a higher average level. My study does, however, not support any notion that the structure of cognitive abilities differs for males and females, only that there are mean differences in some of the dimensions. Ve (1982) has emphasised that males and females develop the same type of rationalities, but which type they will base their action on will depend on the activities they involve in.

The patterns of differences in my studies may, on the other hand be viewed as a sign of different focus. The overall pattern indicates that females to a larger degree than males have developed broad ability dimensions, which they can use on almost any cognitive task. Males on the other hand seem to a larger degree to specialise on narrow ability dimensions, which is profitable in specific tasks. In this way the male group becomes less homogeneous than the female group, since the tendency to specialise may be the uniting trait, but the choice of area for specialisation may differ.

The developmental tradition would probably explain the female advantage on general intellectual dimensions at the age levels investigated with females' earlier maturation. However, with the line of reasoning provided above another explanation may be equally reasonable. The difference may have a materiel ground in the social environment, which implies that the difference may be due to both differences in power and different experiences, values, interests and expectations. Such an explanation also indicates that this pattern may change as the relations between males and females change in society. The inconsistent pattern of gender differences in the comparative study illustrates that differences have a socio-cultural basis. As Haavind (1985, 1992) repeatedly reminds us, we must seek the changing meaning in gender as the social condition change. *"What is happening to male dominance and female subordination? The conclusion is: both are still there, but are difficult to grasp."* (Haavind, 1985, p. 27).

My "birds-eye" perspective is based on statistical analyses, which, according to many feminists is incompatible with feminist research. However, in the next section, I intend to examine to what degree such critique may apply to my research.

Feminist critique of quantitative methodology

The critique against quantitative methods has been massive during the last couple of decades. Many equate quantitative methods with positivism, while others argue that positivism is no longer what it used to be (Egidius, 1986). My experience is that quantitative methods and the way they are thought of among researchers today, are quite different from those in the early days of the social sciences. Much of the feminist critique has its roots in the anti-positivistic movement that grew strong at the end of 1960s (Franssén, 1991; Hallberg, 1992).

Nevertheless, critical analysis and self-reflections are an important part of any type of research. It goes on constantly and is as lively within different paradigms as between them. Here, I will comment on the most frequent feminist critique against quantitative methods, taking a starting point in a paper by Rosén and Wernersson (1996). In this article we argue for the necessity of quantitative feminist research in education, and use my first article as an example. We conclude that the feminist critique to some degree is unrejectable, but as far as the critique concerns statistical analysis techniques and quantitative data, it is in essential parts misdirected.

It could be said, that most of the feminist critique has its roots in a perceived failure of traditional research within the social sciences to address and investigate phenomena of importance to women. As Jayaratne (1993) notices:

> *"Anyone who glances through the indices of social science journals for the past thirty or forty years cannot deny that the great majority of research addresses issues of importance to white male academicians*" (Jayaratne, 1993, p. 111)

The controversy between qualitative/quantitative research methods is a part of this more general critique. Others make the distinction between naturalistic and interpretative modes of inquiry (e. g., Moss, 1996), but the critical points are quite similar. Quantitative against qualitative research methods appear in many debates, where they tend to be standpoints between opposite poles, like good and bad, winner and looser or right and wrong.

An illuminating example of this is Reinhartz (1983) table, where quantitative modes of inquiry ("Conventional or patriarchal") are mirrored against qualitative ("Alternative or feminist") ones. Examples of descriptive phrases for the former are "limited, specialised, exclusive, sometimes socially significant, testing hypotheses, detached relation to subjects", while the latter are "broad, inclusive, socially significant problems, development of understanding through grounded concepts and descriptions, the relation to the subject is involved, sense of commitment, participation, sharing of fate". The description of quantitative methods illustrate the prejudice, lack of insights or perhaps bad experience, which so easily leads to incorrect statements that are reproduced in textbooks and debates. Qualitative methods on the other hand, are, in Reinhartz table, glorified, and the author also comes to the conclusion that quantitative methods should be abandoned within the social sciences.

This polarisation, although not always so extreme, is quite common in feminist discourse (Hallberg, 1992). The binary opposition and the accompanying value hierarchy has many similarities with the female/femininity – male/masculinity debate, that seems to have always been going on, but which for the moment is extremely popular in the media, in everyday conversations, and among feminist researchers with an essentialist standpoint (Martin, 1994).

Addelson (1994) proposes that the "we" against "them" is created in action, where some knowledge is made public and others is kept silent, and by privileging some peoples' actions over others. "*The 'we' is a commitment to a way of knowing the past and making the future*" (Addelson, 1994, p. 6). The critique against quantitative method is often argued from those committed to "a way of making knowledge" different from quantitative research. This qualitative/quantitative "we vs them -making" within feminist research illustrates a paradox within contemporary feminist thoughts. As Kelly, Regan and Burton (1995) notice with reference to Harding (1986), while recent feminist theory has questioned binary oppositions and the accompanying value hierarchies within Western patriarchal thought, many discussions of feminist methodology have also reproduced them, albeit reversing the value hierarchy.

Within feminist research the choice of methods is not only a question of scientific conduct; many feminist researcher classify methods as either in support of or in opposition to women's interests. Quantitative methods are simply connected to hegemonic masculinity (Reinharz, 1979: 1983; Fox Keller, 1985; Harding, 1986; Lather, 1991; Hallberg, 1992; Benhabib, 1994). Although others have pleaded against methodological essentialisms (Martin, 1994; Jayaratne, 1995; Jayaratne & Stewart, 1995; Kelly, Regan & Burton, 1995; Elisasson, 1987, 1994; Rosén & Wernersson, 1996; Wernersson & Ve, 1997) the limited number of feminist researchers in the field bear witness of the persuasive power of arguments against quantitative methods. This acknowledgement has not yet removed the connection between quantitative methods and patriarchy or the hegemonic masculinity. Use of quantitative methods when having feminist interests therefore easily becomes a moral dilemma.

Certain parts of the feminist critique of quantitative methods and traditional social research is today widely accepted:

- Most social researchers are aware that hers/his social background (sex, class, ethnicity, upbringing and so forth) as well as her/his ideological standpoints may (or has to) influence their research praxis one way or another.
- Few researchers believe that any method, quantitative or qualitative, is truly objective or that any method can give us the absolute and pure truth.
- Male dominance in research has a long history, and although the proportion of female researchers has increased over the years, males still dominate both in terms of numbers and in terms of voices being heard. Of all the ascendant and influential perspectives used in research men have developed the major part. Still, the implications of this fact are seldom an issue within traditional research.

However, the fact that research is male dominated and has omitted female perspectives, does not imply that females would always have framed different questions or would have arrived at different conclusions. It is an exaggeration to claim that research questions raised and investigated by males are of no interest to females. However, women certainly have an obligation to be involved in all research areas today

In the early days, when finally a few women were allowed into academia, they soon recognised flaws in the research about gender differences, methodological ones as well as interpretative ones (e. g., Hollingworth, 1914). Woolley (1910, cited in Shields, 1975), as one early witness, commented on the quality of research on sex differences in psychology, and thus alerted the importance of female participation when females are investigated:

> *"There is perhaps no field aspiring to be scientific where flagrant personal bias, and logic martyred in the cause of supporting a prejudice, unfounded assertions, and even sentimental root and drivel, have run riot to such extent as here*". (Woolley, p.340)

As history has shown, and feminists have emphasised (e. g., Benhabib, 1994), in times of methodological monopoly certain important questions will never be asked. However, it does not follow that quantitative methods are "male corrupted" or that qualitative methods would be less limiting in the choice of research questions than quantitative methods, which is more or less implicit in some of the critique against quantitative methods.

Davies and Esseveld (1989) summarise in their textbook on qualitative research on women the feminist critique against quantitative research methods, which they claim led to a complete rejection of such methods among feminist scholars from the mid and late 1970s. The critique was based on the general critique of the social sciences that was put forward in the 1960s. It stated that the conception of reality within the social sciences was characterised by five fundamental elements; distinction, order, quantification, manipulation, and control. Furthermore, it was argued that the methods themselves were based on specific hierarchies. Feminists of the early 1970s stated that this view is a distinctively male view of knowledge, an active but impersonal standpoint, which is consistent with the male ethic within the public sphere, and that quantitative research methods reflected that view. Since concepts and questions often reflect the researchers' personal social world, i.e. white, male, elite, prejudice assumptions and specific (male) points of departure are likely to be reproduced. The more concrete problems of quantitative methods are identified in the following way:

1. The questions are predefined.
2. Concepts and the analytical basis are not founded within the sphere of experience.
3. Quantitative methods are male.
4. The use of quantitative methods (unduly) objectifies people.

5. Since the researcher is governed by the method, there is only a limited set of questions that are asked.
6. Quantitative methods are unable to uncover occurrences that are not immediately observable, and fail to capture the complexity of reality.

Davies and Esseveld (1989) have, as many other feminist researchers (e. g., Martin, 1994) acknowledged, that methodological essentialism is problematic. The feminist research literature is now more concerned with the problems of qualitative research. It has for example been realised that certain questions can not be asked within the framework of qualitative methods alone, and there is a tendency to investigate the lives of females and males separately, which also tends to reinforce sex stereotype images. Furthermore, many of the problems identified above are equally relevant for qualitative research. The problems thus seem to be a part of research praxis in general. The feminist critique has now left the quantitative/qualitative discourse and moved on to discuss which research problems are meaningful to women and those that are not, and the importance of epistemological viewpoints (Lather, 1991; Hallberg, 1992). Despite the official acceptance of quantitative methods, the critique does not disappear just because qualitative methods have similar problems; on the contrary the critique against quantitative methods is still strongly influential. A deeper analysis of the critique may thus be useful.

In discussing the issues, it is useful to differentiate between data collection and data analysis. Fixed answer formats, which allow simple quantification, are for some types of research interests, like mine, necessary and/or preferable. Studies of variations within and between groups require information from many individuals. Such information is also needed if one is interested in studying structures and interrelationships between variables. The quality and validity of à priori formulated question with fixed response formats is, of course, dependent on the insights of the researcher, as is any data collection method. Several qualitative phases precede any serious large-scale survey study, in terms of interviews, various kinds of observations, consultations with literature and expertise and pre-tests. Furthermore, the quality and validity of on à priori formulated questions are in many studies investigated in the quantitative analysis of the data.

Quantitative analysis does not require data to be numerical from the start, but it has to be numerically coded before any analysis can be done. However, if one wants to extend the analysis beyond description of frequencies, then quantitative analysis is not intuitively obvious, which for many researchers within the social sciences is an obstacle. Beyond experience to interpret and understand results, which is needed in all research approaches, statistical techniques require additional competence of a different kind than that which is usually needed for understanding the substantive aspects of educational research. Anyone who masters the technique of advanced statistical analysis realises that these methods can not be used mechanically. They require theoretical insights, intuition and interpretative ability in order to yield interesting results, which is true for qualitative analysis as well.

It is true that the answering formats in the tests used in my studies are constructed to enable quantification, but it is not true that the research questions are socially irrelevant for people or practitioners in the field. Furthermore, for the ability studies, related questions and concepts are based on more than 100 years of experience. During the course of time they have been elaborated, adjusted and reframed as new methodology has been developed. Most of the tests have been developed in close collaboration with students, teachers and other practitioners in experimental and quasi-experimental settings and with other methods than purely quantitative. There has virtually always been an interest in issues related to learning and knowledge. For "the woman" question the use of "objective tests" came as a relief, given the general belief of women's inferior intellectual capacity. (For a recent parallell, see Wennerås & Wold's (1997) analysis of sexism in academic competence ratings). However, it may well be that the tests used are too narrowly defined. This may both be due to the lack of female/feminist researchers in the field, and to the fact that many theoretical notions of knowledge are very hard to translate into operational measures.

A consequence of the feminist critique and the social history that surrounds quantitative methods is a widespread and healthy suspiciousness against numbers, but this suspiciousness has gone too far when numbers are interpreted as evil and/or male, as the third critical statement indicates. Numbers in themselves are not evil to women or female perspectives and numbers are certainly not male. Feminist researchers need to conquer the tools that traditionally have belonged to men and utilise them for our own purposes.

The fourth point of criticism attaches moral values with quantitative methods, as if choice of method would reveal the researcher's valuation, understanding and experience of humans. Quantitative methods are said to improperly objectify people, since the researcher seldom gets to know the people in large-scale studies personally. It certainly is true that personal involvement with all participants in such studies usually is impossible. However, the anonymity of participants in large-scale studies does not indicate that the researcher is uninterested in these individuals' interests and welfare. Clearly choice of method cannot be regarded as a valid indicator of the researcher's ideological perception of people or knowledge. One may even claim, that large-scale studies show the participants greater respect in terms of integrity and time than methods which require the researcher to get close to the research subject (Bertilsson, 1995; Bernstein, 1977).

The critique from feminist qualitative researchers is widely heard, valued and discussed among feminists and others. However, as was concluded by

Rosén and Wernersson (1996) the arguments often take the form of "guilt by association" which illustrates how easy it is to over-generalise observable contemporary occurrences (quantitative methods number males hegemonic masculinity), and interpret them as if they were inherently tied together, which easily leads to incorrect conclusions.

It is my hope that this essay has illustrated the necessity to continue the research on gender differences in patterns of knowledge, for both pedagogical and feminist interests. Furthermore, by connecting different research traditions and perspectives, new patterns of knowledge emerge.

References

Addelson, K. Payne (1994). *Moral Passages. Toward a collectivist moral theory*. New York: Routledge.

Anastasi, A. (1958). *Differential psychology*. New York: MacMillan.

Andersson, A. (in press). Differential selectivity to upper secondary school with respect to gender and previous achievement. *Nordisk Pedagogik*.

Barr, R., Kamil, M. L., Mosenthal, P. B., & Pearson, P. D. (1992). *Handbook of reading research*, (Vol. 2). New York: Longman.

Belenky, M., Clinchy, B., Goldberger, N. & Tarule, J. (1986). *Women's Ways of Knowing. The Development of Self, Voice and Mind*. New York: Basic Books.

Benbow, P. C., & Stanley, J. (1980). Sex differences in mathematical ability: Fact or artifact? *Science, 210,* 1262-1264.

Benbow, P. C., & Stanley, J. (1987). Sex differences in mathematical reasoning ability: More facts. In M. Roth Walsh (Ed.), *The psychology of women: ongoing debates* (pp. 336-340). New Haven: Yale University Press.

Benhabib, S. (1994). *Autonomi och gemenskap. Kommunikativ etik, feminism och postmodernism* [Autonomi and fellowship. Communicative ethics, feminism and postmodernism] Göteborg, Sweden: Daidalos.

Bernstein, B. (1977): *Class, Codes and Control. Vol 3. Towards a Theory of Educational Transmission*. London: Routledge & Kegan Paul

Bertilsson, M. (1995): Till Kritiken av den Kvalitativa Metoden.[To the critique of the qualitative method] *Sociologisk forskning, 4*, s. 18-31.

Binet.A., & Simon, T.(1905): Méthodes nouvelles pour le diagnostique du niveau intellectuel des anormaux (New methods for diagnosing the intellectual level of abnormals*). L'Anné Psychologique, 11*, 191-336.

Bleier, R. (1984). *Science and gender: A critique of biology and its theories of women*. New York: Pergamon Press.

Block, N. J., & Dworkin, G. (1976). *The IQ controversy. Critical Readings*. New York: Pantheon Books.

Bollen, K., & Long, J. S. (Eds.) (1993). *Testing structural equation models*. Newbury Park, CA: Sage.

Burnett, S. A. (1986). Sex-related differences in spatial ability: Are they trivial? *American Psychologist, 41,* 1012-1013.

Caplan, P. J., MacPherson, G. M., & Tobin, P. (1985). Do sex-related differences in spatial ability really exist? A multilevel critique with new data. *American Psychologist, 40*, 786-799.

Carlstedt, B. (1997). *Test complexity and measurement of the general factor*. Manuscript submitted for publication.

Carroll, J. B. (1993). *Human cognitive abilities*. Cambridge: Cambridge University Press.

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

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

Cattell, R. B. (1971). *Abilities: Their structure, growth and action.* Boston: Houghton Mifflin.

Cattell, R. B. (1987). *Intelligence: Its structure, growth and action*. Amsterdam: Elsevier Science Publishers.

Chipman, S. F. (1988). Far too sexy a topic. In S. Hyde & M. C. Linn (Eds.), [A review of The psychology of Gender: Advances through meta analysis 1986. Baltimore MD: John Hopkins University Press.] *Educational Researcher, 17*(3), 46-49.

Chipman, S.F. Brush, L.R. & Wilson, D.M. (1985). *Women and mathematics: Balancing the equation*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cleary, A. T. (1992). Gender differences in aptitude and achievement test scores. In J. Pfleiderer (Ed.), *Sex equity in educational opportunity, achievement and testing. Proceedings of the 1991 ETS inivitational conference* (pp. 51- 90). Princeton, N.J.: Educational Testing Service.

Cole, N. S. (1997). The ETS gender study: How females and males perform in educational settings. ETS May 1997.

Connell, R. W. (1987). *Gender & Power*. London: Basil Blackwell, Ltd

Davies, K. & Esseveld, J. (1989). *Kvalitativ kvinnoforskning*. Stockholm: Arbetslivscentrum.

Deaux, K. (1985). Sex and gender. *Annual Review of Psychology, 36,* 49-82.

Dijkstra, B. (1986). *Idols of perversity: Fantasies of feminine evil in fin-de-siècle culture*. New York: Oxford University Press.

Ds (1994). *Vi är alla olika. en åtgärdsrapport om jämställdhet i skolan som en pedagogisk fråga och ett kunskapsområde från arbetsgruppen Kvinnligt och manligt i skolan*. Departementskrivelse 1994:98.

Eagly, A. E. (1997). Comparing women and men: Methods, finding and politics. In M. Roth Walsh (Ed.), *Women, Men & Gender: ongoing debates* (pp. 24- 32). New Haven: Yale University Press.

Eccles, J. S. (1987). Gender roles and achievement patterns: An expectancy value perspective. In J. M. Reinisch, L. A. Rosenblum, & S. A. Sanders (Eds.), *Masculinity/Femininity Basic Perspectives* (pp. 240-280). New York: Oxford University Press.

Eccles, J., & Jacobs, J. (1987). Social forces shape math attitude and performance. In M. Roth Walsh (Ed.), *The psychology of women: Ongoing debates* (pp. 341-354). New Haven: Yale University Press.

Egidius, H. (1986). *Positivism - fenomenologi - hermeneutik.*. [Positivism - fenomenology - hermeneutics] Lund, Sweden: Studentlitteratur.

Einarsdottir, T. (1997). *Läkaryrket i förändring. En studie av den medicinska professionens heterogenisering och könsdifferentiering.* [Medical profession in change. A study of the heterogenizing and ? of the medical profession]. Göteborg, Sweden: Göteborg University, Monograph from the Department of Sociology.

Eliasson, R. (1987). *Forskningsetik och perspektivval.* [The ethics of research and choices of perspective] Lund, Sweden: Studentlitteratur.

Eliasson, R. (1994). Metodvalet – en fråga om kön och moral? [The choice of methods - a matter of gender and moral] I B. Starrin & P-G. Svensson (Eds.), *Kvalitativ metod och vetenskapsteori.* (pp. 139-162). Lund, Sweden: Studentlitteratur.

Elley, W. B. (Ed.). (1994). *The IEA study of reading literacy: Achievement and instruction in thirty-two school systems. International studies in educational achievement.* Great Britain, Exeter: Pergamon.

Emanuelsson, I., & Fischbein, S. (1986). Vive la difference? A study of sex and schooling. *Scandinavian Journal of Educational Research, 30,* 71-84.

Emanuelsson, I., & Svensson, A. (1989). Does the level of intelligence decrease? A comparison between thirteen-year-olds tested in 1961, 1966, and 1980. *Scandinavian Journal of Educational Research, 30* , 25-37.

Emanuelsson, I., & Svensson, A. (1990). Changes in intelligence over a quarter of a century. *Scandinavian Journal of Educational Research, 34* (3), 171-187.

Entwisle, D. R., Alexander, K. L., & Olson, L. S. (1997). *Children, schools & inequality.* Oxford: Westview Press.

Fausto-Sterling, A. (1985). *Myths of gender. Biological theories about women and men.* New York: Basic Books.

Fausto-Sterling, A. (1992). *Myths of gender. Biological theories about women and men.* (2:nd ed). New York: Basic Books.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American* Psychologist, 43(2), 95-103.

Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at an old controversy. Rewiev of Educational Research, 62 (1), 61-84.

Feingold, A. (1994). Gender differences in variability in intellectual abilities: A cross-cultural perspective. Sex Roles, 30,(Nos.1/2), 81-92.

Fennema, E & Peterson, P (1985). Autonomous Learning Behavior: A Possible Explanation of Gender-Related Differences in Mathematics. In Wilkinson Louice Cherry & Cora B. Marrett (Eds.) *Gender Influences in the Classroom.* New York/London: Academic Press.

Florin, C., & Johansson, U. (1993). "Där de härliga lagrarna gro…" Kultur, klass och kön i det svenska läroverket 1850-1914. [Culture, class and gender in the Swedish secondary grammar school 1850-1914]. Kristianstad, Sweden: Tiden.

Fox, L. H.,Brody. L., & D. Tobin, D. (1980): Women and the mathematical mystique. Baltimore: John Hopkins University Press

Franssén, A. (1991). Observera, deltaga och intervjua. Rapport från ett *fältarbete om kvinnor i vården* [Observe, participate and interview. Report from a field work of women in the caring field]. Lunds University, Sweden, Department of Sociology.

Friedman, L. (1989). Mathematics and the gender gap : a meta-analysis of recent studies on sex differences in mathematical tasks. *Review of educational research, 59*, 2, s. 185-213

Gilligan, C. (1982). *In a different voice.* Cambridge, MA: Harward University press.

Gould, S. J. (1981). *The mismeasure of man.* New York: W. W. Norton & Co.

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

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

Gustafsson, J-E. (1992a). The relevance of factor analysis for the study of group differences. *Multivariate Behavioral Research, 27*(2), 239-247

Gustafsson, J-E. (1992b,). *General intelligence and analytical ability.* Paper presented in the symposium "Individual Differences in Intelligence" at the International Congress of Psychology, Brussels July 19-24.

Gustafsson, J.-E. (1994). Hierarchical models of intelligence and educational achievement. In A. Demetriou & A. Efklides (Eds.), *Intelligence, mind and reasoning: Structure and development.* Elsevier.

Gustafsson, J-E. (1997). *Measuring and understanding G: Experimental and Correlational approaches*. Paper presented at the conference "The Future of Learning and Individual Differences research: Processes, Traits and Content", University of Minnesota, October 9-12.

Gustafsson, J-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28*(4), 407-434.

Gustafsson, J-E., & Stahl, P. A. (1997). *STREAMS User's Guide*, Version 1.7. Mölndal, Sweden: MultivariateWare.

Gustafsson, J-E., & Undheim, J. O. (1996). Individual differences in cognitive functions. In D. Berliner & R. Calfee (Eds.), *Handbook of Educational Psychology* (pp. 186-242). New York: Macmillan.

Gustafsson, J-E., Lindström, B., & Björck-Åkesson, E. (1981). *A general model for the organization of cognitive abilities* (Report 1981:06). Göteborg, Sweden: Göteborg University, Department of Education and Educational Research.

Hallberg, M. (1992). *Kunskap och kön. En studie av femininstisk vetenskapsteori* [Knowledge and gender. A study of feministic epistemology] Göteborg, Sweden: Daidalos.

Halpern, D. F. (1986). A different answer to the question, Do sex-related differences in spatial abilities exist?. *American Psychologist, 41*, 1014-1015.

Halpern, D. K., (1992). *Sex differences in cognitive abilities*. 2nd edition. Hillsdale, N.J.: Erlbaum.

Harding, S. (1986). *The science question in feminism*. London: Cornell University Press.

Haavind, H. (1985). Förändringar i förhållandet mellan kvinnor och män [Changes in the relation between women and men]. *Kvinnovetenskaplig Tidskrift, Årg 6,* (3), 17-27.

Haavind, H. (1992). Vi måste söka efter könets förändrade betydelse [We must look for the changed meaning of the gender]. *Kvinnovetenskaplig Tidskrift, Årg13,* 16-33.

Hirdman, Y. (1987). Makt och kön. (Historiska institutionen, Stockholms universitet) [Power and gender. (Institution of history, Stockholm university)]. In O. Peterson (Ed.), *Maktbegreppet* (pp. 188-206). Stockholm: Carlssons förlag.

Hirdman, Y. (1988). Genussystemet – reflektioner kring kvinnors sociala underordning [The gender system - reflections on the social sub-ordination of women]. *Kvinnovetenskaplig tidskrift, 3,* 49-63.

Hirdman, Y., (1993) Skev läsning: till debatten om genussystem. [Twisted reading: to the debate about gender system] *Kvinnovetenskaplig tidskrift. 14*, 2, s. 57-61

Hollingworth, L. S. (1914). Variability as related to sex differences in achievement. A critique. *The American journal of sociology, 19,* 510-530.

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

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

Hyde, J.S. (1981). How large are cognitive gender differences? A meta-analysis using w^2 and d. *American Psychologist, 36,* 892-901.

Hyde, J. S. (1990). Meta-analysis and the psychology of gender differences. *Signs: Journal of Women in Culture and Society,16*(1), 55-73.

Hyde, J. S., & Linn, M. C. (1986). *The psychology of gender: Advances through meta-analysis*. Baltimore, MD: John Hopkins University Press.

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 194*(1), 53-69.

Hyde, J. S., Fennema, E., & Lamon, J. S. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*, 139- 155.

Härnqvist, K. (1997). Gender and grade differences in latent ability dimensions. *Scandinavian Journal of Psychology, 38*(1), 55-62.

Jacklin, C. N. (1989). Female and Male: Issues of Gender. *American Psychologist, 44*(2), 127-133.

Jayaratne, T. E. (1993). The value of quantitative methodology for feminist research. In Hammersly (Ed.), *Social Research: philosophy, politics and practice* (pp. 109-123). London: Sage University Press.

Jayaratne, T. E., & Stewart, A. J. (1995). Quantitative and qualitative methods in social sciences: Feminist issues and practical strategies. In J. Holland & M. Blair with S. Sheldon (Eds.), *Debates and issues in feminist research and pedagogy* (pp. 217-234). Bridgend, Great Britain: Multilingual Matters, The Open University, WBC.

Jöreskog, K. G., & Sörbom, D. (1993a). *LISREL 8: Structural equation modeling with the SIMPLIS Command language*. Hillsdale, N.J.: Erlbaum.

Jöreskog, K. G., & Sörbom, D. (1993b). *LISREL 8: User's Reference Guide* (2nd ed.). Chicago: Scientific Software.

Keller, E. Fox (1985). *Reflections on Gender and Science.* New Haven: Yale University Press.

Kelly, A (1986). *Gender differences in the teacher-pupil interaction. A meta-analytic review*. University of Manchester, U.K.: Department of Socilogogy.

Kelly, L., Regan, L., & Burton, S. (1995). Defending the indefendable? Quantitative research methods and feminist research. In J. Holland & M. Blair with S. Sheldon (Eds.), *Debates and issues in feminist research and pedagogy* (pp. 235-247). Bridgend, Great Britain: Multilingual Matters, The Open University, WBC.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (a little more than) working-memory capacity. *Intelligence, 14*(4), 389-433.

Lather, P. (1991). *Getting Smart. Feminist research and pedagogy with/in the postmodern.* New York & London: Routledge.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterisation of sex differences in spatial ability: A meta-analysis. *Child Development, 56,* 1479-1498.

Ljung, B-O. (1965). The adolescent spurt in mental growth. Acta Universitatis Stockholmiensis. Stockholm Studies in Educational Psychology 8. Stockholm: Almqvist & Wiksell

Lott, B. (1997). Cataloging gender difference: Science or politics. In M. R. Walsh (Ed.), *Women, Men & Gender. Ongoing debates*.(pp 19-23). London: Yale University Press.

Lpo (1994). *Lpo 94 (Läroplan för det obligatoriska skolväsendet)* [Curriculum for the compulsory school]. Stockholm:Skolverket.

Lundberg, I. (1991). Cognitive aspects of reading. *International Journal Of Applied Linguistics, 1,* 151-163.

Lundberg, I., & Høien, T. (Eds.). (1991). *Literacy in a world of change.* The Hauge: IEA.

Maccoby, E, D. (1966). *The development of sex differences.* Stanford University Press, Stanford California.Maccoby. E, D,. & Jacklin, C. (1974). *The psychology of sex differences.* Stanford, California: Stanford University Press.

Martin. J, R. (1991). The contradiction and the challenge of the educated woman. In I. Elgqvist-Saltzman (Ed.), Education and the Construction of Gender (pp. 6-25). Umeå university, Sweden: *Kvinnovetenskapligt Forums Rapport* (1991:2).

Martin. J, R. (1994). Methodological essentialism, false differences and other dangerous traps. *Signs,* 630-657.

Matarazzo, J. D. (1972). *Wecshler's measurement and appraisal of adult intelligence.* New York: Oxford University Press.

McNemar, Q. (1942). *The revision of the Stanford-Binet Scale.* Boston: Houghton-Mifflin.

Mirza, H. S. (1998). Race, gender and IQ: the social consequence of a pseudo-scientific discourse. *Race, Ethnicity and Education, 1*(1), 109-126.

Moss, P. A. (1996). Variability: A pernicious hypothesis. *Review of Educational Researcher, 25*(1), 20-28, 43.

Muthén. B.(1997). Latent variable modeling of longitudinal and multilevel data. In A. Raftery (Ed.) *Sociological methodology,* pp. 453-480. Blackwell Publishers: Boston

Muthén. B., Kaplan. D, & Hollis. M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52* (3), 431-462.

Noddings, N. (1992). Variability: A pernicious hypothesis. *Educational Research, 62*(1), 85-88.

Olofsson, S-S. (1996). Kvinnor i svensk skola. En studie av kvinnliga lärare och skolledare [Women in Swedish schools. A study of women teachers and school -leaders]. *Pedagogisk orientering och debatt, nr. 105.* Malmö, Sweden: Pedagogiska Punkten.

Peterson, P. L., & Fennema, E. (1985). Effective teaching, student engagement on classroom activities, and sex-related differences in learning mathematics. *American Educational Research Journal, 22*(3), 309-335.

Reinharz, S. (1979). *On becoming a social scientist.* San Francisco: Jossey-Bass.

Reinharz, S. (1983). Experimental analysis: A contribution to feminist research. In G. Bowles and R. D. Klein (Eds.), *Theories of Women's Studies*. London: Routledge and Kegan Paul.

Reinharz, S. (1992). *Feminist methods in social research*. Oxford: University Press.

Resnick, L. B. (1976). *The nature of intelligence*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Reuterberg, S-E. (1998a). On Differential Selection in the Swedish Aptitude Test. *Scandinavian Journal of Educational Research, 42*(1), 81-97.

Reuterberg, S-E. (1998b). *Effects of gender mean differences and variability on the Swedish Scholastic Aptitude test*. Unpublished manuscript, Göteborg, Sweden: Göteborg University, Department of Education and Educational Research.

Reuterberg, S-E., & Svensson, A. (1998). *Vem väljer vad i gymnasieskolan. Förändringar i rekryteringsmönstret efter den senaste gymnasiereformen* [Educational choices in upper secondary school. Changes in the pattern of recruiting after the latest reform] (Rapport 1998:06). Göteborg, Sweden: Göteborg University, Department of Education and Educational Research.

Rosén, M. (1995). Gender differences in structure, means and variances of hierarchically ordered ability dimensions. *Learning and Instruction, 5*, 37-62..

Rosén, M. (1998). Gender differences in hierarchically ordered ability dimensions. The impact of missing data. *Structural Equation Modeling, 5*(1), 37- 62.

Rosén, M. (in press). Gender differences in reading performance on documents across countries. *Reading & Writing*.

Rosén, M., & Wernersson, I. (1996). Kunskapsmönster och kön [Patterns of knowledge and gender]. Om nödvändigheten kvantitativ feministisk forskning i pedagogik. *Pedagogisk forskning i Sverige, Årg 1*(1), 8-24.

Rossiter, M. W. (1982). *Women scientists in America: Struggles and strategies to 1940*. Baltimore, MD: John Hopkins University Press.

Sanders, B., Soares, M. P., & D'Aquila, J. M. (1986). The sex difference on one test of spatial visualization: A non-trivial difference. *Child Development, 53*, 1106-1110.

Sandqvist, K. (1995). Verbal Boys and Mathematical Girls – family background and educational careers. *Scandinavian Journal of Educational Research, 39*, 1, pp. 5 – 36.

SCB (1988). SCB U 20. *Statistiska meddelanden 98 01* [Messages from Statistics]. Stockholm: Statistiska centralbyrån.

SCB (1995). *Utbildningsstatistisk årsbok*. [Yearbook of educational statistics/ Statistics Sweden]. Stockholm: Statistiska centralbyrån.

SCB (1997a). *Education in Sweden 1997*. Statistics Sweden.

SCB (1997b). *Arbetskraftsbarometetern 1997* [Measurement of man power]. Stockholm: Statistiska centralbyrån.
Schafer, A. (1981). Sex and mathematics. *Science,* No. 211, 392.
Scott, J. W. (1988). *Gender and the politics of history.* Oxford: Columbia University Press.
Shields, S. A. (1975). Functionalism, Darwinism, and the psychology of women. *American Psychologist, 30*(7), 739-754.
Siegwald, H. (1944). *Experimentella undersökningar rörande intellektuella könsdifferencer* [Experimental investigations concerning intellectual sex differences]. Unpublished doctoral dissertation, University of Lund, Sweden.
Snow, R. E., & Lohman, D. F. (1989). Implications of cognitive psychology for educational measurement. In R. Linn (Ed.), *Educational measurement* (3rd., pp. 263-331). New York: MacMillan.
SOU (1995). *Viljan att veta och vilja att förstå. Kön , makt och den kvinnovetenskapliga utmaningen i högre utbildning.* Statens offentliga utredningar 1995:110. Stockholm: Utbildningsdepartementet
SOU (1998). Ty makten är din ... Myten om det rationella arbetslivet och det jämställda Sverige. Betänkande från Kvinnomaktutredningen [Thine is the power ... The Myth about the rational working life and the equal Sweden]. *Utredningen om fördelningen av ekonomisk makt och ekonomiska resurser mellan kvinnor och män.* Statens offentliga utredningar 1998:6. Stockholm: Utbildningsdepartementet.
Spearman, C. (1904). 'General intelligence' objectively determined and measured. *American Journal of Psychology, 15,* 201-293.
Spearman, C. (1923). *The nature of 'intelligence' and the principles of cognition.* London: MacMillan.
Spearman, C. (1927). *The abilities of man.* London: MacMillan.
Spender, D. (1982). *Invisible women. The schooling scandal.* London: Writers and readers.
Staberg, E-M. (1992). *OLIKA världar skilda VÄRDERINGAR. Hur flickor och pojkar möter högstadiets fysik, kemi och teknik.* [Different Worlds, Different Values. How girls and boys meet physics, chemistry and technology at the upper level of compulsory school]. Umeå, Sweden: University of Umeå. Dissertation of the Faculty of Social Sciences.
Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review, 84,* 353-378.
Tannen, D. (1990). *You just don't understand.* New York: William Morrow.
Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs,* No. 1.
Tavris, C. (1992). *The mismeasure of women.* New York: Simon & Schuster

Ullman, A. (1997). *Rektorn. En studie av en titel och dess bärare* [The school principal. A study of a title and its carrier]. Stockholm: HLS förlag.

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

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

Ve, H. (1982). Makt, intresse och socialisation [Power, interest and socialisation]. *Kvinnovetenskaplig tidskrift*, (1982:2), 23-329.

Ve, H. (1991). Women's experience – Women's rationality. In I. Elgqvist-Saltzman (Ed.), Education and the Construction of Gender (pp. 48-57). Umeå university, Sweden: *Kvinnovetenskapligt Forums Rapport* (1991:2).

Wagemaker, H. (Ed.). (1996). *Are girls better readers? Gender differences in reading literacy in 32 countries.* The Hague: IEA.

Walsh, M. Roth (1987). *The psychology of women: ongoing debates.* New Haven: Yale University Press.

Walsh, M. Roth (1997). *Women, men & gender: Ongoing debates.* New Haven: Yale University Press.

Wennerås, C., & Wold, A. (1997). Nepotism and sexism in peer-review. *Nature, 387,* 341-342.

Wernersson, I. (1977) *Könsdifferentieringen i grundskolan.* [Gender differentitation in compulsory school] (Göteborg Studies in Educational Sciences, 22). Göteborg, Sweden: Acta Universitatis Gothoburgensis.

Wernersson, I. (1989). Olika kön samma skola? En kunskapsöversikt om hur elevernas könstillhörighet påverkar deras skolsituation [Different gender same school? A research overwiev about how the students gender affect their school situation]. Skolöverstyrelsen, *Vad säger forskningen F 89:1.*

Wernersson, I. (1992a). *Gender differences in social interaction in the classroom setting: Alternative explanations.* Paper presented at "Alice in Wonderland" – First International Conference on Girls and Girlhood: Transitions and Dilemmas, Amsterdam, The Netherlands, 16-19 June, 1992.

Wernersson, I. (1992b). Könsdifferentieringen i grundskolan. I R. Åberg (Red.), *Social bakgrund, utbildning och livschanser*[Social background, education and chances in life] (pp. 89-116). Stockholm: Carlssons.

Wernersson, I., & Ve, H. (1997). Research on gender and education in the nordic countries. *Scandinavian Journal of Educational Research, 41*(3-4), 295- 317.

Wikander, U. (1992). Delat arbete, delad makt: om kvinnor underordning i och genom arbetet. En historisk essä. In G. Åhström & Y. Hirdman (Red.), *Kontrakt i kris: om kvinnors plats i välfärdsstaten* [Agreement in crisis:

about womens place in the welfare state] (pp. 21-84). Stockholm: Carlsson.

Wilder, G. Z., & Powell, K. (1989). Sex Differnces in test performance. A survey of the literature. *College Board Report, No. 89-3.*

Wilkinson, L. C., & Marrett, C. B. (1985). *Gender influence in the classroom interaction.* New York: Academic press.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment.* Mahwah, N.J.: Lawrence Erlbaum Associates.

Willis, S. (1990). The power of mathematics: For whom? In S. Kenway & S. Willis (Eds.), *Hearts and minds. Self-esteem and the schooling of girls.* London: Falmer Press.

Woolley Thompson, H. (1910). Psychological literature: A review of recent literature on the psychology of sex. *Psychological Bulletin, 7,* 335-342.

Öhrn, E. (1990). *Könsmönster i klassrumsinteraktion. En observations- och intervjustudie av högstadieelevers* lärarkontakter[Gender patterns in classroom interactions. Observations and interviews concerning students' interactions with teachers in grade nine of the comprehensive school] (Göteborg Studies in Educational Sciences, 77). Göteborg, Sweden: Acta Universitatis Gothoburgensis.

Part II

Pergamon

Learning and Instruction, Vol. 5, pp. 37–62, 1995

0959–4752/95 $29.00

0959–4752(95)00002–X

GENDER DIFFERENCES IN STRUCTURE, MEANS AND VARIANCES OF HIERARCHICALLY ORDERED ABILITY DIMENSIONS

MONICA ROSÉN

Göteborg University, Sweden

Abstract

The study investigates gender differences in cognitive abilities using a multivariate, latent variables approach. A test battery of 13 ability tests and 3 standardized achievement tests was given to a sample of 1224 13-year-olds, and a hierarchical model with 13 latent ability variables was fitted to the covariance matrix. No gender differences were found in the structure of cognitive abilities. There were mean differences in favor of females in general intelligence and in the broad crystallized intelligence factor. Males had higher means on most spatial dimensions, for which they also were more variable. Males also had higher means on narrow numerical and verbal ability dimensions.

Introduction

Many studies have been published on gender differences in cognitive abilities. However, such differences are typically investigated with a single variable comparison. In the present study, this method is contrasted with a multivariate latent variable approach. This methodology takes into account the fact that any cognitive task reflects more than one cognitive dimension.

Previous Research on Gender Differences in Cognitive Abilities

Research on gender differences in intellectual functioning has been going on for more than hundred years (e.g., Siegwald, 1944; Fausto-Sterling, 1992). Work from the 19th century often both assumed and "proved" female inferiority in cognitive functioning. Much of this work would not be accepted as scientific today, but it was fully accepted at that time. More recent research shows a different picture even though traces of historical assumptions still may be recognized in some studies.

Address for correspondence: Dr M. Rosén, Department of Education and Educational Research, Göteborg University, P.O. Box 1010, S-431 26 Mölndal, Sweden.

Several studies during the present century show that females excel in the verbal domain, but there are also studies which demonstrate an absence of gender differences in verbal ability. Anastasi (1958, first published 1937) and Tyler (1965) concluded that females are superior in verbal and linguistic functions from infancy to adulthood. Tyler found that girls are superior in verbal fluency but not in vocabulary. Several researchers (e.g., Maccoby, 1966; Maccoby & Jacklin, 1976; Halpern, 1986; Wernersson, 1989) have reached the same conclusion and it has also been shown that females have an advantage in written production, which seems to hold true for vocabulary tasks too. From the 1989 National Evaluation study in Sweden, Balke (1990) analyzed achievement in English as a foreign language and found that girls achieved better on reading comprehension and vocabulary tests which required productive answers. On an essay test girls outscored boys. Halpern (1992) concluded on the basis of a review of differences in verbal ability, that females outperform males in fluent speech production, anagrams, and on general and mixed tests of verbal abilities, whereas males outperform females on solving analogies. In a large-scale study of several cohorts of 13-year-olds in Sweden there were in 1961 only small differences between the genders in the verbal domain but in 1985 girls were clearly ahead (Emanuelsson & Svensson, 1990). Hyde and Linn (1988) performed a meta-analysis of 165 studies of gender differences in verbal ability and concluded that there are no gender differences in verbal ability. However, the studies seem to support the general conclusion that there are differences favoring girls within the verbal domain, whenever the tasks require written production.

Most reviews of findings in the spatial domain have concluded that boys on the average have greater visual-spatial ability (Anastasi, 1958; Tyler, 1965; Maccoby & Jacklin, 1976; Hyde, 1981; Halpern, 1986). However, in a meta-analysis of 172 studies of gender differences in spatial ability Linn and Petersen (1985) found different amounts of gender differences for three types of spatial ability. For spatial perception tasks a moderate advantage for males was found; for mental rotation tasks there was a clear superiority for males, and for spatial visualization no gender difference was found. Emanuelsson and Svensson (1990) found that the pattern of differences in spatial reasoning ability has changed over time. Boys were superior in the early 1960s, but 25 years later the difference had disappeared. It thus seems that there is a tendency for males to excel in spatial tasks even though the difference varies over time and tasks.

Males seem on the average to perform better than females in mathematical, quantitative achievement tests as well (e.g., Maccoby & Jacklin, 1976; Halpern, 1990). However, some studies show that the male superiority is limited to certain aspects of mathematical achievement: mathematical reasoning (e.g., Tyler, 1965) and problem solving (e.g., Hyde, Fennema, & Lamon, 1990). Hyde (1981) reanalyzed the studies collected by Maccoby and Jacklin with meta-analysis methodology and concluded that the differences were not as large as the previous studies have indicated. Friedman (1989) performed a meta-analysis of more recent studies on gender differences in mathematical tasks. The analysis included 98 studies published between 1974 and mid-1989. Friedman found that the average gender difference is very small, and that differences in performance are decreasing over time. Halpern (1990) suggests that a part of the male advantage in quantitative ability is due to the male advantage in spatial-visual ability. Evidence supporting this notion has been presented by Fennema and Sherman (1977), Hills (1957), Burnett, Lane, and Dratt (1979) and Hyde, Geiringer, and Yen (1975). However, Linn and Hyde (1989) argue that the evidence is only correlational and not sufficient for such conclusions. Evidence against the hypothesis has been presented by for example Fennema and Tarte (1985).

Feingold (1992) assessed variance ratios (VRs) of gender differences in several large and well established test batteries and claimed that males are more variable than females in general knowledge, mechanical reasoning, quantitative ability, spatial visualization, and spelling. There was essentially homogeneity of variance for most verbal tests, short term memory, abstract reasoning, and perceptual speed.

As pointed out by Feingold (1992) and Noddings (1992), the question about gender differences in variability in performance has a long history which includes an ongoing controversy about the "greater male variability". The controversy has several parts; one concerns existence of a difference in variability between the genders (McNemar & Terman, 1936), and one concerns the interpretation and use of the greater male variability findings (see Feingold, 1992 for an overview). Noddings (1992) calls it a "pernicious hypothesis" due to the social history that surrounds it. The often found greater male variability has usually been given biological explanations (e.g., Dijkstra, 1986; Rossister, 1982). Statements of women's lesser variability such as "females are all alike" were given by prominent male professionals (Dijkstra, 1986). Noddings (1992) affirms that evolution theory exacerbated this trend; it was widely held that "the male has been shaped for thought and creativity, the female for reproduction. In her increasingly specialized role, woman has little need for the variation that characterizes man" (p. 85). Psychologists and educators in the U.S.A. used the variability hypothesis to explain why men were more often than women found in the rank of genius. Females were (are?) considered genetically inferior, an opinion shared by leading social scientists such as G. Stanley Hall, Edward Thorndike and James McKeen Cattell (Rossister, 1982). Noddings argues that the variability hypothesis should not be referred to as an explanation at all, it is rather the variability itself that should be explained.

From this brief review we can conclude that gender differences are found with some regularity in at least three domains: in verbal ability, in visual-spatial ability and in mathematical or quantitative ability. Meta-analyses have been conducted in all three domains and Hyde (1990) concludes that gender differences in cognitive abilities are generally not large. Contradictory results are common, however, and it seems that different results are obtained when different aspects within these domains are focused upon. The need for a theory for the abilities and their internal structure is obvious. Different results are also obtained for different samples. A fair amount of the studies referred to have used selected and/or very small samples which may have caused bias in either direction (Hyde, 1981; Becker & Hedges, 1984).

The main point of the present paper, however, is that it is not only unclear which ability is measured by a particular test, but that tests always tend to measure more than one ability. If different abilities are involved in test performance, the observed difference in level of performance will reflect a mixture of differences in several abilities. It is thus necessary first to decompose the complex observed variables into latent variables, and then investigate gender differences in these ability dimensions. This perspective is closely linked to a hierarchical view of the organization of cognitive abilities, so before we go any further some recent research on the structure of abilities will be described.

Hierachical models of ability

During the last couple of decades hierarchical models of ability, which include both broad and narrow factors of ability have gained increasing popularity (see, e.g., Carroll, 1993; Gustafsson, 1988, 1989). This is at least partly due to disenchantment with results achieved in the research on primary abilities in the tradition of Thurstone and Guilford (Lohman, 1989). Hierarchical models have primarily been developed as higher-order factor analytic models. In these, correlations between first-order factors are analyzed, with second-order factors as an outcome. The relations between the second-order factors can in turn be analyzed in terms of third-order factors, and so forth. Cattell (1971) and Horn (1986) has formulated the influential *Gf* ("fluid intelligence") — *Gc* ("crystallized intelligence") model on the basis of results of second-order analyses of "primary" ability factors (Thurstone, 1938). The model includes several broad abilities of which the following are the most important:

Fluid intelligence (Gf) is reflected in tasks requiring abstraction, concept formation, perception and eduction of relations, and in which tasks the stimulus material is new to the examinees. *Gf* is supposed to represent biological factors and incidental learning on intellectual development. *Gf* is most strongly involved in primaries such as Induction (*I*), General Reasoning (*R*) and Cognition of Figural Relation (*CFR*).

Crystallized intelligence (Gc) is shown in tasks requiring abstraction, concept formation, perception and eduction of relations too, but where the stimulus material primarily is verbal-conceptual in nature. In contrast to *Gf*, *Gc* is established through cultural pressures, education and experience. *Gc* is most strongly shown in primary factors such as Verbal Comprehension (*V*), Cognition of Semantic Relations (*CMR*), and in measures of school achievement.

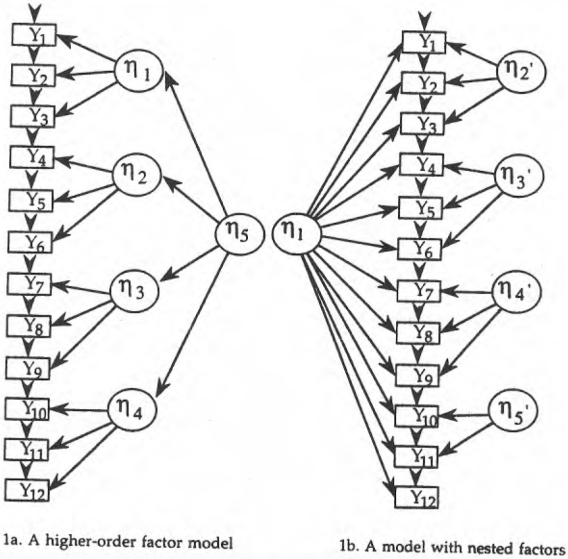
General Visualization (Gv) is involved in "visualizing the movements and transformations of spatial patterns, maintaining orientation with respect to objects in space, unifying disparate elements and locating a given configuration in a visual field" (Horn & Cattell, 1966, p. 254). *Gv* is loaded by primary abilities such as Visualization (*Vz*), Spatial orientation (*Sr*), Flexibility of Closure (*Cf*), Speed of Closure (*Cs*), and Figural Adaptive Flexibility (*DFT*).

In addition to the Cattell–Horn model there exist other higher-order models. Carroll (1993) has reanalyzed almost 500 correlation matrices using higher-order factor analysis and has formulated a "Three-Stratum" model, which at the first level (stratum) defines narrow primary abilities. At the second stratum about a dozen broad abilities of the kind identified by Cattell and Horn are identified, and at the third stratum there is a single, general, factor.

Gustafsson (1984, 1988; see also Undheim & Gustafsson, 1987) showed that the results from exploratory factor analysis can be replicated with the confirmatory factor analytic methodology which has been developed by Jöreskog (e.g., Jöreskog & Sörbom, 1989). At the lowest level, the model represents as first-order factors abilities similar to the primary mental abilities identified by Thurstone (1938), Guilford (1967), and other researchers working within the multiple factor tradition. At an intermediate level, the model identifies the factors *Gf*, *Gv* and *Gc* that are easily interpreted as the broad abilities within the Cattell–Horn model. At the highest level, the model includes a factor of general intelligence (*g*), on which all the broad abilities has loadings. This more powerful and flexible technique also makes it possible to carry the analysis to models with a general third-order factor, even when the number of variables is relatively limited. In several studies the general factor has been found to be perfectly correlated with the *Gf* factor (Undheim & Gustafsson, 1987; Gustafsson, 1988; Undheim, 1981), which makes it possible to identify a *g*-factor in as invariant a manner as the *Gf*-factor may be identified.

As was pointed out by Gustafsson (1989; see also Gustafsson & Balke, 1993) the hierarchical model implies that the instruments which are used to measure cognitive achievements are influenced by all the dimensions which appear in the hierarchical structure above the variable. Each instrument is therefore an indicator of several dimensions. This conclusion is not immediately deducible from a higher-order factor model, because the higher-order factors in such models have only an indirect relation to the observed variables (see Figure 1a). However, as has been shown by Gustafsson and Balke (1993) such models may be reformulated in terms of an orthogonal model with nested factors (NF-model). In this kind of model it is immediately seen that every observed variable is complex in the sense that it is related to more than one latent variable (see Figure 1b).

When gender differences in cognitive performance are analyzed, this is typically done by looking at differences between means for one test at the time. However, this univariate method relies on the assumption that performance on a particular task is indicative of only one underlying ability. If the hierarchical approach is adopted for understanding the structure of abilities the differences in performance on a task are due to several ability dimensions with different degrees of generality. According to



1a. A higher-order factor model 1b. A model with nested factors

Figure 1. Alternative hierarchical models.

this model an observed group difference may be due to differences in one or more of these underlying abilities. This suggests that group differences should be studied within a multivariate latent variable model. In order to compare the two approaches gender differences will be studied both traditionally and within a hierarchical latent variable model.

How Should Group Differences in Cognitive Abilities be Investigated?

Feingold (1992) argued that contemporary research on gender differences in cognitive performance has based its conclusions upon mean differences, or standardized mean differences *d*s (the difference between male and female means divided by the pooled within-gender standard deviation; Cohen, 1977), whereas no concern has been taken to differences in variability. He argued that gender differences in variability and in central tendency must be considered together to form correct conclusions about the magnitude of cognitive gender differences. If the genders differ in variability in performance on cognitive test, the more variable gender would be overrepresented at both high and low levels of performance when the average scores of males and females are the same. If the genders differ in both average performance and variability, the gender differences at lower and upper levels of ability would be different from what would be expected from the standardized mean difference.

The critique in this paper against univariate analysis is based on the fact that the observed performance on any cognitive tasks reflects a mixture of several ability dimension of different degree of generality. It is thus necessary to first decompose the variance into latent variables (preferably in a hierarchical model with orthogonal nested factors), and then compare means on the latent variables. The same critique applies to Feingold's approach, the variances he compares are not reflections of single ability dimensions, but reflect a mixture of several abilities.

Theory and previous research on cognitive abilities argue for a hierarchical model with three levels (cf. Gustafsson, 1988; Carroll, 1993), and methodology and software for estimating such models is available (e.g., LISREL by Jöreskog & Sörbom, 1989; LISCOMP by Muthén, 1988; EQS by Bentler, 1989).

However, a multivariate analysis with latent variables typically makes a homogeneity assumption through restricting the factor structure and the factor variances to be invariant over the groups. In the univariate case the homogeneity assumption can be investigated by comparing the variances for the different groups and by comparing the shape of the variability curve. If a multivariate approach is adopted, using a hierarchical latent variable model in a multiple group analysis, many more questions concerning homogeneity in cognitive abilities may be answered.

The primary purpose of the present paper is to contrast the traditional approach for investigating gender differences with a multivariate hierarchical approach. Another purpose is to investigate the assumption of invariance between the genders in a hierarchical latent variable model of cognitive abilities in grade 6. Differences in patterns of loading, means and variances on the latent variables will thus be investigated. A third purpose, finally, is to investigate the robustness of estimates of mean differences in latent variables in the presence of structural differences in the measurement model.

Method

Sample and Variables

The sample consisted of most 6th graders (age 12–13) in two communities, Mölndal and Kungsbacka (51 classes) close to Göteborg in the western part of Sweden. The first data collection took place in 1980. The test battery of 13 ability test along with some questionnaires was administered to 1254 pupils. For different reasons the battery of cognitive tests could not be administered to 30 pupils, which left a final sample of 1224 pupils, 602 boys and 622 girls. The sample and subsamples were followed up in different respects during the following three years. This has resulted in a longitudinal data base, suited for studies of individual differences in the domain of cognition. Several studies have been conducted and published on the data; some of them will be referred to later on in this article.

A description of variables collected in grade 6 is presented below. The procedure for the data collection and the aptitude variables are described in great detail elsewhere (Gustafsson, Lindström, & Björck-Åkesson, 1981; Gustafsson, 1984), so only a short presentation of the instruments and the latent variables which they originally were hypothesized to measure is given here.

The Test Battery in Grade 6

The test battery was assembled in such way that enough primary factors would be represented to make possible identification of the second-order factors *Gf*, *Gc* and *Gv*. The test battery consists of two parts; one with 13 rather well established ability tests, and one with 3 standardized achievement tests.

The following ability tests were included in the battery in grade 6. Bold style indicates the short labels used for the tests and italic style the short labels for the factors in tables and figures:

1. Opposites **Op** — The task is to select the word which is the antonym of a given word (40 items). This test has been found to measure Verbal Comprehension (*V*), which loads on *Gc*.

2. Number Series II **NS** — 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 (20 items). Tests of this type have been shown to load on the primary factor Induction (*I*), which in turn loads on *Gf*.

3. Letter Grouping II **LG** — 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 (20 items). This is a test of Induction (*I*), which loads on *Gf*.

4. Auditory Number Span **ANS**	This is a conventional digit-span test, with digits in series of varying length being read for immediate reproduction (19 items). The test measures Memory span (*Ms*), which has a weak relation to *Gf*.
5. Auditory Letter Span **ALS**	This test is identical with the Auditory Number Span test, except that letters are used instead of digits (19 items). This test too loads on Memory span (*Ms*).
6. The Raven Progressive Matrices **Ra**	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 (45 items). The test measures (*I*) (cf. French, Ekstrom, & Price, 1963) or Cognition of Figural Relations (*CFR*) (cf. Horn & Cattell, 1966) which factors load on *Gf*.
7. Metal Folding **MF**	The task is to find the three-dimensional object which corresponds to a two-dimensional drawing (30 items). Tests of this kind load on Visualization (*Vz*), which loads on the second-order factor *Gv*.
8. Group Embedded Figures **GEFT**	The task is to find a simple figure within a more complex figure (9 items out of 16 given as recommended by Witkin, Oltman, Raskin, & Karp, 1971). This test has been found to measure Flexibility of Closure (*Cf*) which loads on *Gv*.
9. Hidden Patterns **HP**	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 simpler configuration (200 items in each of two parts). This is another test of *Cf* and *Gv*.
10. Copying **Co**	Each item consists of a given geometrical figure, which is to be copied onto a square matrix of dots (32 items in each of two parts). This test too measures *Cf* and *Gv*.
11. Card Rotations **CR**	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 (14 items in each of two parts). The test defines the Spatial orientation (*Sr*) primary, which loads on *Gv*.
12. Disguised Words **DW**	In this test words are presented with parts of each letter missing, and the task is to identify the word (12 items in each of two parts). This test measures the primary factor Speed of Closure (*Cs*), which has been found to load on *Gv*.

13. Disguised Pictures **DP**	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 (12 items in each of two parts). This is another test of *Cs*.

These 13 tests of cognitive ability were administered on one occasion. The second data collection in grade 6 yielded scores on three Standardized Achievement tests (SA), that is, mathematics, English and Swedish. Results were available at the subtest level for most of the pupils in the sample.

The SA test in mathematics is composed of 5 subtests and they are hypothesized to load on the *Gc* factor and on the primary factor Numerical Achievement (*Num Ach*):

1. Per Cent Calculation **Ma PCC**	Tests the ability to carry out calculations involving the per cent concept. There are 16 items in the test, which are to be solved in 25 min.
2. Estimates **Ma Est**	Tests the ability to make rapid estimates of approximate results of an expression. The test is a multiple-choice test with 21 items and a time-limit of 10 min.
3. Geometry and Diagrams **Ma GD**	Consists of 8 geometry items presenting tasks such as computing the area of rectangles, and 6 items assessing the ability to understand information presented within graphs and tables.
4. Applied Computation **Ma AC**	Presents 12 verbally stated problems, most of which require a mixture of the rules of arithmetic.
5. Numerical Calculation **Ma NC**	Presents 20 items testing understanding of the number line, the ability to carry out additions, subtractions, multiplications, divisions and calculation with fractions. The time-limit is 35 min.

In the SA test in English (En) there are 4 subtests and they are also hypothesized to load on the *Gc* factor:

1. Vocabulary **En Vo**	Consists of 40 multiple-choice items which present a one- or two-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 min.
2. Listening Comprehension **En LC**	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 items the alternatives are verbal, and for 20 questions the alternatives are pictorial. The time-limit is 30 min.
3. Forms and Structures **En FS**	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 min.
4. Reading Comprehension **En RC**	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. The 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 approximate answer from a list with 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 is 30 min.

The SA test in Swedish (Sw) language consists of 6 subtests and they are all hypothesized to load on the *Gc* factor:

1. Spelling **Sw Sp**	Contains 25 items in which the task is to correctly spell dictated words.
2. Reading Comprehension **Sw Rc**	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. There are 21 items in the test, which is administered with a time-limit of 35 min.
3. Words of Relation **Sw WR**	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 28 min.
4. Vocabulary **Sw Vo**	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 is 12 min.
5. Word List **Sw WL**	Tests the pupils ability to use a word-list to find the meaning, spelling and flexion of words. There are 11 items in the test and the time-limit is 10 min.
6. Sentence Construction **Sw SC**	Presents a text lacking punctuation, and the task is to add the 18 punctuation marks which are missing. The time-limit is 15 min.

Subjects and Attrition

The results on the standardized achievement tests were collected a few months after the first test battery. The standardized achievement tests were administered by the regular teacher who also decided whether the class should or should not participate. This caused attrition of some classes and students.

Out of 1224 pupils 981 had results on all the standardized achievement tests, while 113 (9.3%) had not taken any of these. The frequency of attrition is about the same among boys and girls, about 20% in each group.

Previous Research on the Data Collected

As mentioned before the test battery was assembled in such way that enough primary factors would be represented to make possible identification of the second-order factors *Gf*, *Gc* and *Gv*.

Gustafsson and Balke (1993) analyzed the same data in order to find out to what degree ability factors could predict school achievement. The factors were fitted in an order determined by their degree of generality in a nested factor model.

First the *g*-factor was fitted, and then more narrow factors were successively added. Their final model of the test battery included 20 observed variables (13 ability test, some split in halves and 3 SA tests) and 10 latent factors (*g*, *Gv'*, *Gc'*, *V'*, *Ms'*, *Num Ach'*, *CFR'*, *Vz'*, *Sr'* and *Cs'*; the prime is added to indicate a residual factor), and fitted rather well. The most striking difference between the HO-model and the NF-model was that the

Table 1
Aptitude Variables Included in the Analysis in Grade 6
(Gustafsson & Balke, 1993)

Test name	Label	Factors		
1. Opposites — odd items	**Op-O**	*g*	*Gc'*	*V'*
2. Opposites — even items	**Op-E**	*g*	*Gc'*	*V'*
3. Auditory Number Span	**ANS**	*g*		*Ms'*
4. Auditory Letter Span	**ALS**	*g*		*Ms'*
5. Number Series II	**NS**	*g*		*Num Ach'*
6. Letter Grouping II	**LG**	*g*		
7. Raven — odd items	**Ra-O**	*g*	*Gv'*	*CFR'*
8. Raven — even items	**Ra-E**	*g*	*Gv'*	*CFR'*
9. Metal Folding — odd items	**MF-O**	*g*	*Gv'*	*Vz'*
10. Metal Folding — even items	**MF-E**	*g*	*Gv'*	*Vz'*
11. Group Embedded Figures	**GEFT**	*g*	*Gv'*	
12. Hidden Patterns	**HP**	*g*	*Gv'*	
13. Copying	**Co**	*g*	*Gv'*	
14. Card Rotations Part I	**CR-I**	*g*	*Gv'*	*Sr'*
15. Card Rotations Part II	**CR-II**	*g*	*Gv'*	*Sr'*
16. Disguised Words	**DW**	*g*	*Gv'*	*Cs'*
17. Disguised Pictures	**DP**	*g*	*Gv'*	*Cs'*
18. SA test Swedish	**Sw Ach**	*g*	*Gc'*	
19. SA test English	**Eng Ach**	*g*	*Gc'*	
20. SA test Math	**Ma Ach**	*g*	*Gc'*	*Num Ach'*

latter contained fewer factors (10 vs 14). The reason for that is that the narrow factors in the NF-model may be identified only if there is any remaining residual variance when the broader factors have been identified. In Table 1 summary information is presented about the Gustafsson and Balke model.

As has already been pointed out there are at least two ways to formulate a hierarchical model, HO-modelling and NF-modelling. One problem with the HO-approach is that it does not offer procedures for analysis of differences between groups in means on higher-order factors (Gustafsson, 1992). With NF-models it is easy to analyze group differences in means on the latent factors. These reasons make a nested factor model preferable for this study.

Strategy for the Analysis

Some of the cognitive tests (Opposites, Ravens, Card Rotation and Metal Folding) have in the following multivariate analysis been split into two subscores (one for odd and one for even items), following the same procedure as Gustafsson (1989) and Gustafsson and Balke (1993). This makes it possible to extract more narrow dimensions, which thus tend to mix test-specific variance with variance due to narrow abilities.

In the first step, gender differences are analyzed in the traditional way through looking at the observed test-scores. The second step in the analysis will be to fit a hierarchical model with latent variables to the data and after doing so, study gender differences in means on the latent variables with a simple and straight-forward method, and to contrast the results of the latent variable approach with the univariate approach. In the next step previous assumptions about invariance in structure and variances are tested by performing the hierarchical model in a multiple group analysis. The final step in the analysis will be to evaluate the two latent variable approaches in order to find out to what extent the simple approach is valid and reliable.

The tools used for conducting the analysis is SPSS (SPSS Inc., 1988) for the univariate part, and LISREL technique for the confirmatory factor analysis (Jöreskog & Sörbom, 1988).

The LISREL program computes several goodness of fit indices as an aid in evaluating model fit. Conventional fit measures in LISREL have always been the χ^2-test and the Goodness of Fit Index (GFI). However, LISREL 8 (Jöreskog & Sörbom, 1993, pp. 120–131) reports several others, among them Normed Fit Index (NFI), Non Normed Fit Index (NNFI) and Root Mean Square of Approximation (RMSEA). For accepting a model the RMSEA should be below .05. The new measures are considered to be better indicators of fit than the conventional ones. Thus the χ^2 measure indicates in large samples almost always a significant deviation between the model and data. However, the conventional measures are also reported for the sake of comparability.

Results

Gender Differences in Observed Test Performance

The univariate analysis has been performed by correlating test scores with gender for comparison of means. Positive numbers indicate female advantage, negative male

Table 2
Relations Between Gender and Test Performance in Grade 6. $N = 981$

	Mean comparison			Variance comparison		
	r	r^2*100	t	VR	F	$p<0.05$
Opposites (1 & 2)	.057	0.32	1.79	0.99	1.01	
Auditory Number Span	.023	0.05	0.67	1.40	1.40	**
Auditory Letter Span	.034	0.10	1.06	1.08	1.08	
Number Series (1 & 2)	−.071	0.50	−2.22**	1.33	1.33	**
Letter Grouping (1 & 2)	.233	4.97	7.49**	0.98	1.02	
Raven Progressive Matrices	.106	1.12	3.33**	1.36	1.36	**
Metal Folding	.022	0.05	0.69	1.28	1.28	
Group Embedded Figures (1 & 2)	.035	0.12	1.08	1.07	1.07	
Hidden Pattern	.160	2.56	5.07**	0.94	1.06	
Copying	.061	0.37	1.91	0.90	1.11	
Card Rotation (1 & 2)	−.061	0.37	−1.90	0.94	1.07	
Disguised Words	−.019	0.04	−0.61	1.03	1.03	
Disguised Pictures	−.052	0.27	−1.61	1.06	1.06	
SA test in Swedish total score	.160	2.56	5.08**	1.14	1.14	
— Spelling	.209	4.37	6.68**	1.31	1.31	**
— Reading Comprehension	.075	0.56	2.35**	1.23	1.23	
— Words of Relation	.182	3.31	5.48**	1.01	1.01	
— Vocabulary	−.045	0.20	−1.42	1.03	1.03	
— Word List	.162	2.62	5.14**	1.03	1.03	
— Sentence construction	.226	5.11	7.27**	1.29	1.29	
SA test in English total score	.137	1.88	4.33**	1.33	1.33	**
— Vocabulary	.125	1.56	3.93**	1.34	1.34	**
— Listening Comprehension	.040	0.16	1.25	1.30	1.30	**
— Forms and Structures	.159	2.53	5.03**	1.14	1.14	
— Reading Comprehension	.157	2.46	4.96**	1.34	1.34	**
SA test in Math total score	−.076	0.58	−2.39**	1.12	1.12	
— Per Cent Calculation	−.059	0.35	−1.85	0.94	1.07	
— Estimates	−.157	2.46	−4.98**	1.24	1.24	
— Geometry & Diagrams	−.028	0.08	−0.87	1.14	1.14	
— Applied Computation	−.049	0.24	−1.52	1.11	1.11	
— Numerical Calculation	−.016	0.03	−0.51	1.04	1.04	

For significance at the .05 level t-values should > 1.96 or < −1.96.

advantage. Variance ratios (VR) have been obtained by dividing the male variance with the female variance, VRs greater than 1 indicate higher male variability, and VRs less than 1 higher female variability. An F-test investigates if the variance is significantly different between the genders. With 479/500 degrees of freedom the F_{crit} is approximately 1.31 on the 5% level.

There are some significant mean differences, most of which favor girls. Boys achieve better on Number Series and on one of the subtests on the SA Math test. When all the subscores are added up, the total score on the SA Math test is favoring boys. Girls on the other hand achieve better results on most subtests on the SA tests in Swedish and English, which is also reflected in the total score on these two tests. Girls also achieve better on Hidden Patterns, Raven Progressive Matrices and on Letter Grouping. In most of the tests gender explains less than 1% of the variance. Exceptions are most of the subtests in Swedish and English and the Estimates subtest in Math where gender accounts for 1.5 to 5.1% of the variance. This is also the case with Letter Grouping and

Hidden Patterns. These results correspond reasonably well to the findings in previous research.

A few variances differ significantly between the genders. Boys vary more than girls in performance on Number Series, on Ravens Progressive Matrices, on Auditory Number Span, on one of the subtest in SA test in Swedish and on three out of four subtests in the SA test in English.

The Hierarchical Measurement Model

The next step in the analysis will be to fit the variables into a hierarchical model with latent variables. The Gustafsson and Balke (1993) model has been taken as a starting point. Their model fitted the data quite well, but there is still room for some improvement. By identifying a general speed factor (*Gs'*), identified by tests with narrow time limits (Copying, Hidden Patterns, Card Rotations and Group Embedded Figures) model fit was improved. Another problem with the Gustafsson and Balke model is

Table 3
Loadings of the Aptitude Variables on the Ability Factors in Grade 6. N = 981

	g	*Gv'*	*Gc'*	*Ms'*	*Cfr'*	*Vz'*	*Sr'*	*Cs'*	*Gs'*	*Num Ach'*	*Eng Ach'*	*V'*	*Words'*
Op-O	.56		.31									.32	.42
Op-E	.54		.33									.33	.44
NS	.75									.24			
LG	.70												
ANS	.28			.57									
ALS	.34		.12	.70									
Ra-O	.54	.20			.73								
Ra-E	.57	.18			.64								
MF-O	.48	.64				.40							
MF-E	.50	.62				.38							
GEFT	.57	.30							.23				
HP	.56	.24							.44				
Co	.53	.25							.48				
CR-I	.43	.19					.58		.32				
CR-II	.50	.25					.69		.25				
DW	.33							.69	.11				
DP	.22	.23						.43					
Ma PC	.55	.10	.25							.50			
Ma Est	.53		.10							.44			
Ma GD	.65	.07	.19						.11	.44			
MA AC	.70		.14							.40			
Ma NC	.63		.23							.50			
Eng Vo	.58		.51								.45	.09	
Eng LC	.52		.34								.51	.24	
Eng FS	.60		.50								.35		
Eng RC	.59		.49								.39	.17	
Sw Sp	.50		.54										
Sw RC	.60		.35									.36	
Sw WR	.61		.41									.19	
Sw Vo	.53		.39									.47	.12
Sw WL	.65		.32										
Sw SC	.58		.38										

that the *Num Ach'* factor was poorly identified (see Table 1). In order to improve that factor the SA Math test was split into its 5 subtest. The other SA tests were also divided into their subtests in order to allow a more fine-grained analysis of verbal performance. The model is estimated from the covariance matrix and an acceptable model fit was obtained (χ^2 = 726, df = 406, RMSEA = .03, GFI = .96, NFI = .96, NNFI = .98). The standardized factor loadings are presented in Table 3.

By squaring the standardized loading the proportion of variance in the tests accounted for by the latent dimension is obtained. The *g*-factor accounts for a sizeable proportion of variance in most tests. The highest loadings are obtained with the non-verbal reasoning test Number Series and Letter Grouping, and with the Applied Computation subtest in the SA Math test which involves problem solving as well.

Gv' appears in the model as a broad factor which has a relatively weak relation to every test which involves figural content, except for the relationship to the Metal Folding test where *Gv'* explains about 39% of the variance in the variable. Two of the Ma subtests also include figural content which results in tiny but significant loadings on *Gv'*.

Gc' has its strongest relations with the SA subtests in Swedish and English as expected, and with Opposites which demands both correct interpretations of words and a rich vocabulary. *Gc'* also has relations with all the subtests in SA Math and gives a small contribution to the Auditory Letter Span test.

MS', *CFR'*, *Vz'*, *Sr'* and *CS'* only have their relations to the tests that are hypothesized to measure these ability dimensions. The amount of variance explained by these factors in the tests varies between 17 and 45%.

The *Num Ach'* factor is defined by all the subtests in SA Math and the *Eng Ach'* factor is defined by all subtests in SA English.

V' is a narrow verbal factor with relations to all tests that requires some vocabulary knowledge, and *Word'* is an even more narrow factor with relations to Swedish vocabulary tests.

Gender Differences in Latent Ability Dimensions

Gender differences in level of performance on latent variables may be studied in different ways. The most obvious procedure is to perform a two-group analysis and investigate mean differences in latent variables with the procedure described by Jöreskog and Sörbom (1989). Another way is to enter gender as a dummy variable in a one-group analysis. A dummy variable is a dichotomous variable with values of zero and one; in this case zero represent males and one represent females. The gender variable naturally lacks measurement error. By letting this dummy variable have relations to the latent variables in the model, one obtains estimated coefficients which are equivalent to the mean difference on the latent variable. This is exactly the same procedure as when gender is entered in an ordinary regression analysis.

The dummy procedure is much less tedious than the two-group approach, but it does make the assumption that the pattern of relations between observed and latent variables is invariant for the genders. Results obtained with both approaches will be presented.

Table 4 presents relations between gender, coded as a dummy variable, and the true scores of the latent variables in grade 6. Significant relations between gender and the latent factors have been entered in the model. Positive correlations and t-values indicate

Table 4
Relations Between Gender and Latent Variables in a One-group Model for Grade 6. $N = 981$

	Grade 6		
	r	r^2*100	t
g	.22	5.00	5.41
Gv'	–.15	2.00	–2.70
Gc'	.26	7.00	4.58
V'	–.46	21.00	–6.37
Ms'			
Num Ach'	–.55	30.00	–11.28
CFR'			
Vz'			
Sr'	–.24	6.00	–5.93
Cs'	–.14	2.00	–3.04
Gs'			
Eng Ach'	–.19	4.00	–3.33
Word'			

results to girls' advantage, and negative values to boys' advantage. By squaring the correlation and multiplying it with 100, the per cent variance explained by gender is obtained.

There are several rather striking results in Table 4. There is quite a large difference in the *g*-factor favoring girls which is also true for the broad factor *Gc'*. There are also differences favoring boys in the broad factor *Gv'* and the narrow factors *V'*, *Sr'*, *Cs'*, *Eng Ach'* and *Num Ach'*. The difference in *Num Ach'* is quite considerable, gender explaining some 30% of the variance in the factor. The same goes for *V'* where gender explains 21%. This multivariate analysis thus gives results which are strikingly different from those obtained in the univariate analysis.

Differences in the Structure of Cognitive Abilities

An equivalent model to the one-group model is obtained for the two groups when all parameters are constrained to be equal over the groups. A two-group model in which all parameters are free over groups will have equal or better model fit, and the modification indices for each gender will tell if the pattern should be changed for any of the two groups. The modification indices for the two groups did not suggest any pattern of relations between observed and latent variables different from the pattern presented in the original model.

The following fit indices are obtained for the model where all free parameters are constrained to be equal between the groups; $\chi^2 = 1425$, df = 934, RMSEA = .02, GFI = .92, NFI = .93, NNFI = .97. This model is to be compared to the one-group model. When all free parameters are allowed to vary between the groups the fit indices are as follows; $\chi^2 = 1189$, df = 812, RMSEA = 0.02, NFI = 0.94, NNFI = 0.98. The improvement is significant ($\chi^2 = 236$, df = 122). However, among all the differences between the genders that can be seen in the model now, only a few are significant.

The easiest way to find out which ones are significantly different is to be guided by the modification indices from the model with all free parameters invariant between the genders and then calculate a t-test on the parameters that are released from the invariance restriction. The t-test may be calculated by the following formula,

$$\frac{est_{males} - est_{females}}{\sqrt{\dfrac{(n_{males} \times SE^2_{est_{males}}) + (n_{females} \times SE^2_{est_{females}})}{n_{males} + n_{females}}}} \qquad (1)$$

where *est* is the unstandardized estimate, SE_{est} the standard error for the estimate, *n* is the number of subjects in each group. When a few parameters were allowed to vary between the genders, the model fit improved $\chi^2 = 1326$, df = 925, RMSEA = 0.02, NFI = 0.93, NNFI = 0.98, which matches the model fit for the model where all parameters were allowed to vary.

Significant differences between the genders were found in 3 out of 13 factor variances, 3 out of 100 factor loadings and 2 out of 32 error variances. The differences are presented in Table 5.

Table 5
Significant Gender Differences in Various Variance Components in the Hierarchical Latent Variable Model

	Males	Females	T-value	VR
Variance in latent variables				
Gv'	5.59	2.88	3.76	1.94
SE_{est}	0.82	0.63		
Ms'	2.76	1.86	3.46	1.48
SE_{est}	0.29	0.22		
CFR'	5.15	3.61	4.28	1.43
SE_{est}	0.41	0.31		
Factor loading				
NS on *g*	1.88	1.50	3.45	
SE_{est}	0.12	0.10		
MA Est on *g*	1.39	1.03	3.60	
SE_{est}	0.11	0.10		
Sw SC on *V'*	–0.41	0.18	–3.47	
SE_{est}	0.19	0.14		
Error variance				
ANS	4.96	3.57	3.97	
SE_{est}	0.40	0.29		
Sw SC	9.09	6.14	5.0	
SE_{est}	0.72	0.43		

The variance in the male group turns out to be greater than among females for the broad *Gv'* factor, and for the primary ability factors *CFR'* and *Ms'*. From the mean comparison above a male advantage is seen in *Gv'*, whereas no such difference is visible in *Ms'* and *CFR'*.

The 3 significant differences in factor loadings are hard to interpret in any substantial way. It seems that the Number Series test and the Math Estimates subtest are to some extent better measures of *g* for males than for females, and that the SA Swedish in Sentence Construction subtest is a better predictor for *V'* for females than males. These scattered results obviously need to be cross validated before any conclusions can be drawn, because they just as well may be due to chance. The same conclusion must be drawn about differences in error variance. Still the question is if these differences affect the mean differences in any substantial way. In the final step in this analysis it is investigated how stable the mean differences are between the different models.

In Table 6 t-values for mean differences in the latent variables are compared for the one-group model, the two-group model with all parameters invariant between the genders, the two-group model where some parameters are allowed to vary between the groups and the model where all parameters are allowed to vary between the groups.

Table 6
Comparing T-values on Mean Differences in the Latent Variables for Different Models Between Males and Females. *N* of Males = 480, *n* of Females = 501

	T-values			
	One-grp model Invariant assump	Two-grp model All invariant	Two-grp model Some free pm's	Two-grp model All pm's free
g	5.41	5.45	6.22	5.88
Gv'	–2.70	–3.09	–3.36	–3.76
Gc'	4.58	4.58	4.11	4.84
V'	–6.37	–6.44	–6.37	–5.37
Ms'				
Num Ach'	–11.28	–11.39	–11.16	–10.13
CFR'				
Vz'				
Sr'	–5.93	–5.99	–5.99	–5.23
Cs'	–3.04	–3.25	–3.31	–2.83
Gs'				
Eng Ach'	–3.33	–3.31	–3.11	–3.0
Word'				

The mean differences between the genders turn out to be fairly stable when different types of approaches and assumptions are applied. The model that is most different from the others is the one where all parameters are allowed to vary between the genders. This is due to the fact that there is a large number of parameters in the model which show small random variations between the genders. The one-group approach does not offer any different result in terms of mean differences on latent variables than does the two-group analysis in which some parameters are allowed to vary. There is, however, a small tendency for this model to yield more conservative estimates of the group difference.

Discussion and Conclusions

As has been demonstrated, the methodology and the pattern of gender differences in cognitive abilities are closely linked so both methodological issues and results will be attended to the discussion. The pattern of results obtained in the univariate and multivariate analyses are strikingly different, so the study is an illustration of Halpern's observation that "In fact, the experimental and statistical methods used may be more important in determining the answers we get to empirical questions than the underlying phenomenon being investigated" (1986, p. 16). This does not imply, of course, that the differences between methods are arbitrary, or that one method may not be preferable over other methods.

Mean Differences in Cognitive Abilities

The pattern of mean differences on the latent variables between the genders is very different from the ones on the observed test scores where girls performed better on most tests. The multivariate analysis showed a substantial and unexpected difference in the general ability dimension favoring girls. A major reason for this is, of course, that in the univariate analysis there is no single test which represents the construct "general ability", so even if there are differences in such an ability, they cannot be detected in the univariate analysis. This also explains why so few findings concerning gender differences in general ability have been reported in previous research. However, in a parallel study to the present one on other data, Härnqvist (1994) found the same pattern, e.g., a substantial advantage on *g* for females.

Another explanation why no gender differences favoring girls in general ability have been reported before is to be found in the development of cognitive tests. In the standardization of modern so called IQ tests it has simply been assumed or decided, that the general intelligence is equally distributed between the genders at each age level. Test items which have shown gender bias have either been excluded or matched by an item with the opposite bias (Matarazzo, 1972; McNemar, 1942). The matching strategy permits questions about gender differences to be answered, whereas the exclusion strategy does not, as long as analysis is restricted to observed test scores. One can not simply assume that an item or a test is biased just because one group performs better than another on it. The findings from the multivariate analysis and the hierarchical model indicate that item selections as well as selections of tests are to be determined on better and more substantial grounds when the purpose of the test is to investigate cognitive abilities. It should also be noted that an IQ score is something quite different from the construct of *g* (Gustafsson, 1984). The *g*-dimension is highly involved in IQ scores, but in addition such tests involve a substantial amount of *Gc'*.

If the finding in the multivariate analysis is accepted as valid, it may be asked how the difference in favor of girls in general ability can be interpreted. One should keep in mind that this study investigates a sample of 12–13-year-olds in Sweden. This age-group may be in a dynamic phase particularly in terms of intellectual development. One possible explanation is that the difference is a reflection of gender differences in the process of development, and that the earlier onset of puberty among girls

(Ljung, 1965) may be accompanied by an earlier spurt in mental growth for girls than for boys. Waber (1976) showed interesting relations between physical maturation and verbal and spatial skills at age 13, and concluded that the differences are due to physical maturation and not to gender. However, no such relation was found in a more recent and rigorous study by Waber, Mann, Merola, and Moyland (1983) considering whether systematic associations between maturation rate and cognitive performance exists prior to adolescence. It seems hard to extract biological factors from social ones when addressing gender differences in cognitive performance. Future research of various age-groups, preferably using longitudinal approaches with data from a wider range of age groups, may give us answers regarding the dynamics of intellectual development and its relation to age.

Girls show a similar advantage on the broad *Gc'* factor, and this is, for example, reflected in performance on the achievement tests. Performance on verbal tests have often been interpreted as a female advantage in verbal ability. However, *g* and *Gc'* explains most of the variance in those test, and when the variance is partitioned into latent variables it becomes clear that the observed gender differences in those tests are due to gender differences in those broad dimensions. Girls advantage on *Gc'* may also depend on earlier maturity and on expectations of gender specific behavior. *Gc'* is rather well defined in this analysis with 17 indicators, that is all 15 subtest of the standardized achievements tests, and Opposites and Letter Grouping from the ability test battery. The univariate analysis shows significant differences on 8 of the language subscores favoring girls, some of them substantial, and on one of the subscores in math, which favor boys.

In the univariate analyses of the tests selected to measure *Gv'*, no single test shows a significant difference in favor of boys in grade 6. Thus, even though the pattern of gender differences varies somewhat over the different tests these analyses seem to support the conclusion that there are no gender differences in spatial ability among 13-year-olds. However, in the multivariate analysis there is a significant difference in favor of boys not only in the broad *Gv'*-factor, but also in the narrow spatial abilities *Sr'* and *Cs'*. The only ability in the spatial domain which does not display a significant difference is *Vz'*. The reason why the methods of analysis give contradictory results is, of course, that the spatial tests are quite highly related to the *g*-factor, and the difference in favor of girls in the *g*-factor cancels the difference in favor of boys on the spatial factors, leaving a non-significant difference in observable test performance.

There are also mean differences in some of the more narrow dimensions, that is in the residual factors after the broad dimensions of *g*, *Gc'* and *Gv'* are taken into account. On the *Num Ach'* factor and on the primary *V'*-factor the differences favoring boys are huge, gender explaining some 30% in the former and some 20% of the variance in the latter. Boys also show some advantage on the *Eng Ach'* factor. The univariate analysis indicated the opposite pattern for most language tests, and only a small difference favoring boys on one of the tests in math. One reason for these contradictions is, that a part of the difference in favor of girls on those tests is captured by the difference in the *g*-factor. It may also be that the difference in favor of boys in *V'* and *Eng Ach'* is due to test-bias, i.e., the tests involved in the factors were developed with an eye towards gender differences, and that the difference in *g* in favor of girls unknowingly was compensated for by choice of words that favor boys. Similar results are reported by Härnqvist (1994).

In further research of gender and the quantitative ability dimension it maybe of interest to try to distinguish between math knowledge and math reasoning in order to shed more light over the gender gap in the numerical domain as well as over the quantitative ability dimension. However, the present study does not support such distinctions which may be due to the limitations of the standardized achievement test in math.

As a second explanation to the male advantage on quantitative tasks some researchers have claimed that visual-spatial abilities are related to quantitative abilities (Halpern, 1992; Anderson, 1990; Fennema & Sherman, 1977; Hills, 1957; Hunt, 1985; Fennema & Tarte, 1985). Some of them have used this notion as an explanation for why males perform better on quantitative tasks, or why males more often than females choose education with quantitative content. A relation between spatial and quantitative ability is so far rejected by the hierarchical model of cognitive abilities, since all abilities are assumed to be orthogonal. However, since every task does reflect more than one ability dimension, there is still a possibility that some topics within the mathematical domain like topology, trigonometry and geometry may have some spatial components. None of the spatial dimensions in the present study explained any part of the variance in the Ma subtest in geometry. It is reasonable to assume that a substantial part of the relationship between quantitative and spatial performance, is a reflection of *g* which is highly involved in both quantitative and spatial tasks.

Another hypothesis why females perform so poorly compared to males on the *Num Ach'* factor is, that the whole quantitative area is a well established male domain, where females are not supposed nor expected to excel. Males on the other hand are both expected and supported to do so (Boswell, 1985; Stallings, 1985; Leder, 1990; Schatz Kohler, 1990; Fennema, 1990).

No mean differences are found in the broad *Gs'* or in the primary ability dimensions *Ms'*, *CFR'*, *Vz'* or in *Word'*.

Variance Differences in Cognitive Abilities

The pattern of gender differences in variances on latent variables is also strikingly different from the pattern observed for the observed variables. The univariate approach showed a somewhat greater male variability for 3 out of 16 different tests of abilities, namely Number Series, Auditory Number Span, Ravens Progressive matrices and on 4 of the 10 language subscores. The multivariate analysis on the other hand showed almost twice as large a variance for boys on the broad spatial ability dimension *Gv'*, and about 45% larger variance for boys in *CFR'* (Cognition of Figural Relations, a primary spatial ability dimension) and in *Ms'* (Memory span), another narrow ability dimension.

The variance difference in spatial dimensions is considerable. Biological and cultural hypotheses have been offered to explain mean differences on spatial dimensions, but none of these seem reasonable for explaining the difference in variance. It is an extremely interesting question why there are variance differences on these particular ability dimensions and not on others. Both biological and social explanations have been suggested, however, none of those seem unproblematic in the light of available evidence. Future research may contribute to the understanding of this phenomenon.

It is also of great interest to investigate how spatial ability dimensions relate to different kinds of tasks in real life. These findings may also be of importance for the question concerning equality between the genders, since spatial tasks are frequently used for selection to different types of educations and positions.

No gender differences in variance were obtained in the general ability dimension, *g*, in the broad abilities *Gc'* and *Gs'*, or in the other narrow and primary ability dimensions represented in this study (*Vz'*, *Sr'*, *Cs'*, *V'*, *Num Ach'*, *Eng Ach'* and *Words'*. The "greater male variability hypothesis" is thus not supported in this study, for most cognitive abilities. The data in the present study suggest a revised and more focused examination of both genetic and environmental effects, but the focus should be on *Gv* and other more narrow spatial dimensions and not on *g*.

Differences in the Structure of Cognitive Abilities

The structure of abilities has never been an issue in the univariate analysis, but the hierarchical view on cognitive abilities suggests that possible structural differences should be investigated when comparing groups. As for the genders the two-group analysis did not show any difference in the structure of cognitive abilities, even though a few parameter estimates differed significantly between the genders. None of those differences proved to affect the overall pattern of mean differences on the cognitive abilities. However, more research is needed before any conclusions can be drawn about gender differences in the structure of cognitive abilities. The conclusion here is that there do not appear to be any *qualitative* between-gender differences in the structure or organization of cognitive abilities. The same conclusion has been reached by Hertzog and Carter (1982), who used an oblique confirmatory model.

The hierarchical approach

When utilizing the multivariate hierarchical approach, it becomes clear that cognitive tests are quite highly related to the *g*-factor and to other broad factors like *Gc'* and *Gv'*, which is one reason why contradictory results are obtained compared to the univariate approach. Non-significant gender differences in observed test performance may hide differences in those broad dimensions. The SA subtests in math, for example, hides a huge difference on the *Num Ach'* factor favoring boys. Significant differences on observed test performance may be due to differences on other latent variables than the test is intended to capture. Girls often perform better on verbal tests, for example, but when the hierarchical approach is adopted, this difference disappears because it is due to higher means for girls on *g* and *Gc'*. Differences in latent dimensions may thus cancel or change directions of gender differences in observed test performance.

The overall pattern of results from the multivariate analysis thus seems reasonable, and also more interesting than the pattern of results obtained from the univariate analysis, but it may be asked if the results are trustworthy. It is easy to see that the results from the multivariate analysis are closely linked to the choice of model, and to the correctness of the definition of the latent variables in the hierarchical model. Since an infinite number of models may be fitted to any set of variables it could be argued that the results reported here are arbitrary and could be replaced by a number of other sets of results. However, even though this argument is true in a general sense, it could also be argued that the hierarchical model relied upon here is well established in a large number of previous studies.

In the multivariate approach adopted here it is crucial that the *g*-factor is identified in an invariant manner, because the results for all the latent variables are affected by the definition of this latent variable. In future research it is essential to determine to what extent this allows identification of an invariant *g*-factor, and it is also essential to investigate to what extent the pattern of gender-differences is affected by choice of indicators for the *Gf*-factor (Gustafsson, 1994).

The choice of indicators for the broad dimensions on the intermediate level in the hierarchical model may also affect the pattern of gender differences. They too should be identified in an invariant manner since they affect all the narrow dimensions on the first level in the hierarchy.

Gender Differences in Means vs Gender Differences in Variability

This paper both agrees and disagrees with Feingold's (1992) claims that comparing means is not enough for making reliable conclusions of gender differences in cognitive abilities. Feingold argues that variability within each group should be considered as well. The result of the present study shows that different methods give different results. This paper has demonstrated that observed test performance is determined by several latent variables, which may differ with respect to both means and variances. It thus implies that the observable performance is a complex function of the characteristics of the latent variables, which should be considered in the analysis too.

Despite the fact that the two-group analysis rejected the invariance assumption between the genders, the differences in means on the latent variables remained fairly stable when compared to the analysis in which invariance was assumed. A two group analysis, where some relations are allowed to vary, will of course still contribute with a much more complex and complete picture of the differences, but it is still appropriate to use the invariant approach when focusing on mean differences between the genders on latent ability dimensions.

Acknowledgements — The work reported here was financed by the Swedish Council for Research in the Humanities and Social Sciences (HSFR). The analysis is a part of the research project Models for Analysis of Group Differences in Cognitive Abilities (MAGIC), under the leadership of professor Jan-Eric Gustafsson. I am grateful for his valuable support, advice and comments on my draft. I also would like to express my gratitude to professor emeritus Kjell Härnqvist's helpful comments on my draft.

References

Anastasi, A. (1958). *Differential psychology*. New York: MacMillan.

Andersson, J. R. (1990). *Cognitive psychology and its implications* (3rd ed.). New York: Freeman.

Balke, G. (1990). Pojkar, flickor och Engelska i årkurs 5. Delrapport inom den nationella utvärderingen av grundskolan. (Boys, girls and English in grade 5. A progress report from the national evaluation of the compulsory school). Report from the Department of Education, 1990: 25, University of Göteborg.

Becker, B. J., & Hedges, L. V. (1984). Meta-analysis of cognitive gender differences: A comment on an analysis by Rosenthal and Rubin. *Journal of Educational Psychology*, **76**(4), 583–587.

Bentler, P. M. (1989). *EQS structural equations program manual*. Los Angeles, CA: BMDP Statistical Software.
Boswell, S. L. (1985). The influence of sex-role stereotyping on women's attitudes and achievement in mathematics. In S. F. Chipman, L. R. Brush, & D. M. Wilson (Eds.), *Women and mathematics: Balancing the equation* (pp. 175–198). Hillsdale, NJ: Lawrence Erlbaum.
Burnett, S. A., Lane, D. M., & Dratt, L. M. (1979). Spatial visualization and sex differences in quantitative ability. *Intelligence*, **6**, 57–68.
Cattell, R. B. (1971). *Abilities: Their structure, growth and action*. Boston: Houghton-Miffin.
Carroll, J. B. (1993). *Human cognitive abilities*. New York: Cambridge University Press.
Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (rev. ed.). New York: Academic.
Dijkstra, B. (1986). *Idols of perversity: Fantasies of feminine evil in fin-de-siècle culture*. New York: Oxford University Press.
Emanuelsson, I., & Svensson, A. (1990). Changes in intelligence over a quarter of a century. *Scandinavian Journal of Educational Research*, **34**(3), 171–187.
Fausto-Sterling, A. (1992). *Myths of gender: biological theories about women and men* (2nd ed.). New York: Basic Books.
Feingold, A. (1992). Sex differences in variability in intellectual abilities: A new look at old controversy. *Review of Educational Research*, **62**(1), 61–84.
Fennema, E. (1990). Teachers' beliefs and gender differences in mathematics. In E. Fennema & G. C. Leder (Eds.), *Mathematics and gender* (pp. 149–168). New York: Teachers College Press, Colombia University.
Fennema, E., & Sherman, J. (1977). Sex-related differences in mathematics achievement, spatial visualization, and sociocultural factors. *Journal of Educational Research*, **14**, 51–71.
Fennema, E., & Tarte, L. A. (1985). The use of spatial visualization in mathematics by girls and boys. *Journal for Research in Mathematics Education*, **16**, 184–206.
Friedman, L. (1989). Mathematics and the gender gap: A meta-analysis of recent studies on sex differences in mathematical tasks. *Review of Educational Research*, **59**(2), 185–213.
French, J. W., Ekstrom, R. B., & Price, L. A. (1963). *The description of aptitude tests in terms of rotated factors*. Princeton, NJ: Educational Testing Service.
Guilford, J. P. (1967). *The nature of human intelligence*. New York: McGraw-Hill.
Gustafsson, J.-E. (1984). A unifying model for the structure of intellectual abilities. *Intelligence*, **8**, 179–203.
Gustafsson, J.-E. (1988). Hierarchical models of individual differences in cognitive abilities. In J. R. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol. 4, pp. 35–71). Hillsdale, NJ: Erlbaum.
Gustafsson, J.-E. (1989). Broad and narrow abilities in research on learning and instruction. In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation and methodology: The Minnesota Symposium in Learning and Individual Differences* (pp. 203–237). Hillsdale, NJ: Erlbaum.
Gustafsson, J.-E. (1992). The relevance of factor analysis for the study of group differences. *Multivariate Behavioural Research*, **27**(2), 239–247.
Gustafsson, J.-E. (1994). Hierarchical models of intelligence and educational achievement. In A. Demetriou & A. Eftklides (Eds.), *Intelligence, mind, and reasoning: Structure and development* (pp. 45–73). Amsterdam: Elsevier Science B.V.
Gustafsson, J.-E., & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioural Research*, **28**(4), 407–434.
Gustafsson, J.-E., & Undheim, J. (1992). Changes in broad and narrow factors of intelligence from ages 12 to 15. *Journal of Educational Psychology*, **84**(2), 141–149.
Gustafsson, J.-E., Lindström, B., & Björck-Åkesson, E. (1981). A general model for the organization of cognitive abilities. University of Göteborg: Report from the department of education and educational research, 1981: 6.
Halpern, D. K. (1986). *Sex differences in cognitive abilities*. Hillsdale, NJ: Erlbaum.
Halpern, D. K. (1992). *Sex differences in cognitive abilities* (2nd ed.). Hillsdale, NJ: Erlbaum.
Härnqvist, K. (1994). Gender and grade differences in latent ability dimensions. Unpublished manuscript, University of Göteborg, Department of Education and Educational Research.
Hertzog, C., & Carter, L. (1982). Sex differences in the structure of intelligence: A confirmatory factor analysis. *Intelligence*, **6**, 287–303.
Hills, J. R. (1957). Factor analyzed abilities and success in college mathematics. *Educational Psychological Measurement*, **17**, 615–622.
Horn, J. L. (1986). Intellectual ability concepts. In J. R. Sternberg (Ed.), *Advances in the psychology of human intelligence* (Vol 3, pp. 35–78). Hillsdale, NJ: Erlbaum.
Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized intelligence. *Journal of Educational Psychology*, **57**, 253–270.
Hunt, E. (1985). Verbal ability. In R. J. Sternberg (Ed.), *Human abilities: An information processing approach* (pp. 33–58). New York: W. H. Freeman.

Hyde, J. S. (1981). How large are cognitive gender differences? A meta-analysis using w2 and d. *American Psychologist*, **36**(8), 892–901.
Hyde, J. S. (1990). Meta-analysis and the psychology of gender differences. *SIGNS. Journal of Women in Culture and Society*, **16**, 55–73.
Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin*, **104**(1), 53–69.
Hyde, J. S., Geiringer, E. R., & Yen, W. M. (1975). On the empirical relation between spatial ability and sex differences in other aspects of cognitive performance. *Multivariate Behavioral Research*, **10**, 289–309.
Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin*, **107**(2), 139–155.
Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7. A guide to the program and applications*. Chicago: SPSS.
Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Hillsdale, NJ: Erlbaum.
Leder, G. C. (1990). Teacher/student interactions in the mathematics classroom: A different perspective. In E. Fennema & G. C. Leder (Eds.), *Mathematics and gender* (pp. 149–168). New York: Teachers College Press, Colombia University.
Linn, M. C., & Hyde, J. S. (1989). Gender, mathematics and science. *Educational Researcher*, **18**(8), 17–19, 22–27.
Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of differences in spatial ability: A Meta-Analysis. *Child Development*, **56**, 1479–1498.
Ljung, B-O. (1965). *The adolescent spurt in mental growth*. Acta Universitatis Stockholmiensis. Stockholm Studies in Educational Psychology 8. Stockholm: Almqvist & Wiksell.
Lohman, D. F. (1989). Human Intelligence: An introduction to advances in theory and research. *Review of Educational Research*, **59**, 333–373.
Maccoby, E. D. (1966). *The development of sex differences*. Stanford, CA: Stanford University Press.
Maccoby, E. D., & Jacklin, C. (1976). *The psychology of sex differences*. Stanford, CA: Stanford University Press.
Matarazzo, J. D. (1972). *Wechler's measurement and appraisal of adult intelligence*. New York: Oxford University Press.
McNemar, Q. (1942). *The revision of the Stanford-Binet Scale*. Boston: Houghton-Miffin.
McNemar, Q., & Terman, L. M. (1936). Sex differences in variational tendency. *Genetic Psychology Monographs*, **18**, 1–65.
Muthén, B. O. (1988). *LISCOMP Analysis of Linear Structural Equations with a Comprehensive Measurement Model. A program for advanced research* (2nd ed.). Mooresville: Scientific Software Inc.
Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrica*, **52**(3), 431–462.
Noddings, N. (1992). Variability: A pernicious hypothesis. *Review of Educational Research*, **62**(1), 85–88.
Rossister, M. W. (1982). *Women scientists in America: Struggles and strategies to 1940*. Baltimore, MD: John Hopkins University Press.
Schatz Koehler, M. (1990). Classrooms, teachers, and gender differences in mathematics. In E. Fennema & G. C. Leder (Eds.), *Mathematics and gender* (pp. 128–148). New York: Teachers College Press, Colombia University.
Siegwald, H. (1944). Experimentella undersökningar rörande intellektuella könsdifferencer (Experimental investigations concerning intellectual sex differences). Unpublished doctoral dissertation at University of Lund, Sweden.
Stallings, J. (1985). School, classroom and home influences on women's decisions to enroll in advanced mathematics courses. In S. F. Chipman, L. R. Brush, & D. M. Wilson (Eds.), *Women and mathematics: Balancing the equation* (pp. 199–224). Hillsdale NJ: Lawrence Erlbaum.
SPSS Inc. (1990). *SPSS Reference Guide*. Chicago. SPSS.
Thurstone, L. L. (1938). Primary mental abilities. *Psychometric Monographs, No. 1*.
Tyler, L. E. (1965). *The psychology of human differences* (3rd ed.). New York: Appelton-Century-Crofts.
Undheim, J. O. (1981). On intelligence IV: Toward a restoration of general intelligence. *Scandinavian Journal of Psychology*, **22**, 251–265.
Undheim, J. O., & Gustafsson, J.-E. (1987). The hierarchical organization of cognitive abilities: restoring general intelligence through the use of linear structural relations (LISREL). *Multivariate Behavioral Research*, **22**, 149–171.
Waber, D. P. (1976). Sex differences in cognition: A function of maturation rate? *Science*, **192**, 572–573.
Waber, D. P., Mann, M. B., Merola, J., & Moylan, P. M. (1976). Physical maturation rate and cognitive performance in early adolescence: A longitudinal study. *Developmental Psychology*, **21**(4), 666–681.

Wernersson, L. (1989). Olika kön samma skola? En kunskapsöversikt om hur elevernas könstillhörighet påverkar deras skolsituation. [Different genders, the same school? A research overview about how students' gender affect their school situation.] *Skolöverstyrelsen, Vad säger forskningen F 89:1*.

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

Part III

STRUCTURAL EQUATION MODELING, 5(1), 37–62

Gender Differences in Hierarchically Ordered Ability Dimensions: The Impact of Missing Data

Monica Rosén
Department of Education and Educational Research
Göteborg University

In this study, the impact of missing data on estimates of gender differences in hierarchically ordered ability dimensions is investigated. The data consist of 13 ability tests on which the whole sample of 1,224 13-year-old students has information and 3 standardized achievement tests on which a reduced sample of 981 participants has information. Utilizing missing data techniques for latent variable models, the study also becomes a validation of previously reported gender differences in latent hierarchical ability dimensions. In the previous analysis, the 243 students lacking data were assumed to be missing at random and thus excluded from the analysis. The attrition was found to have an impact on both the hierarchical model and on gender differences in latent dimensions. The attrition appeared biased with respect to general achievement and gender. When the cases with missing data were included in the analysis, the structure of the model remained stable and strengthened in some respects. The female advantage on *g* increased, whereas their advantage on Crystallized Intelligence decreased. The male advantage on spatial dimensions increased, whereas their advantage on narrow achievement factors decreased. About half of the attrition was due to missing classes, and the other half was due to missing individuals. There was no gender difference in proportions of the two types of attriters. It was mainly male individual attriters who contributed to the bias by performing much worse than the male completers.

In a previous study of 13-year-olds, gender differences in structure, means, and variances of cognitive abilities were investigated within the framework of a

Requests for reprints should be sent to Monica Rosén, Department of Education and Educational Research, Göteborg University, P.O. Box 1010, S–431 26 Mölndal, Sweden. E-mail: Monica.Rosen@ped.gu.se

hierarchical model (Rosén, 1995). A multivariate latent variable model approach was adopted, and some rather striking gender differences were revealed. The model was based on information collected in two waves of data collection in the sixth grade. In the first wave, information from 13 rather-well-established ability tests were collected. Results from three standardized achievement (SA) tests (Math, Swedish, and English) were collected in the second wave.

No differences were found in the structure of the hierarchical model fitted to the data. At the apex of the hierarchy is *g* or General Intelligence, in which a substantial difference was found favoring girls. At the intermediate level, two broad factors were identified: Crystallized Intelligence (*Gc'*) and General Visualization (*Gv'*). On average, girls showed a higher mean on *Gc'*, whereas for boys there was both a higher mean and a greater variability on *Gv'*. Ten narrow ability dimensions were identified at the lowest level, and gender differences were found in 7 of them. Differences favoring boys were found in a narrow Verbal Ability factor (*V'*), Spatial Orientation factor (*Sr'*), Flexibility of Closure factor (*Cs'*), a narrow English Achievement factor (*EngAch'*), and a narrow Numerical Achievement factor (*NumAch'*). In *NumAch'*, the difference was quite large, and gender explained some 30% of the variance in the factor. Furthermore, the male group showed greater variability on Memory Span (*Ms'*) and Cognition of Figural Relations (*CFR'*). These results are quite striking because they seem to contradict much previous research on gender differences in the cognitive domain. As discussed more fully by Rosén (1995), the reason for this is that the multivariate approach with latent variables decomposes the observed variables into multiple components, whereas the analysis of observed variables only captures net effects of the differences in multiple variables. This approach is thus much more powerful in revealing patterns of performance in cognitive dimensions.

However, the study described may have problems related to missing data. The second data collection reduced the sample from 1,224 to 981 cases with complete data. The missingness was assumed to be random and thus not dealt with. If this assumption is not correct, the results found may at least partly be due to bias in the missing data. It is thus of greatest interest to investigate the influence of missing data in latent variable models. Both the pattern of gender differences and the structure of the hierarchical model itself may be affected.

Cases missing data, completely or partially, are a perennial problem in almost all studies within the social sciences. Various statistical methods and imputation techniques for handling these problems have been available for some time, although not so frequently applied (e.g., Little & Rubin, 1987; Madow, Nisselson, & Olkin, 1983; Madow & Olkin, 1983; Madow, Olkin, & Rubin, 1983; Rovine, 1994; Rubin, 1987). Some sources of bias are well known; it is, for example, often true that less successful individuals are more likely to lack data than others. However, the bias is typically assumed to be the same over groups—for example, gender, which in practice means that the missing data are handled as if they were random. A fair amount of the studies investigating gender differences in the cognitive domain have used samples that are reduced or restricted because of missing data or selection mechanisms, which may have caused bias in either the female or male direction (Becker & Hedges, 1984; Hyde, 1981).

The purpose of my study is to investigate the data missing at random (MAR) assumption. There are several reasons for assuming the pattern of gender differences to be affected by selection mechanisms, which are discussed later. In this study, the impact of missing data is investigated in both a univariate and multivariate manner. The main contribution is the approach to test the MAR assumption in a latent variable model, by utilizing the theory and methodology for improving validity in studies with missing data, as discussed in Allison (1987); Allison and Hauser (1991); and Muthén, Kaplan, and Hollis (1987). Two aims are in focus: to validate the hierarchical model of cognitive abilities from the previous analyses and to investigate the impact of missing data on reported estimates of gender differences in those latent hierarchically ordered ability dimensions (Rosén, 1995).

PROBLEMS OF MISSING DATA

The problem of missing data is often regarded as purely statistical, but the problem may also be addressed from a more substantial point of view, and in this way important mechanisms for understanding gender-related phenomena may be discovered. Thus, whether to take part in a study or not may be a decision that interacts more strongly with gender than gender interacts with the performance studied.

The data missing in this study are results on SA tests, which are given to help teachers assign comparable grades. In this study, it is possible to identify sources of missing data at two levels: missing individuals and missing classes.

Individuals may, of course, be prevented from taking part in the data collection occasion for different valid reasons (e.g., sickness). However, they may have "chosen" not to participate for one reason or another. Perhaps they did not want to go, or their parents or teachers did not want them to go. Gender-stereotyped expectations are frequently reported, particularly in math and science (e.g., Chipman, Brush, & Wilson, 1985; Fennema & Leder, 1990), which may be one of the reasons not to participate in standardized tests.

At the time of the data collection, the schools were free to decide whether to give the tests, so when whole classes are missing there may be other mechanisms involved in addition to those on the individual level. About half of the missing cases in my study come from missing classes. Fear of exposing achievement levels of individual classes or ideological reasons may underlie a school teacher's or a principal's decision to not participate. Opinions of parents could also be important in such a decision.

Variability differences, for example, also indicate that the gender groups are unequal (Cleary, 1992; Feingold, 1992a, 1992b, 1993; Hedges & Friedman, 1993a, 1993b).

The variability in achievement on some cognitive tests has often been found to be greater for boys than for girls. If it is mainly the low achievers who are missing in both groups, the male performance will be overestimated more than the female performance, given that the scores are approximately normally distributed in both groups.

Some recent findings concerning self-selection to the scholastic aptitude test in Sweden also provide support for the notion of gender bias in missing data (Mäkitalo & Reuterberg, 1996; Reuterberg, Westerlund, & Gustafsson, 1992). Although about as many boys as girls participate in the test, they do not seem to be selected in the same way with respect to ability. Boys seem to be more positively selected than girls.

There is also research that stresses lower self-esteem in learning ability among girls, which tends to be unrelated to their cognitive achievement (Byrne & Shavelson, 1989). Fennema (1980) argued that one reason for girls to avoid mathematics is lower confidence in learning mathematics and a belief that mathematics is not useful for them, together with the male belief that mathematics is a male domain. If these are reasons for girls to not participate in a math test situation, then it indicates that the girls with missing data may not be at the same achievement level as those boys with missing data. Language learning, however, is frequently reported as a female subject (e.g., Wernersson, 1989) although it has not been reported to affect boys' self-esteem in the subject. Still, indications of male avoidance of a language test were found in a large-scale study of grammar and vocabulary in foreign language (Balke-Aurell & Lindblad, 1980), in which there were more missing data for boys than for girls. Teachers' expectations may be one of the factors involved in these gender-correlated behaviors.

Male dominance in the classroom and girls' lesser visibility is frequently reported (Wernersson, 1977, 1989, 1991; Wilkinson & Marrett, 1985; öhrn, 1990, 1993), which supports claims that teaching and learning in relation to boys becomes the norm (Einarsson & Hultman, 1984; Hirdman, 1987). If male performance is more influential in determining whether a class will participate in a test, the missing data pattern will be biased and will be another reflection of the male norm.

In this section, I addressed the notion that missing data generating mechanisms may be interacting with gender and that several underlying mechanisms at different levels may be working simultaneously. Thus, it is reasonable to expect bias with respect to gender; however, it was ignored in my previous research. It thus is necessary to investigate the impact of missing data.

DESIGN AND METHODOLOGY

Sample and Variables

The sample consists of most sixth graders (age 12 to 13) in two communities, Mölndal and Kungsbacka (50 classes) close to Göteborg in the western part of

Sweden. The first data collection took place in 1980. The test battery of 13 ability tests with some questionnaires was administered to 1,254 pupils in the sixth grade (age 12–13). As a second data collection, results from 3 SA tests in the same grade were collected.

For different reasons the battery of cognitive tests could not be administered to 30 pupils, which left a final sample of 1,224 pupils—602 boys and 622 girls. There is no information on these 30 students, and they are not included in any further analyses.

The test battery was assembled in such a way that enough primary factors would be represented to make possible identification of the broad factors *Gf* (Fluid Intelligence), *Gc*, and *Gv*. It consists of two parts: one with 13 rather well established ability tests that covers primary factors mainly for *Gf* and *Gv*, and one with three SA tests in Swedish, English, and mathematics, which capture the *Gc* dimension. A description of variables collected in 1980 is presented next. The procedures for the data collection and the aptitude variables are described in great detail elsewhere (Gustafsson, 1984; Gustafsson, Lindström, & Björck-Åkesson, 1981). A short presentation of each instrument's name, short labels, number of items, and the latent variables that they originally were hypothesized to measure are given in Table 1.

The scores on all subtests of the three SA tests (Swedish, English, and mathematics) were collected a few months after the test battery had been administrated. The purpose of the SA tests is to calibrate the grades on class level in Sweden. However, in the research setting these tests serve as indicators of primarily *Gc* and perhaps some more narrow dimensions. As has already been mentioned, the second wave of data collection caused some attrition. At this time in Sweden in some of the schools, the teacher could decide whether the class should participate in the SA test. In Table 2, the proportion of boys and girls for different types of attrition is presented.

Out of the 1,224 students, there were 243 students who had not completed all three SA tests. The frequency of missing data is about the same among the sexes–20% in each group, or 122 boys and 121 girls. Out of these 243 missing students, 47% are missing individuals, and 53% come from missing classes (five classes). It should be mentioned that out of these five classes, three miss all three SA tests, whereas two have data on one or two of the SA tests. The same pattern is true for the missing individuals. However, in this study, participants with incomplete data on any of the three SA tests are coded as missing.

This leaves a sample of 981 students with complete data on 13 ability tests and 3 SA tests, and a second subsample of 243 subjects with complete data on 13 ability tests. Small differences between the sexes in proportions of missing data on individual and class level do not suggest any gender bias in terms of the origin of the missing data.

As in previous multivariate analysis with a hierarchical latent variable model fitted to these data, four of the ability tests have been split in half in the multivariate

TABLE 1
Description of the Test Battery in Grade 6

Test Name	*Short Label*	*Number of Items*	*Factors*	*Comment*
Opposites–odd items	Op–O	20	*g, Gc, V*	Half of one test
Opposites–even items	Op–E	20	*g, Gc, V*	Half of one test
Auditory Number Span	ANS	19	*g, Ms*	
Auditory Letter Span	ALS	19	*g, Ms*	
Number Series II	NS	20	*g, NumAch*	
Letter Grouping II	LG	20	*g*	
Raven Matrices–odd items	Ra–O	23	*g, Gv, CFR*	Half of one test
Raven Matrices–even items	Ra–E	22	*g, Gv, CFR*	Half of one test
Metal Folding–odd items	MF–O	15	*g, Gv, Vz*	Half of one test
Metal Folding–even items	MF–E	15	*g, Gv, Vz*	Half of one test
Group Embedded Figures	GEFT	9	*g, Gv*	9 out of 16 items given[a]
Hidden Patterns	HP	400	*g, Gv*	
Copying	Co	64	*g, Gv*	
Card Rotations Part I	CR–I	14	*g, Gv, Sr*	First part of one test
Card Rotations Part II	CR–II	14	*g, Gv, Sr*	Second part of one test
Disguised Words	DW	24	*g, Gv, Cs*	
Disguised Pictures	DP	24	*g, Gv, Cs*	
Swedish standardized achievement tests				
Spelling	Sw Sp	25	*g, Gc*	
Reading Comprehension	Sw RC	21	*g, Gc*	
Words of Relation	Sw WR	12	*g, Gc'*	
Vocabulary	Sw Vo	25	*g, Gc*	
Word List	Sw WL	11	*g, Gc*	
Sentence Construction	Sw SC	18	*g, Gc*	
English standardized achievement test				
Vocabulary	Eng Vo	40	*g, Gc, EngAch*	
Listening Comprehension	Eng LC	35	*g, Gc, EngAch*	
Forms and Structures	Eng FS	40	*g, Gc, EngAch*	
Reading Comprehension	Eng RC	24	*g, Gc, EngAch*	
Math standardized achievement tests				
Percent Calculation	Ma PC	16	*g, Gc, NumAch*	
Estimates	Ma Est	21	*g, Gc, NumAch*	
Geometry and Diagrams	Ma GD	6	*g, Gc, NumAch*	
Applied Computation	Ma AC	12	*g, Gc, NumAch*	
Numerical Calculation	Ma NC	20	*g, Gc, NumAch*	

Note. Aptitude variables and their relation to factors included in the hierarchical model in Grade 6.
[a]As recommended by Witkin, Oltman, Raskin, and Karp (1971).

TABLE 2
Proportion of Girls and Boys With Different Types of Attrition

	Complete Data[a]			*Missing Classes*[b]			*Missing Individuals*[c]			
	n	*Col%*	*Row%*	*n*	*Col%*	*Row%*	*n*	*Col%*	*Row%*	*Total*
Girls	501	51.1	80.5	70	54.3	11.3	51	55.3	8.2	100
Boys	480	48.9	79.7	59	45.7	9.8	63	44.7	10.5	100

Note. Col% = column percent.
[a]$n = 981$. [b]$n = 129$. [c]$n = 114$.

analysis to identify some more narrow ability dimensions (Opposites, Metal Folding, Raven Progressive Matrices, and Card Rotations).

Analysis Procedures

The first part of the analysis consists of ordinary univariate analysis in terms of descriptives, *t* test, variance ratios (VRs), and correlations. A multiple-regression model is fitted to the data for investigating the interaction effect of gender and missing data on the observed variables. Dependent variables are the 13 ability tests, and independent variables are gender, missing data (also labeled *attrition* or *the attriters*), and the cross-product of gender and attrition. The group with complete data is also referred to as *the completers*. All independent variables are handled as dummy variables, where the completers are coded 0 and the attriters coded 1, and boys are coded 0 and girls are coded 1. The reference group in the regression equation thus is boys in the group with complete data. The mean of the reference group equals the constant (a) in the regression equation.

$$y = a + b_1\text{Gender} + b_2\text{Attrition} + b_3\text{Gender} \times \text{Attrition}$$

The *y*-mean may be estimated for each of the subgroups by summing those coefficients for which a bullet (•) has been assigned in Table 3.

The regression analysis will inform us if there are differences between the completers and the attriters, if there are gender differences, and if such differences vary between the two groups. The model also informs us about within-male differences between completer and attriters. Significant within-gender differences are investigated by computing *t* test for means and *F* test for VRs between the two groups.

The group with missing data is also divided into a "missing classes" group and a "missing individuals" group. Mean and VRs for these two groups are compared with the complete data group, and gender differences within and between these groups are investigated in the same fashion.

TABLE 3
The Multiple Regression Model for Investigating the Effects of Gender and Attrition

Group			*Gender*	*Attrition*	*Gender × Attrition*
Gender	*Missing Group*	*a*	b_1	b_2	b_3
0 boys	0 complete data	•			
0 boys	1 missing data	•	•		
1 girl	0 complete data	•		•	
1 girl	1 missing data	•	•		•

The second part of the analysis investigates the impact of missing data in a multivariate fashion with latent variables further described next. First, the model as such is investigated in terms of validity, and then gender differences are investigated. Thus, beyond investigating the impact of missing data on gender differences in latent dimensions, this part also serves as a validation study of the previously reported hierarchical model for these data (Rosén, 1995).

Handling Missing Data in Structural Equation Modeling

It has been demonstrated that missing data techniques may be utilized on several types of data patterns—for example, monotone (Rovine, 1994) and matrix sampling designs (Muthén, Khoo, & Nelson Goff, 1994). Rovine's review indicated that although several methods for dealing with missing data exist, there are very few simulations or real examples of how they work. This study utilizes the method discussed by Allison (1987) and Muthén et al. (1987), which Rovine referred to as "direct parameter estimation" (p. 204).

Two types of model estimators for maximum likelihood are to be compared. One is referred to as *listwise quasi likelihood* (LQL; see Muthén et al., 1987), which are maximum likelihood estimates based on the sample with complete data (n = 981). The LQL model assumes data to be missing completely at random (MCAR). In this case, this is the same model that was analyzed in the previous study (Rosén, 1995). The other estimator is referred to as *full quasi likelihood* (FQL), which is the correctly specified model where the 241 attriters are included (n = 1,224) in the maximum likelihood estimates on those variables they do share with the group that has complete data. In Figure 1, the LQL and FQL samples are represented.

The LQL estimation utilizes only the shaded area, whereas the FQL estimation utilizes all available information (i.e., from all 1,224 cases). This is a very common situation of missing data, and one of the more uncomplicated ones to be handled in the proposed fashion of Muthén et al. (1987).

To obtain the correct chi-square and degrees of freedom (*df*) for testing the fit of the FQL model, two models have to be fitted to the data: The so-called H_0 model,

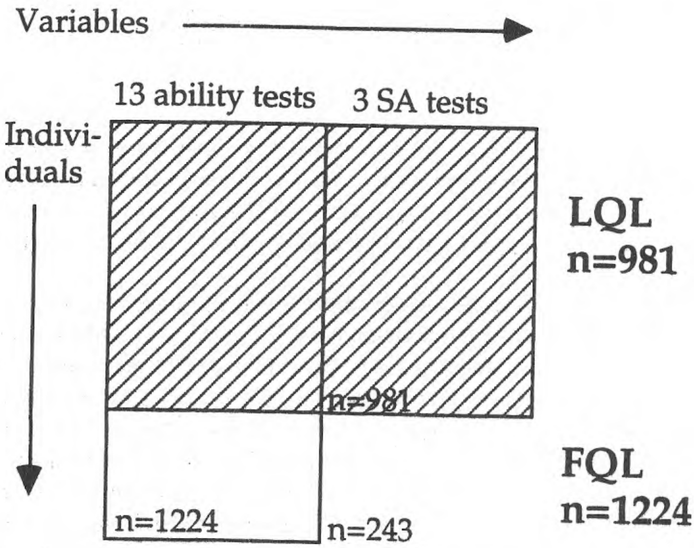


FIGURE 1 Graphical representation over the data pattern in the sample. The sample size is 981 for the subsample with complete data for listwise quasi likelihood (LQL) estimation. For the full quasi likelihood (FQL) estimation, the 243 cases lacking three SA tests are included through a multiple-group analysis.

on which model restrictions are imposed on the data, and the H_1 model, which tests if the covariance matrices and the mean vectors for the groups come from the same population. The correct fit statistic is obtained when the chi-square and *df* of the H_1 model is subtracted from the H_0 model. According to Muthén et al. (1987), the FQL H_1 model informs us if our MCAR assumption is true in our LQL model. This is true if the chi-square in the H_1 model is insignificant. If the chi-square is significant, then the LQL is misspecified—that is, it does not give maximum likelihood estimates, and MCAR assumption does not hold.

RESULTS

The results of the analysis are presented in three steps. In the first step, the impact of gender and missing data on the observed results on the 13 ability tests is investigated. In the second step, the latent hierarchical measurement model is investigated by comparing the LQL model estimates with the FQL model estimates. The impact of missing data on gender differences is investigated in the third step through comparison of correlations between gender and the latent cognitive ability dimension in the LQL and FQL models.

Impact of Missing Data on Observed Achievement on 13 Ability Tests

Several questions can be answered: First, are there differences between those with complete and incomplete data on the following tests? Furthermore, are there gender or Gender × Attrition interaction effects? Table 4 presents descriptive information, and in Table 5, results from the regression model where gender, attrition, and the Gender × Attrition interaction cross-product are used to predict the ability tests scores.

In Table 4, means and standard deviations for the different subsets of the sample are presented. The results for the subsample of 981 participants with complete data do not seem alarmingly different from the results for the full sample of 1,224. Knowing that about as many girls as boys are missing, the MAR assumption in the previous study (Rosén, 1995) does not appear unreasonable. However, the means are slightly lower among the attriters, and the standard deviations are slightly larger.

As reported earlier for the completers (Rosén, 1995), significant gender differences favoring girls can be seen on Letter Grouping, Raven Progressive Matrices, Opposites, and Hidden Patterns, and there is a tendency toward female advantage on Copying. Male completers perform better on Number Series and have a tendency toward higher performance on Card Rotations. The missing data group shows a significantly lower performance level on all the tests, except on four Gv tests. The

TABLE 4
Means and Standard Deviations on the Ability Tests in Grade 6 for Subsets of the Sample for Which Different Variables Have Been Observed

	13 Ability Tests[a]		*Includes 3 SA Tests*[b]		*3 SA Tests Missing*[c]	
	M	*SD*	*M*	*SD*	*M*	*SD*
Number Series II	7.82	3.75	8.04	3.78	6.93	3.51
Letter Grouping II	10.96	3.61	11.18	3.56	10.10	3.96
Raven Progressive Matrices	24.16	5.96	24.44	5.83	23.05	6.36
Auditory Number Span	4.41	2.68	4.49	2.68	4.11	2.64
Auditory Letter Span	4.52	2.17	4.57	2.20	4.33	2.04
Metal Folding	18.53	6.13	18.70	6.00	17.82	6.58
Group Embedded Figures	10.60	2.89	10.69	2.79	10.26	3.24
Hidden Patterns	70.42	23.99	71.08	23.73	67.74	24.91
Copying	22.42	8.19	22.67	8.12	21.44	8.42
Card Rotations	96.39	38.05	97.37	37.32	92.45	40.68
Disguised Words	11.72	3.41	11.89	3.36	11.02	3.56
Disguised Pictures	12.68	3.25	12.79	3.28	12.23	3.09
Opposites	21.81	5.75	22.08	5.69	20.72	5.83

Note. SA = standardized achievement.
[a]$n = 1{,}224$. [b]$n = 981$. [c]$n = 243$.

TABLE 5
Effect of Gender, Attrition, and Gender × Attrition Interaction on the Achievement on 13 Ability Tests

	Gender[a]		*Attrition*[b]		*Gender × Attrition Interaction*		*% Explained Variance*
	Beta	*p*	*Beta*	*p*	*Beta*	*p*	$R^2 \times 100$
Number Series II	–.07	.02	–.18	.00	–.11	.01	2.02
Letter Grouping II	.23	.00	–.16	.00	.08	.06	8.17
Raven Progressive Matrices	.10	.00	–.13	.00	–.05	.20	2.47
Auditory Number Span	.02	.50	–.09	.03	.05	.28	0.56
Auditory Letter Span	.03	.28	–.08	.04	.06	.15	0.66
Metal Folding	.02	.50	–.09	.03	–.05	.27	0.56
Group Embedded Figures	.03	.30	–.09	.02	–.05	.25	0.70
Hidden Patterns	.15	.00	–.06	.11	.02	.71	2.98
Copying	.06	.06	–.06	.13	.01	.95	0.73
Card Rotations	–.06	.06	–.04	.32	–.02	.67	0.71
Disguised Words	–.01	.55	–.14	.00	.06	.13	1.21
Disguised Pictures	–.05	.10	–.06	.14	–.01	.74	0.80
Opposites	.05	.00	–.17	.07	.11	.01	2.29

Note. Boys are coded 0, girls are coded 1. Complete data are coded 0, attrition are coded 1. Reference group is boys with complete data. Beta is the standardized regression coefficient.
[a]$n = 981$. [b]$n = 243$.

Gender × Attrition interaction indicates a deviating pattern of gender differences in the group lacking data compared to the completers. Among attriters, a greater female advantage can be seen on Opposites and Letter Grouping. Both these tests are indicators of *g* or *Gf.* Otherwise, the pattern of gender differences in the group lacking data is the same as in the complete data group.

In Table 6, observed variances are compared within and between boys and girls among both attriters and completers. VRs have been obtained by dividing the variance for one group by the variance for the other. In the two left columns, variances for completers has been divided by variances for attriters for boys and girls separately. VRs below 1 indicate higher variability for the attriters. In the two right columns, variances for boys have been divided by variances for girls for completers and attriters separately. Here, VRs below 1 indicate higher variability for the girls. *F* tests are computed for test of significance of the found variability differences.

TABLE 6
Variance Comparison Between and Within the Genders in the Group With Complete Data and in the Group Lacking Data

	Boys		*Girls*		*Completers*		*Attriters*	
	Completers/Attriters		*Completers/Attriters*		*Boys/Girls*		*Boys/Girls*	
	Df 479/121 $F_{crit} = 1.27$		*Df 500/120* $F_{crit} = 1.27$		*Df 479/500* $F_{crit} = 1.16$		*Df 121/120* $F_{crit} = 1.36$	
Test	*VR*	*F*	*VR*	*F*	*VR*	*F*	*VR*	*F*
Number Series II	1.30	1.30*	1.02	1.02	1.33	1.33*	1.03	1.03
Letter Grouping II	1.05	1.05	0.94	1.96*	0.98	1.02	0.88	1.14
Raven Progressive Matrices	0.92	1.05	0.79	1.27*	1.36	1.36*	1.17	1.17
Auditory Number Span	1.00	1.00	1.12	1.12	1.39	1.39*	1.57	1.57*
Auditory Letter Span	1.14	1.14	1.23	1.23	1.08	1.08	1.16	1.16
Metal Folding	0.99	1.01	0.70	1.43*	1.29	1.29*	0.91	1.09
Group Embedded Figures	0.70	1.43*	0.82	1.22	1.08	1.08	1.26	1.26
Hidden Patterns	1.00	1.00	0.83	1.20	0.94	1.06	0.78	1.28
Copying	0.72	1.40*	1.26	1.26	0.90	1.11	1.59	1.59*
Card Rotations	0.92	1.09	0.85	1.17	1.04	1.04	0.97	1.03
Disguised Words	0.80	1.25	1.01	1.01	1.03	1.03	1.29	1.29
Disguised Pictures	0.92	1.08	1.47	1.47*	1.05	1.05	1.68	1.68*
Opposites	1.01	1.01	0.99	1.01	0.97	1.01	0.97	1.01

Note. Variance ratios (VRs) > 1 indicate greater variability for the group with complete data in the two left-most columns, and for boys in the two right-most columns.
*$p = .05$ (approximate).

Variability differences among boys are found in four of the tests: three spatial tests and one test of *g* or *Gf*. In the spatial tests, the male attriters group is the more variable, and in the *g* or *Gf* test, the male completers are more variable.

Variability differences are also found among girls, again in three spatial tests and in one test of *g* or *Gf*, although not in the same tests as the male group. In two of the spatial tests, female completers are more variable, and in the other two the female attriters are more variable.

When male and female variability is compared, differences are found in four tests among the completers and in three tests among the attriters, all showing a greater male variability. Again, it is not the same tests that show variability differences in the complete data group as in the group missing data. It should also be noted that no difference in variability is found in most of the tests. Because the tests that show variability differences vary over groups, it is difficult to interpret the pattern, and some of the differences may be random effects.

It may be concluded from this that the attrition is not entirely random. The 243 missing cases achieve somewhat differently in terms of means, variability, and gender differences than do those who have complete data. It appears that male attriters are less successful on most ability tests than are female attriters, which may have an impact on the pattern of gender differences in the complete data group.

Comparing Class- and Individual-Level Attrition

In Tables 7, 8, and 9, attrition has been divided into two categories: *class-level attrition* and *individual-level attrition.* In Table 7, the class-level attrition and the individual-level attrition are compared in terms of means and variability with the groups having complete data. No mean differences are found between the class-level attrition and the complete data group. In five tests, variability differences are found—four of them indicating that the group with complete data has a greater variability. The reverse pattern is found when the individual-level attrition is compared to the complete data group. Mean differences favoring the completers

TABLE 7
Group Comparison of Means and Variances on 13 Ability Tests

	Group With Complete Data[a] Versus Group Lacking Data on the Class Level[b]			*Group With Complete Data[a] Versus Group Lacking Data on the Individual Level[c]*		
	t	*VR*	*F*	*t*	*VR*	*F*
Number Series II	–.1.37	1.18	1.38*	4.84*	1.01	1.03
Letter Grouping II	1.55	1.05	1.55*	4.84*	0.90	1.22
Raven Progressive Matrices	0.82	1.04	1.08	4.49*	0.86	1.36*
Auditory Number Span	1.16	1.13	1.27*	1.83	0.92	1.18
Auditory Letter Span	0.74	1.19	1.42*	1.56	0.95	1.05
Metal Folding	0.17	0.96	1.11	2.79*	0.87	1.31*
Group Embedded Figures	–0.98	0.88	1.29*	4.30*	0.89	1.26*
Hidden Patterns	–0.33	1.04	1.08	3.34*	0.90	1.19
Copying	–0.72	1.06	1.13	3.97*	0.92	1.18
Card Rotations	–0.08	1.00	1.03	2.59*	0.88	1.31*
Disguised Words	1.34	1.09	1.20	4.05*	0.84	1.42*
Disguised Pictures	1.27	1.12	1.25	2.32*	1.00	1.01
Opposites	1.09	1.00	1.09	3.59*	0.99	1.08

Note. t values > 1.96 or < –.1.96 indicate significance at the .05 level. Positive signs on t values indicate higher means, and variance ratios (VRs) < 1 indicate greater variance for the group with complete data. With 1000/100 degrees of freedom the F_{crit} is approximately 1.26 on the 5% level.
[a]$n = 981$. [b]$n = 129$. [c]$n = 114$.
*$p = .05$ (approximate).

TABLE 8
Within-Gender Comparison of Means and Variances on 13 Ability Tests

	Girls With Complete Data[a] *Versus Girls Lacking Data on the Class Level*[b]			*Girls With Complete Data*[a] *Versus Girls Lacking Data on the Individual Level*[c]		
	t	*VR*	*F*	*t*	*VR*	*F*
Number Series II	0.94	1.10	1.21	0.85	0.91	1.09
Letter Grouping II	0.62	1.05	1.05	2.03*	0.89	1.25
Raven Progressive Matrices	−0.47	1.33	1.41*	2.79*	0.77	1.83*
Auditory Number Span	0.52	1.16	1.34	0.53	0.95	1.11
Auditory Letter Span	0.37	1.32	1.73*	2.03*	0.89	1.25
Metal Folding	−0.21	0.95	1.12	1.29	0.77	1.66*
Group Embedded Figures	−0.90	0.90	1.24	2.17*	0.95	1.11
Hidden Patterns	0.34	1.00	1.00	1.34	0.82	1.49*
Copying	0.59	1.19	1.42*	1.71	1.05	1.10
Card Rotations	0.82	0.97	1.07	1.47	0.90	1.25
Disguised Words	1.18	1.05	1.11	1.00	0.94	1.13
Disguised Pictures	1.84	1.39	1.95*	1.07	1.04	1.09
Opposites	−0.26	1.01	1.03	1.05	0.97	1.09

Note. t values > 1.96 or < −1.96 indicate significance at the .05 level. Positive signs on t values indicate higher means, and variance ratios (VRs) < 1 indicate greater variance for girls with complete data. With 500/70 degrees of freedom the F_{crit} is approximately 1.37 on the 5% level. With 500/50 degrees of freedom the F_{crit} is approximately 1.46 on the 5% level.

[a]$n = 501$. [b]$n = 70$. [c]$n = 51$.

*$p = .05$ (approximate).

are found in every test except two. Variability differences are found in five of the tests, indicating greater variability among individual attriters. In the next step, the same comparison is performed for each gender. In Table 8, the results for the girls are presented.

There are no mean differences between girls of the class-level attrition group and female completers, although differences in variability are found in four of the tests, with all four indicating greater variability for the female completers. In four tests, mean differences are found when the girls of the individual-level attrition group are compared with the female completers. Differences in variability are found in three tests, all indicating greater variability among the female attriters on the individual level. Results from similar analysis for boys are presented in Table 9.

There is a tendency for mean differences between boys of the class-level attrition group and male completers in two of the tests, both in favor of the latter. Significant variability differences are found in one test only, indicating greater variability among boys with complete data. Boys of the individual-level attrition group show significantly lower results on almost every test compared to the male completers. Variability differences are present in three of the tests, indicating greater variability

TABLE 9
Within-Gender Comparison of Means and Variances on 13 Ability Tests

	Boys With Complete Data[a] *Versus Boys Lacking Data on the Class Level*[b]			*Boys With Complete Data*[a] *Versus Boys Lacking Data on the Individual Level*[c]		
	t	*VR*	*F*	*t*	*VR*	*F*
Number Series II	0.94	1.24	1.53*	5.60*	1.19	1.42*
Letter Grouping II	1.94	1.14	1.30	4.49*	0.97	1.07
Raven Progressive Matrices	0.87	0.96	1.16	3.35*	0.99	1.03
Auditory Number Span	1.13	1.09	1.20	1.83	0.92	1.15
Auditory Letter Span	0.72	1.09	1.18	2.29*	1.05	1.11
Metal Folding	0.93	1.07	1.09	2.49*	0.98	1.06
Group Embedded Figures	–0.44	0.85	1.37	3.71*	0.86	1.36
Hidden Patterns	–0.72	1.06	1.13	3.11*	1.00	1.01
Copying	–1.58	0.94	1.12	3.73*	0.84	1.41*
Card Rotations	–0.89	1.05	1.09	2.26*	0.85	1.39
Disguised Words	0.68	1.14	1.30	4.49*	0.80	1.57*
Disguised Pictures	–0.12	0.97	1.30	2.22*	0.97	1.06
Opposites	1.93	1.02	1.18	3.83*	1.00	1.00

Note. *t* values > 1.96 or < –1.96 indicate significance at the .05 level. Positive signs on *t* values indicate higher means and variance ratios (VRs) < 1 indicate greater variance for boys with complete data. With 500/60 degrees of freedom the F_{crit} is approximately 1.41 on the 5% level.

[a]$n = 480$. [b]$n = 59$. [c]$n = 63$.

*$p = .05$ (approximate).

for boys, with complete data in one of the tests and greater variability for the other group in two. The pattern of means and variances indicates that there may exist a restriction of range among male completers due to the attriters' lower means.

The contribution from the class-level attrition to the missing data bias is not very large. It seems that missing individuals are causing most of the observed bias, and boys contribute more to the bias than do girls. The male attriters, particularly on the individual level, deviate more in terms of means from male completers than do the missing girls from female completers.

In terms of variability, girls of the individual-level attrition show greater variability in a few tests than do the female completers. The reverse pattern is found among boys.

Impact of Missing Data on the Hierarchical Model of Cognitive Ability Dimensions

A simple test of the MAR assumption is obtained by comparing covariance matrices of the completers and the attriters, including only the 17 variables they have in common. A test of significance is obtained by running an ordinary two-group analysis in LISREL (Jöreskog & Sörbom, 1988, 1993), with constraints of equality on every element of the covariance matrix and the means. The test resulted in the following fit statistics: $\chi^2(170, N = 1{,}224) = 221, p = .005$, which indicates significant differences between the groups. Thus, the MAR assumption does not hold true.

The analyses that follow concern the structure of the model—that is, whether the attrition suggests any different pattern than does the LQL model (the model fitted to the data for the 981 completers). Modification indices for the FQL model (the model fitted to the data when missing cases are included and the sample size is 1,224) do not indicate any other structure than the LQL model. Fit indices do not differ much in the FQL model from the LQL models. The fit statistics (before correction) for the FQL H_0 model are $\chi^2(934, N = 1{,}224) = 948$, root mean square error of approximation (RMSEA) = .003, RMSEA $p < .05 = 1$, Normed Fit Index (NFI) = .91, and Nonnormed Fit Index (NNFI) = 1.0. For the FQL H_1 model, $\chi^2(560, N = 1{,}224) = 221$. The correct fit statistics for the FQL model are 374 for *df* (934–560) and 727 for χ^2 (948–221). The fit of the FQL model thus is approximately the same as the fit of the LQL model: $\chi^2(406, N = 1{,}224) = 726$, RMSEA = .028, RMSEA $p < .05 = 1$, NFI = .96, and NNFI = .98.

The next question to be answered concerns the estimates of the parameters. In Table 10, LQL estimates are compared to FQL estimates. Most standardized factor loadings are about the same, except for the loadings on *g* and *Gv′*, where almost every loading is slightly higher in the FQL model than in the LQL model. The opposite is true for the loadings on *Gc′* and on most narrow factors. However, the discrepancy between the models is small. Differences between the LQL and FQL standardized factor loadings of 0.02 or more have been marked with bold style.

The amount of variance explained by the factors in each observed variable is obtained by squaring the standardized factor loading. In most cases, the amount of explained variance in the observed variables has either increased or decreased with 1% or less. The largest difference in explained variance is obtained for Disguised Words by *Cs′* (Speed of Closure), which decreases from 48% in the LQL model to 42% in the FQL model. The increase or decrease in explained variance for the rest of the variables ranges from 1% to 4%. The *t* values for the factor loadings are also about the same in the FQL model as in the LQL model.

The conclusion of the comparison of factor loadings is that the differences are small but systematic. Most of the FQL estimates for *g* and *Gv′* are higher than the LQL estimates, whereas the FQL estimates for the rest of the factors are equal or lower than the LQL estimates. A similar pattern is obtained for the factor variances, as can be seen in Table 11. The overall deviation from the LQL model is small. The factor variance increases somewhat in *g* and *Gv′*, whereas the rest remains fairly stable or decreases a little. The *t* values for the factor variances follow the same pattern.

TABLE 10
Loadings of the Aptitude Variables on the Ability Factors in Grade 6 in the LQL and FQL Models

	g		*Gv′*		*Gc′*		*Ms′*		*Cfr′*		*Vz′*		*Sr′*	
	LQL	*FQL*	*LQL*	*FQL*	*LQL*	*FQL*	*LQL*	*FQL*	*LQL*	*FQL*	*LQL*	*FQL*	*LQL*	*FQL*
Op–O	.56	.57			.31	.29								
Op–E	**.54**	**.57**			.33	.31								
NS	.75	.75												
LG	.70	.70												
ANS	.28	.28					.57	.56						
ALS	.34	.35			**.12**	**.10**	.70	.70						
Ra–O	**.54**	**.56**	.20	.20					**.73**	**.71**				
Ra–E	**.57**	**.60**	.18	.18					.64	.63				
MF–O	**.48**	**.50**	**.64**	**.61**							.40	.40		
MF–E	.50	.51	.62	.63							.38	.38		
GEFT	**.57**	**.59**	**.30**	**.32**										
HP	.56	.57	.24	.25										
Co	**.53**	**.55**	**.25**	**.27**										
CR–I	**.43**	**.45**	**.19**	**.21**									.58	.57
CR–II	.50	.51	.25	.26									.69	.68
DW	**.33**	**.37**												
DP	**.22**	**.25**	**.23**	**.25**										
Ma PC	.55	.55	.10	.10	.25	.25								
Ma Est	.53	.53			.10	.10								
Ma GD	.65	.66	.07	.08	.19	.18								

(*Continued*)

TABLE 10
(*Continued*)

	g		*Gv′*		*Gc′*		*Ms′*		*Cfr′*		*Vz′*		*Sr′*	
	LQL	*FQL*	*LQL*	*FQL*	*LQL*	*FQL*	*LQL*	*FQL*	*LQL*	*FQL*	*LQL*	*FQL*	*LQL*	*FQL*
MA AC	.70	.70			.14	.14								
Ma NC	.63	.63			.23	.22								
Eng Vo	.58	.59			.51	.51								
Eng LC	**.52**	**.54**			.34	.33								
Eng FS	.60	.61			.50	.49								
Eng RC	.59	.60			.49	.48								
Sw Sp	.50	.51			.54	.54								
Sw RC	.60	.61			**.35**	**.33**								
Sw WR	.61	.62			.41	.40								
Sw Vo	.53	.54			.39	.38								
Sw WL	.65	.66			.32	.31								
Sw SC	.58	.59			.38	.37								
Op–O									**.32**	**.36**	.42	.41		
Op–E									**.33**	**.28**	.44	.44		
NS					.24	.24								
LG														
ANS														
ALS														
Ra–O														
Ra–E														

(*Continued*)

MF–O												
MF–E												
GEFT			**.23**	**.21**								
HP			.44	.43								
Co			**.48**	**.46**								
CR–I			.32	.31								
CR–II			.25	.25								
DW	**.69**	**.65**	.11	.12								
DP	**.43**	**.41**										
Ma PC					.50	.50						
Ma Est					.44	.44						
Ma GD			.11	.10	.44	.43						
MA AC					.40	.40						
Ma NC					.50	.50						
Eng Vo							.45	.45	.09	.08		
Eng LC							.51	.51	.24	.23		
Eng FS							.35	.34				
Eng RC							.39	.39	.17	.17		
Sw Sp												
Sw RC									.36	.36		
Sw WR									.19	.19		
Sw Vo									.47	.47	.12	.12
Sw WL												
Sw SC												

Note. Refer to Table 1 for test name definitions. LQL = listwise quasi likelihood; FQL = full quasi likelihood.

Impact of Missing Data on Gender Differences in Hierarchically Ordered Ability Dimensions

In the final step, gender has been entered into the model as a dummy variable, in which boys are coded 0 and girls are coded 1. Positive estimates indicate female advantage and negative estimates indicate male advantage. The FQL model does not indicate a very different pattern of gender differences on the latent variables from the LQL model, but the magnitude of the differences has changed somewhat.

As can be seen in Table 12, the FQL estimates differ slightly from the LQL estimates. The difference with respect to *g*, favoring girls, increases somewhat, as does the difference in *Gv′*, *Sr′*, and *Cs′* favoring boys. Every difference with respect to the other ability dimensions decreases somewhat. Gender explains 1% more of *g*, *Gv′*, *Sr′*, and *Cs′* in the FQL model than in the LQL model. The amount of explained variance drops 4% in *V′*, 2% in *NumAch′*, and 1% in *Gc′*. In some cases, rather large differences in the *t* value for the two models are observed. This is likely to be due to the standard errors not being correct under the LQL model (Muthén et al., 1987). The differences in *t* values between the LQL and the FQL models seem to be a function of the amount explained variance.

The final conclusion about the gender differences among the broad ability dimensions is that the FQL model increases the advantage for girls with respect to g and for boys with respect to *Gv′*, and it decreases the female advantage in *Gc′*.

TABLE 11
Comparison Between LQL and FQL Estimates of Maximum Likelihood and *t* Values for the Factor Variances

	Factor Variance		*t Values*	
	LQL	*FQL*	*LQL*	*FQL*
g	3.17	3.41	8.81	10.14
Gv′	0.33	0.35	2.69	3.10
Gc′	0.97	0.87	5.00	4.96
Gs′	1.11	1.04	5.29	5.55
V′	1.77	1.79	4.68	4.74
Ms′	2.33	2.28	12.43	13.79
NumAch′	4.23	4.24	8.38	8.44
CFR′	4.36	4.27	16.70	18.42
Vz′	1.53	1.57	2.23	2.74
Sr′	1.60	1.59	14.50	15.99
Cs′	1.94	1.79	9.64	10.18
Eng Ach′	1.20	1.19	8.50	8.44
Word′	1.79	1.74	7.26	7.32

Note. LQL = listwise quasi likelihood; FQL = full quasi likelihood.

TABLE 12
Significant Relations Between Gender and Latent Variables in a One-Group Model for Grade 6: Comparison Between LQL and FQL Estimates of Maximum Likelihood

	r		$r^2 \times 100$		*t Value*	
	LQL	*FQL*	*LQL*	*FQL*	*LQL*	*FQL*
g	.22*	.25*	5	6	5.41*	4.57*
Gv′	–.15*	–.18*	2	3	–2.70*	–2.93*
Gc′	.26*	.24*	7	6	4.58*	3.51*
Gs′						
V′	–.46*	–.41*	21	17	–6.37*	–4.18*
Ms′						
NumAch′	–.55*	–.53*	30	28	–11.28	–5.10
CFR′						
Vz′						
Sr′	–.24*	–.26*	6	7	–5.93	–4.56
Cs′	–.14*	–.16*	2	3	–3.04	–3.24
Eng Ach′	–.19*	–.18*	4	4	–3.33	–2.84
Word′						

Note. Negative sign indicates advantage for boys, positive figures advantage for girls. *t* values > 1.96 or < –1.96 indicate significance at the .05 level. LQL = listwise quasi likelihood; FQL = full quasi likelihood.
*$p = .05$.

Among the more narrow dimensions, the FQL model decreases the male advantage on both achievements factors *NumAch′* and *EngAch′* and on *V′*, and it increases the male advantage on the narrow spatial dimensions *Cs′* and *Sr′*.

DISCUSSION AND CONCLUSIONS

In this study, I demonstrated that the pattern of gender differences on the 13 ability tests varies depending on whether the attriters in SA tests are included. It has also been demonstrated that the attrition group, which performs lower, alone has an impact on test performance and that there is a Gender × Attrition interaction for some variables. Thus, the group of participants with missing data on some variables is not as successful in general performance as those who have complete data. The analysis of differences between the completers' and attriters' covariance matrices and mean vectors on common variables also showed significance. In the analyses, the LQL model was thus rejected and replaced with the FQL model. The differences between the results of the LQL and FQL models in this study may be regarded as modest but interesting in terms of possible interpretations and consequences. The assumption that data are MCAR in this study thus is not tenable.

The deviation from the hierarchical model from the previous analysis, in which participants with missing data were excluded, is very small. The influence of *g* increases between .05 and .10 on most variables, and so does the influence of *Gv′* on the spatial tasks. The influence from the rest of the factors remains stable or decreases somewhat. Differences in standardized loadings between the LQL model and the FQL model are never larger than .05. The same pattern was found for the factor variances (e.g., *g* and *Gv′* increased somewhat), whereas the rest of the factor variances remained fairly stable. Thus, it may be concluded that the model remains stable, even under attritions.

Once again it may be concluded that the pattern of gender differences is strikingly different in the analysis of latent dimensions compared to the pattern of differences in the observed variables. However, this time the observation is made with respect to the impact of missing data. When attrition and the Gender × Attrition interaction were taken into account, the pattern of gender differences changed in favor of girls on all tests except Copying, which remained the same. The pattern of gender differences on the latent variables changed differently compared to the observed variables. The female advantage on *g* increased compared to previous analyses, whereas their advantage on *Gc′* decreased. The male advantage on *Gv′*, *Sr′*, and *Cs′* increased, whereas their advantage on *NumAch′*, *EngAch′*, and *V′* decreased. Even though the difference between models in patterns of gender differences is small, it raises a number of questions.

It appears that the male attriters group is much weaker than the female group. Thus, the male performance is more overestimated than for the female performance in the subsample with complete data. This raises questions about what impact a gender-biased attrition may have had on other studies of gender differences in cognitive performance. It also raises questions about "the greater male variability hypothesis" (see Feingold, 1992a, 1992b; Noddings, 1992), which the previous analysis did not confirm, however. Greater male variability was only found in 3 of the 13 ability tests and in 3 of the 13 latent dimensions (*Gv′*, *Ms′*, and *CFR′*). One may ask if the attrition would influence the variance differences found as well. There were only a few tests in which individual attriters showed greater variability than the completers, two among boys and three (others) among girls. However, to answer this question in a multivariate latent fashion, a four-group model (boys vs. girls with complete or incomplete data, respectively) is required. Considering the modest deviation in results between the LQL and FQL models in this study and the rather time-consuming analysis technique, one may ask if it would be worthwhile.

It is interesting to note that the mean difference in favor of boys on spatial dimensions increases when the attrition is taken into account. Even though girls perform better or equal on many spatial tasks, they seem to use their general ability in solving such tasks. It raises the question about where these spatial capacities are developed, as they are not reflected in school performance. One hypothesis is that this is a cultural effect, in which the male context offers more opportunities to train these skills.

Another question is why the female advantage on *Gc′* decreases, as the differences in favor of girls in the attrition group are both larger and affect more variables than in the subsample with complete data. Furthermore, the means are almost the same in the female attriter group compared to the female group with complete data. The male attriter group has a lower mean than their male counterparts on most tests. It may be that the influence of *g* on those tests has increased when the attriters are included. The female increase in advantage of *g* is probably an effect of the attrition—that is, data missing for girls are rather similar to the girls with complete data, whereas the missing boys are very different from both their male counterparts and girls.

One of the central issues, the causes for the gender-biased attrition, also needs to be discussed. Every class in Sweden was supposed to participate in SA tests. As previously mentioned, the purpose is to calibrate the grades at the class level. However, in 1980, when the data in this study were collected, participation in the three SA tests in Grade 6 was not compulsory, which is the case now, even though most classes did take the tests. Because performance on SA tests may also be seen as a measure of the teacher's degree of success, low achievement may be one reason for not participating. However, this line of reasoning assumes that the teacher has the opportunity to compare results of his or her class with other classes.

Another possibility is that classes with very low achieving boys are missing. Such a finding would speak in favor of the hypothesis that boys are considered to be the norm for the class, whereas female performance is relatively invisible to the teacher (Wernersson, 1989; Öhrn, 1990, 1993). None of these hypotheses received strong support in this study. The pattern of gender differences and performance levels in the classes missing was quite similar to the completers' performance, although a tendency toward lower male performance was found on two of the tests.

It is also possible that the attrition of very weak boys is due to more or less open encouragement to certain students from the teacher and classmates—to stay home at the days when these tests are given. The tendency for this to apply to low-achieving boys and not to low-achieving girls may also be due to the boys' higher visibility compared to the low-achieving girls. The boys may be disturbing the order at the day for the test and thereby also affect the class results negatively, so that the teacher feels that the possibilities to give high grades to high achievers are unduly circumscribed. The hypothesis of male performance as a norm for the class level is applicable in this latter situation as well. The analysis showed that it was instead the missing individuals that caused the bias, and boys in particular. They performed lower on the ability tests than the male completers. Female attriters, on the other hand, seem fairly representative. These findings support the idea of differential selection mechanism for the genders and indirectly the male norm hypothesis.

Having taken gender and performance level into account and also having investigated the possibilities of mechanisms at the class level, one still may ask if the data are MAR data. In this particular case, the missingness on the SA tests appears mostly to be due to student absence because of sickness and other "legitimate" causes, which in turn are likely to be random.

An important question is, what are the methodological implications of this study? It is, of course, impossible to make any general conclusions about how one may expect FQL results to differ from LQL results. One may, however, conclude that FQL estimates are almost always less biased and more precise, which implies that this estimator should be used whenever possible.

It may be noted that the tedious multigroup method of Allison (1987) and Muthén et al. (1987), which was used in this study, is no longer required for this type of analysis. At this time, missing data analysis may be conducted with little effort with user-friendly software such as Amos (Arbuckle, 1996, 1997), Mx (Neale, 1995), and STREAMS (Gustafsson & Stahl, 1996). With Mx or Amos, the estimation procedure uses raw data directly. A reanalysis of the present data produced the same results as those just presented. STREAMS is an interface to modeling programs, which prepares the instructions for the multiple-group method for Amos, LISREL, and Mx, as well as the instructions for the raw data method in Amos and Mx.

The final conclusion from the findings in this study is that analysis of missing data is not only needed for more reliable analysis, but it can also contribute to the understanding of gender differences.

ACKNOWLEDGMENTS

The work reported here was financed by the Swedish Council for Research in the Humanities and Social Sciences. The analysis is a part of the research project, Models for Analysis of Group Differences in Cognitive Abilities, under the leadership of Professor Jan-Eric Gustafsson.

I would like to express my gratitude to Bengt Muthén and Ginger Nelson at UCLA for helpful advice during the analysis phase. I am also in great dept to Jan-Eric Gustafsson and Inga Wernersson for their help, encouragement, and invaluable comments on earlier drafts of the article.

REFERENCES

Allison, P. D. (1987). Estimation of linear models with incomplete data. In C. Clogg (Ed.), *Sociological methodology* (pp. 71–103). San Francisco: Jossey-Bass.

Allison, P. D., & Hauser, R. M. (1991). Reducing bias in estimates of linear models by remeasurement of a random subsample. *Sociological Methods & Research, 19,* 466–492.

Arbuckle, J. L. (1996). Full information estimation in the presence of incomplete data. In G. A. Marcoulides & R. E. Schumacker (Eds.), *Advanced structural equation modeling: Issues and techniques* (pp. 243–277). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Arbuckle, J. L. (1997). *Amos users' guide. Version 3.6.* Chicago: SmallWaters.

Balke-Aurell, G., & Lindblad, T. (1980). *Lär de sig något? Ord och grammatikkunskaper i engelska från årskurs 5 i grundskolan till 3/3 av gymnasiet* [Are they learning anything? Knowledge in English vocabulary and grammar from Grade 5 in compulsory school to the third year in upper secondary school] (Rep. No. 99). Göteborg, Sweden: Göteborg University, Department of Education.

Becker, B. J., & Hedges, L. V. (1984). Meta-analysis of cognitive gender differences: A comment on an analysis by Rosenthal and Rubin. *Journal of Educational Psychology, 76,* 583–587.

Byrne, B., & Shavelson, R. (1989). Adolescent self-concept: Testing the assumption of equivalent structure across gender. *American Educational Research Journal, 24,* 365–385.

Chipman, S. F., Brush, L. R., & Wilson, D. M. (Eds.). (1985). *Women and mathematics: Balancing the equation.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cleary, T. A. (1992). Gender differences in aptitude and achievement test scores. In J. Pfleiderer (Ed.), *Sex equity in educational opportunity, achievement and testing. Proceedings of the 1991 ETS Invitational Conference* (pp. 51–90). Princeton, NJ: Educational Testing Service.

Einarsson, J., & Hultman, T. (1984). *Godmorgon pojkar och flickor. Om språk och kön i skolan* [Good morning boys and girls. Focusing communication and gender in school]. Stockholm: Liber förlag.

Feingold, A. (1992a). The greater male variability controversy: Science versus politics. *Review of Educational Research, 62,* 89–90.

Feingold, A. (1992b). Sex differences in variability in intellectual abilities: A new look at an old controversy. *Review of Educational Research, 62,* 61–84.

Feingold, A. (1993). Joint effects of gender differences in central tendency and gender differences in variability. *Review of Educational Research, 63,* 106–109.

Fennema, E. (1980). Sex differences in mathematics achievement: Where and why. In L. H. Fox & D. Tobin (Eds.), *Women and the mathematical mystique* (pp. 76–93). Baltimore: The John Hopkins University Press.

Fennema, E., & Leder, G. C. (1990). *Mathematics and gender.* New York: Teachers College Press.

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

Gustafsson, J-E., Lindström, B., & Björck-Åkesson, E. (1981). *A general model for the organization of cognitive abilities* (Rep. No. 1981:6). Göteborg, Sweden: Göteborg University.

Gustafsson, J-E., & Stahl, P-A. (1996). *STREAM'S users guide. Version 1.7 for Windows.* Mölndal, Sweden: Multivariate Ware.

Hedges, L. V., & Friedman, L. (1993a). Computing gender difference effects in tails of distributions: The consequences of differences in tail size, effect size, and variance ratio. *Review of Educational Research, 63,* 110–112

Hedges, L. V., & Friedman, L. (1993b). Gender differences in variability in intellectual abilities: A reanalysis of Feingold's results. *Review of Educational Research, 63,* 94–105

Hirdman, Y. (1987). Makt och kön [Power and gender]. In O. Peterson (Ed.), *Maktbegreppet* (pp. 188–206). Stockholm: Carlssons förlag.

Hyde, J, S. (1981). How large are cognitive gender differences? A meta-analysis using w2 and d. *American Psychologist, 36,* 892–901

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

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modelling with the SIMPLIS command language.* Chicago: Scientific Software International.

Little, R. J. A., & Rubin, D. R. (1987). *Statistical analysis with missing data.* New York: Wiley.

Madow, W. G., Nisselson, H., & Olkin, I. (1983). *Incomplete data in sample surveys. Volume 1. Report and case studies.* New York: Academic.

Madow, W. G., & Olkin, I. (1983). *Incomplete data in sample surveys. Volume 3. Proceedings of the symposium.* New York: Academic.

Madow, W. G., Olkin, I., & Rubin, D. B. (1983). *Incomplete data in sample surveys. Volume 2. Theory and bibliographies.* New York. Academic.

Mäkitalo, Å., & Reuterberg, S-E. (1996). *Who takes the Swedish Scholastic Aptitude Test? A study of differential selection to the SweSAT in relation to gender and ability* (Rep. No. 1996:03). Göteborg, Sweden: Göteborg University, Department of Education and Educational Research.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modelling with data that are not missing completely at random. *Psychometrika, 52,* 431–462.

Muthén, B., Khoo, S-T., & Nelson Goff, G. (1994, April). *Multidimensional description of subgroup differences in mathematics achievement data from the 1992 national assessment of educational progress.* Paper presented at the meeting of the American Educational Research Association, New Orleans.

Neale, M. C. (1995). *Mx. Statistical modeling* (3rd ed.). Richmond, VA: Department of Psychology, Medical College of Virginia.

Noddings, N. (1992). Variability: A pernicious hypothesis. *Review of Educational Research, 62,* 85–88.

Reuterberg, S-E., Westerlund, A., & Gustafsson, J-E. (1992). Beror gruppskillnaderna på tidigare selektion i utbildningssystemet [Are group-differences affected by earlier selection in the education system]? *Betyg och högskoleprov för män och kvinnor. Rapport från ett seminarium. UHä –RAPPORT1992:3* [Grades and SAT for males and females. Report from a seminar; Rep. No. 1992:3]. Stockholm: National Swedish Board for Universities and Colleges.

Rosén, M. (1995). Gender differences in structure, means and variances of hierarchically ordered ability dimensions. *Learning and Instruction, 5,* 37–62.

Rovine, M. J. (1994). Latent variable models and missing data analysis. In A. von Eye & C. Clogg (Eds.), *Latent variable analysis. Applications for developmental research* (pp. 181–255). Thousand Oaks, CA: Sage.

Rubin, R. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Wernersson, I. (1977). *Köndsdifferentieringen i grundskolan* [Gender differentiation in compulsory school; Göteborg studies in educational sciences 22]. Göteborg, Sweden: Acta Universitatis Gothenburghensis.

Wernersson, I. (1989). Olika kön samma skola? En kunskapsöversikt om hur elevernas könstillhörighet påverkar deras skolsituation. [Different sex same school? A research review of how student gender affects the school situation] *Skolöverstyrelsen, Vad säger forskningen F 89:1* [National Board of Education. What does the Research tell us?; Rep. No. F 89:1].

Wernersson, I. (1991). Könsskillnader i gymnasieskolan. En kunskapsöversikt. [Gender differences in upper secondary school. A research review] *Skolöverstyrelsen, Vad säger forskningen F 91:2* [National Board of Education. What does the Research tell us?; Rep. No. F 91:2].

Wilkinson, L. C., & Marrett, C. B. (1985). *Gender influences in classroom interaction.* New York: Academic.

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

öhrn, E. (1990). *Könsmönster i klassrumsinteraktion. En observations- och intervjustudie av högstadieelvers lärarkontakter* [Gender pattern in classroom interaction, Göteborg studies in educational sciences 77]. Göteborg, Sweden: Acta Universitatis Gothoburghensis.

öhrn, E. (1993). Gender, influence and resistance in school. *British Journal of Sociology of Education, 14,* 147–158.

Part IV

GENDER DIFFERENCES IN READING PERFORMANCE ON DOCUMENTS ACROSS COUNTRIES

Monica Rosén
Department of Education and Didactics
Göteborg University, Sweden

Paper accepted for publication in
Reading & Writing: An Interdisciplinary Journal
Kluwer Academic Press

ABSTRACT

It's a known fact that excel over males in most reading tasks, but so not consistently in tasks that require processing information from maps, tables, charts and diagram, so called "Documents". The IEA Reading Literacy data provides possibilities to investigate gender differences across countries, in such tasks, in two age groups, 9-year olds and 14-year-olds. The general question about cultural influences vs. an invariant pattern of gender differences is of tremendous interest for gender research, and central in this study. The aim of the paper is to describe and analyse gender differences on Document tasks, and investigate if and how the pattern of differences varies across countries. Another aim is to demonstrate the power of using a multivariate analysis technique by contrasting it against traditional univariate approaches. The univariate analysis indicates female advantage as the most common in the younger group, and a mixed pattern in the older ones. The multivariate analysis indicates that Document tasks are not unidimensional, because both general and specific dimensions can be extracted from the raw scores. The traditional univariate analysis often disguised true patterns of differences in the data, both in terms of country differences and in terms of the direction of the gender difference. Raw score differences between the genders proved to be due to differences in both the general and passage specific dimensions. Ten of the countries showed gender differences in both directions in the general dimension among 9-year-olds, while an almost consistent pattern of female advantage was found among 14-year-olds. Many of the specific passage dimensions turned out to favor either males or females. This complex pattern varied over both age groups and across countries, although communalities in the pattern among subgroups of countries were common.

ACKNOWLEDGEMENT

The research reported in this paper has been financially supported by the Swedish Council for Research in the Humanities and Social Sciences (HSFR). The analysis is a part of the research project MALI (Multivariate Analyses of Literacy), under the leadership of Professor Ingvar Lundberg. I am very grateful to him and to Professor Jan-Eric Gustafsson for their support, advice and valuable comments with respect to my draft. I also wish to express my gratitude to Ingrid Munck for always being supportive and willing to share her knowledge regarding the IEA Reading Literacy data.

INTRODUCTION AND PURPOSE

In the 1990-1991 IEA Reading Literacy study (Elley, 1994), reading proficiency was defined by three text types: "Expository" and "Narrative" prose, and so called "Documents", a text type which requires the students to process information organised in matrix formats, such as maps, tables, charts, graphs, schedules, forms and diagrams. A consistent female advantage was found on Expository and Narrative item types, even though it varied across countries, whereas gender differences in performance on Document tasks tended to be either smaller or shift direction (Elley 1992, Wagemaker, 1996). The Document type was not included in the previous IEA Reading Literacy study (Thorndike, 1973). One implication of this new definition of Reading Literacy, is that it may change the pattern of gender differences from the typical pattern of female advantage in reading performance, which in turn may have implications on all levels in society. It is thus of great interest and importance to further investigate the phenomenon from a gender perspective.

The aim of the paper is to deepen the understanding of the gender differences in Document reading across countries reported in Elley (1994) and Wagemaker (1996). The approach adopted relies on a psychometric theory of cognitive abilities according to which, differences in performance on any cognitive task are caused by differences in *several* underlying abilities and contextual dimensions, which in turn have *various degrees of generality* (e.g. Gustafsson & Undheim, 1996). The theory is supported by a vast amount of empirical evidence (e.g. Carroll, 1993). Studies of gender differences are usually performed by analysing the mean difference on one *manifest* variable at a time, which is also the case in Elley's and Wagemaker's studies. In the present study this methodology is contrasted with a multivariate *latent* variable approach, which takes into account the fact that there are several factors influencing performance. The present study aims to decompose the variance on the Document tasks into multiple latent sources before investigating gender differences. The purpose is thus to investigate if, and how, the pattern of gender differences on the dimensions underlying performance on the Document tasks in the IEA reading literacy study, varies across countries. Another aim is to demonstrate the power of using a multivariate analysis technique.

In the following section I elaborate somewhat more on both the rationale for investigating gender differences in performance on document tasks, and on the the chosen perspective and research approach.

The Importance of the Problem

As Wagemaker (1996) points out, gender differences in educational outcome are of great concern for many societies. Therefore many questions need to be raised when females and males appear unequally able to derive benefit from the investment in education. Particular interest has been paid to gender differences in reading and related areas, since reading plays a significant role in achieving educational and vocational outcomes and in promoting the individual's ability to function in society. Document reading is a relatively new category in reading literacy. Consequently will probably gain considerable attention, because it is believed to represent non-school settings, such as society and workplace literacy, which according to Wagemaker "*may have a slightly different emphasis from those tasks commonly encountered in school*" (p.16). The importance of Document reading proficiency is even more emphasised in Mosenthal's (1996) review of literature. According to which positive relations have been reported between performance on Document reading tasks and the ability to function in society, e.g. to follow specific procedures, record significant events and make informed decisions. There are also signs of relations between Document literacy activity, occupational status and participation of adults in society. Last but not least, the gender differences in the IEA Reading Litercy study (Elley, 1994; Wagemaker, 1996) show a lesser female advantage for Documents compared to that of Narrative and Expository prose.

The redefinition of Reading Literacy in the IEA study, and Mosenthal's presentation of Document Literacy as a problem of importance, makes it quite obvious that Document reading proficiency is and/or will be perceived as an important cognitive skill, which makes it important to study it from a gender perspective. Measures of cognitive skills and large-scale studies generally tend to have profound effects on society and on our general belief system, regarding differences between males and females. Many cognitive measures have, furthermore, been shown to be either unfair to certain groups or badly analysed and interpreted (e.g. Gould, 1981; Tavris, 1992).

However, before Document reading proficiency is regarded as a specific cognitive skill other possibilities should be investigated. One reason why females do not have so much of an advantage on Documents as they do on Narrative and Expository, in the IEA Reading Literacy study, may well be due to the fact that Document tasks beside written words, often involve numerical and/or spatial content on which certain types of content, males excel at (e.g. Halpern, 1992). Another possibility is that the performance is affected by the actual topic/subject in each Document task. The opportunity to investigate the effects of these factors, across countries, may enrich our understanding of both our own and other cultures, and it may well reveal patterns that are now oblivious to us. For example, Taube & Munck (1996) argue that explanations of gender differences fall into two categories, either physiological or sociocultural. Support for the former is given if a consistent pattern of differences is found across countries, and support for the latter is given if the pattern of gender differences varies. They thought they had found some support for the former. There are thus several reasons to investigate the sources that may underlie the observed performance across countries before making any strong conclusions about gender differences in Document reading proficiency .

From a feminist point of view of gender equality it is important to observe, that a new definition of Reading Literacy, has been introduced into this global large scale manner, and to investigate what are the consequences. The fact that girls' advantage with respect to reading tests does not seem to benefit completely in equality terms later on in life, may be one indicator that reading proficiency has been too narrowly defined. On the other hand, a more conspiratory interpretation of the shift in definition is that it may also be an attempt to even out this well known imbalance in reading proficiency, so that the inconsistency mentioned can (continue to) be ignored. In either case, gender differences in document reading need to be carefully investigated.

The present study intends to take the above into consideration, and address the problem from a multidimensional psychometric perspective of individual and group differences in performance. This perspective is different from that of Elley (1992) and Wagemaker (1996) who's analyses assume that performance on Document tasks reflects one single underlying dimension. It is also quite different from the perspective chosen by Mosenthal (1996), who focuses item difficulty in Document literacy tasks. Mosenthal uses a cognitive-components approach which aims to develop processing models of existing psychometric measures. From one theoretical model of document processing, four types of strategies ordered along a hierarchy of difficulty were recognised, which when combined with four predictor variables (document complexity, type of information, type of match and plausibility of distractors) accounted for most of the variance in task difficulty. However, as was pointed out by Mosenthal "Although this approach adds precision to one's understanding of what constitutes task difficulty, it leaves unanswered the question of what precisely characterises the nature of individual strategies relative to different levels of task difficulty" (p. 330). Thus, Mosenthals approach may give valuable insights for test and task constructors, but as Lohman (1994) made explicit, attempts to isolate individual differences in performance on homogeneous tasks from component scores cannot succeed. For such goals, studies of individual differences in approaches to tasks are needed, and this is what is offered in the present study.

READING IN RELATION TO COGNITIVE SKILLS

Reading belongs essentially to the verbal domain, even though in practice reading is a key skill in almost any cognitive task. In the tradition of research on cognitive abilities, reading proficiency has not been interpreted as one single abilitiy, but rather as a reflection of several underlying cognitive abilities depending on the nature of the reading material used for the study.

Skills underlying performance on various types of texts may be connected to the domain of individual differences in cognitive abilities. Today, the leading theoretical model of cognitive abilities relies on an empirically well established hierarchical model with three levels (see Carroll, 1995; Gustafsson, 1988, 1994). At the apex there is general intelligence, influencing all cognitive performances. On the intermediate level several broad dimensions are defined, such as crystallised intelligence, which is, mainly a generalised verbal ability or general visualisation, which is a broad spatial ability. On the next level a large number of narrow ability dimensions have been identified, influencing rather specific cognitive tasks.

Reading has an interactive and constructive nature (Lundberg, 1991). Furthermore reading is a product of cultural evolution and relying on cultural transmission for its continued existence (Lundberg, 1995; Lundberg & Høien, 1991). It is thus reasonable to assume that reading is influenced, to a certain extent by general intelligence, to a large extent by crystallised intelligence, and by several more narrow dimensions depending on the type of reading task. Reading may also be given a broader definition than just word recognition and comprehension (Lundberg & Rosén, 1995), for example when the subject also has to decode and comprehend material with figural or pictorial content. With a broader definition it is thus reasonable to assume that reading skills reflect certain verbal abilities as well as other narrow dimensions along with more general (or broader) ability dimensions.

Lundberg & Rosén (1995) describe the three subdomains in the test battery of the IEA study as follows;

Narrative prose refers to continuos text materials in which the writer's aim is to tell a story, whether fact or fiction. They are normally designed to entertain or involve the reader emotionally; they are written in the past tense, and usually have people or animals as their main theme.

Expository prose refers to continuous text materials designed to describe or explain something. The subject of such texts are usually objects, the style is typically impersonal, highlighting such features as definitions, causes, classifications, functions, contrasts and examples, rather than a moving plot with a climax.

Documents refer to structured, tabular texts, such as forms, charts, labels, graphs, lists, and sets of instructions where the reading requirements typically involve locating information or following directions, rather than continuous reading of connected texts.

Balke (1995) adopted a multivariate approach to the data on the item level of the IEA data, and found it impossible to fit a model with separate factors reflecting these three domains. The Narrative and Expository domains were found to refer to the same underlying dimension. Documents, on the other hand, proved, to some extent to require different skills than did Narrative and Expository.

Reading Documents may be a rather complex task as compared to conventional reading of texts, because it requires several skills in addition to printed word recognition. Documents may require the reader to handle spatial layouts in order to locate specific information. The visual search in Documents may be more demanding on attentional perceptual skills. Some Documents call for a familiarity with certain spatial conventions, such as table layouts. Whilst further more Documents require the reader to follow directions. Other

Documents need careful, detailed and exact parsing of syntactic structures. Some may require the reader to relate different pieces of information to each other in order to integrate and compare. Documents also often involve the requirement of processing numbers. For some students, numbers have a strong negative emotional loading, which may trigger avoidance rather than active attention as required by the task. This may be more so for females than for males, and it may affect their performance.

Mosenthal (1996) attempts to examine the relative difficulty in Document tasks by investigating how, what he terms as "document reading-to-do" strategies, combined with different structure and processing conditions. A total of 217 document tasks, from within the USA, involving 5 national surveys of young adults were analysed. The study identified four different strategies (locate, cycle, integrate and generate) which were assumed to be hierarchically ordered with respect to difficulty. Combined with predictors of conditions (document complexity, type of information, type of match & plausibility of distractors), strategy type accounted for 80% of the variance in task difficulty. As mentioned earlier, Mosenthal's approach focuses task difficulty, which is quite different from capturing, describing and analysing differences in performance, as this study aims to show.

In the present study it is assumed that there is some communality among Document tasks which is influenced by both a "general reading skill" and "general document reading proficiency". Document tasks included in the same text passage, may also have an additional influence from "passage or test specific" components. This component is interpreted as interdependency between the items within the passage, which in turn may be due to the students' experience and/or practice of similar tasks. Gender differences may occur in both the general document reading dimension and in each passage specific component. These assumptions are tested in the present study using a hierarchical approach in structural equation modelling.

Gender Differences in Cognitive Skills

Gender differences are found with some regularity in at least three domains: in verbal ability favoring females (e.g., Anastasi, 1958; Tyler, 1947; Maccoby, 1966; Maccoby & Jacklin, 1974; Halpern, 1992; Wernersson, 1989; Emanuelsson & Svensson, 1990), in visual-spatial ability favoring males (Anastasi, 1958; Tyler, 1947; Maccoby & Jacklin, 1974; Hyde, 1981; Linn and Petersen, 1985; Halpern, 1992) and in mathematical or quantitative ability also favoring males (e.g., Maccoby & Jacklin, 1974; Halpern, 1992; Hyde, Fennema & Lamon, 1990). Meta-analyses have been conducted in all three domains and Hyde (1981) concludes that gender differences in cognitive abilities are generally not exessive.

A meta-analysis of 165 studies (Hyde & Linn, 1988) which reported gender differences in verbal ability estimated a weighted mean effect size so small (to the female advantage) that the authors concluded that gender differences in verbal ability no longer exist.

However, contradictory findings are still being reported. For example in a large-scale study of several cohorts of 13-year-olds in Sweden, there were, in 1961, only small differences between the genders in the verbal domain but in 1985 girls were clearly ahead (Emanuelsson & Svensson, 1990).

Gender Differences in Reading Performance

Thompson (1975) reviewed research on gender differences in reading attainment of English-speaking children, and her conclusion supports the idea of gender as being of no importance. Thompson stated that boys are slower at learning to read, but by the age of 10 the female advantage has disappeared.

Contradictory results were reported from an analysis of the 1982 NAEP study, where gender differences in reading achievement were found favoring females. Gates (1961) found a female advantage on three measures of reading (speed, reading vocabulary and level of comprehension) in a large study of more than 13000 students in grades 2 through to 8 (age 8 through 14).

Gender differences in the latest IEA Reading Literacy study have been investigated in several studies (Elley, 1994; Taube & Munck, 1996; Wagemaker; 1996). Wagemaker found a consistent pattern of females performing better than males at both age levels (9- and 14-year-olds), but with the differences tending to diminish in adolescence. This pattern held true for both high-achieving and low-achieving countries, for countries high and low on the Composite Development Index (CDI), and for countries with highly centralised or decentralised educational structures. However, on Document passages males were found to be favoured in both populations, with the differences tending to favour boys even more through time. The male advantage was, however, found to diminish in countries with a high CDI and in decentralised educational systems.

Taube and Munck (1996) reported gender differences as a function of themes for text tasks. A rather consistent pattern was found, with both 8-year-old and 14-year-old females achieving better on most tests, but particularly so on narrative texts, where the theme was about human beings, animals acting as human beings and human activities. Gender differences related to other test [illegible] components were also investigated, such as, main character in the text, [illegible]ive load, text type, length of the passage in terms of words, length of the passage in terms of lines, number of items and number of options. The study supported the hypothesis that the topic addressed in the reading task may be of importance for achievement as well as for the understanding of gender differences. However, these findings should not be viewed as final explanations of gender differences in reading skills, since the approach of Taube and Munck is similar to Mosenthals (1996), and they thus share the same problems. However, their findings may serve as hypotheses of test specific influences on reading performance.

Cultural, social and biological hypotheses are put forward for explaining patterns of gender differences within the cognitive domain. Perhaps tests of reading, to a large degree, indicate habits of reading, although the relation between reading habits and reading proficiency is ambiguous. Reading in terms of voluntary reading may be considered as a typical female habit in many countries, at least during adolescence. Guthrie and Greany (1991) reported that most surveys show that females enjoy reading more than boys do, and that they read more often. Elley (1994) reported gender differences in voluntary reading from the IEA study, supporting this hypothesis for 8-year-olds, but an opposite pattern was found among 14-year-olds. However, 14-year-old females reported higher frequencies of voluntary Document reading, while the opposite was found for the younger sample.

Multivariate Analysis of Gender Differences

Univariate analyses of observed variables, as mentioned previously, do not take into account the fact that the observed performance on any cognitive task reflects a mixture of several ability dimension of varying degrees of generality, and neither is measurement error taken into account (Gustafsson, 1989; Gustafsson & Balke, 1993; Carroll, 1993). For the study of group differences the relevance of using factor analysis is strongly argued for by Gustafsson (1992). The observed variance needs to be decomposed into latent variables, preferably hierarchically ordered and orthogonal, in order to make the interpretation as distinct and comprehensive as possible. Group differences in structure, means and variances may then be investigated. Rosén (1995) used such an approach in a study of gender differences in cognitive abilities in a Swedish sample of 981 13-year-olds. It was demonstrated that the pattern of gender differences was strikingly different when univariate analysis was compared to multivariate one, in which the variance had first been decomposed into a hierarchical model with orthogonal nested factors. Whereas no differences were found in the structure of cognitive abilities, quite substantial differences in variance on a few spatial dimensions were found. Considerable differences were discovered in means on latent dimensions; on *g* (general intelligence) favoring females, on *Gv'* (General visualisation) favouring males, on *Gc'* (Crystallized intelligence) favouring females, and on several narrow ability dimensions all favouring males. The univariate analysis seldom showed any significant gender difference on the 13 ability tests, and mostly a female advantage on standardised achievement tests (particularly on tests in Swedish and English). Thus, Rosén's analysis did not support the conclusion that gender is of no importance for the study of differences in cognitive abilities as proposed by Hyde & Linn (1988), Thompson (1975) and Jacklin (1989).

From this brief review two general questions regarding gender differences becomes vital for the present study: (1) Whether there exists any gender differences at all in document reading "proficiency" and/or if there are (other) gender differences in the underlying dimensions that may explain the observed results. (2) If and how the pattern of differences varies across countries.

Traditional univariate analysis may not be sensitive enough to capture gender differences in underlying dimensions of the tests. Multivariate approaches are more sensitive and capable of capturing complexities, but they are also rather more complicated to perform. The methodological issues connected to the two above questions will thus also be investigated.

METHOD

Sample

The data used in this analysis comes from the IEA Reading Literacy Study (Elley, 1994). The data was collected during 1990/1991 by the International Association for the Evaluation of Educational Achievement (IEA). The original sample is representative and comparable for two populations of age groups from approximately 30 countries: 9-year-olds (also referred to as population A), and 14-year-olds (also referred to as population B).

The samples analysed in the present study were somewhat reduced. The original samples were based on a selection of schools, where one class was selected. However, some countries chose to include all classes in the school. These extra classes are excluded in this study. Furthermore, cases without gender identification were also excluded. Finally, some countries were also excluded due to potential gender problems with their samples or with the data (for example high drop-out rate, low response rate, imbalance of gender, special sampling procedures). Excluded from population A were Indonesia and Cyprus, and excluded from population B were Botswana, Cyprus, France, Nigeria, the Philippines, Thailand, Venezuela, Trinidad & Tobago, and Zimbabwe.

The selected sample for the present study includes 25 countries in population A and 22 countries in population B. It should be noted that the Belgium sample only covers the French-speaking part, that Canada refers to Canada BC, and that the samples of Singapore and Iceland comprise one class from every existing school. Further information regarding the countries and their samples is available in Elley (1994). Descriptive statistics regarding the selected data for population A and B is presented in Table 1.

Table 1 **Number of cases, percentage of males and females per country in Pop A (9-year-olds) and in Pop B (14-year-olds).**

Country	POP A n f cases	POP A % males	POP A % females	POP B n f cases	POP B % males	POP B % females
Belgium F	2623	49,6	50,4	2679	50,5	49,5
Canada BC	2550	51,5	48,5	4019	51,3	48,7
Denmark	2636	50,3	49,7	2512	50,5	49,5
Finland	1502	52,6	47,4	1186	51,7	48,3
France	1843	49,1	50,9	-	-	-
Germany East	1757	50,0	50,0	1818	49,0	51,0
Germany West	2727	52,0	48,0	4052	51,5	48,5
Greece	3496	50,3	49,7	3673	49,7	50,3
Hong Kong	3300	53,7	46,3	3152	49,3	50,7
Hungary	2948	50.6	49,4	3252	49,6	50,4
Iceland	1993	52,0	48,0	1856	51,1	48,9
Ireland	2669	50,1	49,9	3557	52,3	47,7
Italy	2147	52,1	47,9	3056	50,8	49,2
The Netherlands	1590	48,7	51,3	3513	52,5	47,5
New Zealand	2973	51,8	48,2	3023	49,5	50,5
Norway	2416	50,0	50,0	2228	49,9	50,1
Portugal	2574	51,5	48,5	3103	45,7	54,3
Singapore	7326	51,6	48,4	4805	49,7	50,3
Spain	8205	50,3	49,7	8460	48,2	51,8
Sweden	2253	51,5	48,5	3478	51,3	48,7
Switzerland	3264	51,7	48,3	5175	50,6	40,4
Trinidad & Tobago	2683	48,9	51,1	-	-	-
USA	3568	50,1	49,9	3330	49,1	50,9
Venezuela	4286	48,4	51,6	-	-	-
Slovenia	3290	51,3	48,7	3228	47,9	52,1

Sample sizes vary between 1500 and 8200 cases in Pop A, and between 1200 and 8500 cases in Pop B. Most countries have sample sizes between 2500-3500 cases. The proportion of males and females are about equal in every country.

Variables

All sub tests (passages) in the IEA Reading literacy study were divided into two booklets for both populations (Booklet 1 & Booklet 2), and administered to the students on two different occasions. The variables selected for the present study were the sub scores for all the Document passages in both booklets for both populations.

Six Document passages, two in Booklet 1 and four in Booklet 2 were presented to the 9-year-olds. Each passage was followed by three to six questions. The response format was multiple-choice for all passages except one. The "Buses" sub test required the students to write a short answer consisting of a number or a name.

The 14-year-olds were presented with nine Document passages, four in Booklet 1 and five in Booklet 2. Each passage was followed by three to seven questions with the same response formats as those in the Pop A test. An additional response format was represented in the sub test "Directions" which required the student to draw figures in a rectangle according to given directions. A brief description of the passages for Pop A and B is shown in Table 2. An extended description of the tests are given by Elley (1994)

Table 2 **Description of Document passages in Pop A and Pop B**

Documnent passages in Pop A (9-year-olds)			
Document-passage	**Abbrev**	**n of items**	**Description of the task**
1A. Island	**Island**	4	To locate points on a simple map
2A. Maria's time table	**Maria**	3	To locate lessons on a school timetable
3B. Empty bottles	**Bottles**	4	To interpret bar graphs showing number of bottles collected by four classes
4B. Buses	**Buses**	4	To locate places and times on a bus time table
5B. Table of contents	**Content**	3	To identify authors and page numbers in a table of contents
6B. Temperature	**Temp**	5	To locate times and trends on a temperature chart over four days
Documnent passages in Pop B (14-year-olds)			
Document-passage	**Abbrev**	**n of items**	**Description of the task**
1A. Fill in the travellers card	**Trav card**	7	To fill in a traveller's card for a person, using provided information
2A. Resources	**Resou**	3	To use a key to locate resources in a detailed map
3A. Job vacancies	**Job vac**	3	To use job vacancy advertisement to locate jobs with particular features
4A. The snow hare and the Canadian lynx	**Hare & Lynx**	3	To interpret bar graph showing population frequencies of hares and lynxes over 10 years
5B. Bus	**Bus**	3	To locate times, places and routes on a complex bus time table
6B. Directions	**Direc**	3	To insert geometric figures and numbers in a rectangle, as directed, and answer questions about completed figure
7B. Global weather	**Weather**	4	To locate places with particular weather features on a list of 38 world cities
8B. Temperature	**Temp**	5	To locate times and trends on a temperature chart over 4 days
9B. Soluble aspirol with vitamin C	**Aspirol**	3	To answer questions about use, dosage and composition of Aspirol tablets

The number indicates the order of given passages, booklet 1 is denoted "A", booklet 2 is denoted "B".

Gender was considered during the process of construction and selecting passages and items for both populations. The intention was to avoid gender biases in the tests of reading (Wagemaker, 1996).

Procedure and Methodology for the Data Analysis

The first step of the data analysis was a univariate analysis of gender differences in Document passages and also in the total Document scores for each country in each population. Three types of total Document scores were compared, two Rasch scores, and one raw score. The Rasch score is a transformation from the total raw score, where a scaled score is estimated for each student on the basis of the information gathered from responses from all other items. Correlations and/or t-values for mean comparisons between the genders were computed for each country.

In the second step, one measurement model for each population was fitted to separate covariance matrices of all the Document passages for the countries included. First, a general factor was included with loadings on all observed Document passage scores. If the data suggested any additional factors in order to obtain acceptable fit, the factors were then entered after the general factor had been included, in accordance with the hierarchical approach with nested factors suggested by Gustafsson & Balke (1992).

A separate one-factor model was then fitted to the covariance matrices for all countries in population A and in population B. However, for some countries in population A adjustments in terms of correlated errors were made, in order to improve the fit of the model. For population B, a two-factor model was fitted to the covariance matrices in all countries except East Germany, West Germany, Greece, Hong Kong, Italy, Portugal, Singapore, Spain and Slovenia where a one-factor model was sufficient to yield acceptable fit indices. (The models are visualised in Figure 1 & 2 in the result section).

Many measures of fit for latent variable models have been developed and suggested during the past decades. The most adequate measure for the moment appears to be RMSEA (Root Mean Square of Approximation) which should be no higher than .05 even though values up to .07 may be acceptable (Cudeck & Browne, 1993). Still, the Chi-2, Goodness of Fit Index (GFI) and Non Normed Fit Index (NNFI) (Jöreskog & Sörbom, 1993), are also presented as supplementary information. Both GFI and NNFI should be as close to 1.00 as possible, although values around .90 are considered to be quite accebtable.

In the final step, gender was entered into the models as a dummy variable, always with relations to the factors, significant or not, but only to those passage-specific components that indicated significant differences. Entering gender as a dummy variable in a one-group model has many practical and technical advantages as compared to doing a two-group analysis, and it may well give similar results (Rosén, 1995). However, the dummy-procedure imposes the assumption of equal structure and equal variance in the latent constructs and measurement errors, which may or may not be true.

RESULTS

It is important to keep separate manifest variables, such as document passage sub scores and document total scores on the one hand, and latent variables and test specific components on the other. Factors and test specific components are thus noted in italic style, as *Maria*, while observed "manifest" variables are referred to with normal text surrounded by quotation marks, as "Maria".

Gender Differences in Observed Scores on Documents

The univariate analysis of gender differences was based on correlations and t-tests for mean differences on each Document passage sub score and on the total score for all Document items, along with two Rasch scores. Rasch 1 is the score obtained when items not reached are scored as omitted, and Rasch 2 is the score obtained when items not reached are scored as incorrect. Table 3 presents correlations between gender and the two Rasch-scores and the total document raw score for both age groups.

Table 3 Correlations between gender and different types of total Document scores in Pop A (9-year-olds) and Pop B (14-year-olds). Positive values* in light shaded areas indicate female advantage, negative values* in dark shaded areas indicate male advantage at the 5%-level.

	Pop A			Pop B		
Country	**Tot doc**	**Rasch 1**	**Rasch 2**	**Tot doc**	**Rasch 1**	**Rasch 2**
Belgium F	,008	,005	,005	-,011	-,023	-,016
Canada BC	,003	-,013	-,005	,095*	,075*	,075*
Denmark	,087*	-,009	,047*	,032	,001	,005
Finland	,033	,004	,013	,034	,028	,028
France	-,030	-,039	-,043			
Germany East	,031	,000	,017	,030	,004	,005
Germany West	,006	-,017	-,009	-,029	-,059*	-,058*
Greece	-,001	-,012	-,012	-,053*	-,078*	-,077*
Hong Kong	,001	-,034	-,012	-,037*	-,063*	-,063*
Hungary	,003	-,011	-,003	,061*	,039*	,040*
Iceland	,081*	,028	,061*	,089*	,044	,050*
Ireland	,030	,020	,030	,057*	,016	,018
Italy	,021	,009	,009	-,009	-,039*	-,039*
The Netherlands	-,011	-,034	-,015	,056*	,034*	,034*
New Zealand	,075*	,064*	,067*	-,048*	-,077*	-,074*
Norway	,053*	,012	,043*	-,008	-,019	-,024
Portugal	,005	-,014	-,012	-,121*	-,158*	-,158*
Singapore	,031*	,026*	,025*	-,007	-,041*	-,040*
Spain	-,010	-,028*	-,021	-,046*	-,059*	-,055*
Sweden	,031	-,005	,014	,090*	,057*	,055*
Switzerland	-,024	-,033	-,027	-,015	-,044*	-,043*
Trinidad & Tobago	,047*	,044*	,045*			
USA	,010	-,010	-,011	,022	,012	,013
Venezuela	,040*	,031*	,047*			
Slovenia	,054*	,033	,043*	-,008	-,041*	-,041*

Tot doc: Raw scores
Rasch 1: Rasch score when not reached items is scored as omitted
Rasch 2: Rasch score when not reached itmes is scored as wrong

The pattern of gender differences among 9-year-olds (Pop A) does not vary much between the various types of total scores. According to the Tot Doc (raw score) and Rasch 2 females perform significantly better in 9 out of 25 countries. Rasch 1 shows a significant male advantage in Spain, and a female advantage in only in 5 countries. Overall, when an item unreached is scored omitted, females appear to loose and males benefit somewhat. A rather different pattern is observed among 14-year-olds (Pop B). Here, the deviation between the three types of total Document scores is much larger than in Pop A. Out of 22 countries, Tot Doc shows a female advantage in 6 countries and a male advantage in 5 countries. Both Rasch scores seem to benefit males more, showing a significant male advantage in 10 countries, while a female advantage is present in 4 (Rasch 1) or 5 countries (Rasch 2).

In the next step of the analyses the total Doc score was divided into sub scores for each passage. For Pop A t-values and correlation between gender and the raw score for each Document passage is presented in Table 4.

Table 4 Correlations (C) and T-values (T) for gender comparison of observed means on Document passages subtests in Pop A (9-year-olds).

Country / Passage		Content 5B	Maria 2A	Buses 4B	Bottles 3B	Temp 6B	Island 1A
Belgium F	C	,058*	,054*	,040*	-,031	-,024	-,036
	T	2,99*	2,79*	2,02*	-1,60	-1,24	-1,84
Canada BC	C	,095*	,027	-,002	-,019	-,043*	-,014
	T	4,81*	1,35	-0,10	-0,95	-2,18*	-0,71
Denmark	C	,139*	,104*	,057*	,044*	,004	,077*
	T	7,21*	5,39*	2,91*	2,28*	0,21	3,95*
Finland	C	,075*	,082*	-,009	,005	,012	,012
	T	2,90*	3,18*	-0,35	0,21	0,48	0,45
France	C	,096*	,013	-,034	-,032	-,011	-,105*
	T	4,15*	0,57	-1,44	-1,35	-0,49	-4,54*
Germany East	C	,078*	,046	-,001	,018	,028	-,022
	T	3,28*	1,94	-0,04	0,73	1,16	-0,90
Germany West	C	,102*	,050*	-,005	-,004	-,026	-,058*
	T	5,33*	2,59*	-0,25	-0,21	-1,36	-3,01*
Greece	C	,062*	,012	-,007	-,011	-,025	-,007
	T	3,65*	0,72	-0,40	-0,64	-1,45	-0,41
Hong Kong	C	,058*	,036*	-,044*	,024	,038*	-,085*
	T	3,36*	2,09*	-2,54*	1,38	2,18*	-4,87*
Hungary	C	,097*	,063*	-,010	-,033	,027	-,082*
	T	5,26*	3,42*	-0,55	-1,78	1,45	-4,44*
Iceland	C	,144*	,114*	,084*	,070*	-,020	-,001
	T	6,48*	5,11*	3,77*	3,15*	-0,87	-0,04
Ireland	C	,097*	,069*	,051*	-,006	-,029	-,027
	T	5,04*	3,58*	2,65*	-0,30	-1,50	-1,40
Italy	C	,081*	,056*	-,014	-,010	,033	-,027
	T	3,76*	2,59*	-0,64	-0,46	1,54	-1,24
The Netherlands	C	,087*	,020	-,028	-,011	-,053*	-,028
	T	3,49*	0,81	-1,11	-0,45	-2,13*	-1,13
New Zealand	C	,131*	,108*	,053*	,047*	,020	,001
	T	7,22*	5,94*	2,90*	2,54*	1,11	0,05
Norway	C	,128*	,091*	,057*	,026	-,025	-,009
	T	6,35*	4,50*	2,80*	1,25	-1,21	-0,42
Portugal	C	,095*	,004*	,010	-,062*	-,002	-,052*
	T	4,85*	3,25*	0,64	-3,17*	-0,11	-2,66*
Singapore	C	,090*	,098*	,024*	,011	-,009	-,045*
	T	7,69*	8,45*	2,09*	0,90	-0,77	-3,86*
Spain	C	,093*	,047*	,016	-,029*	-,045*	-,085*
	T	8,42*	4,25*	1,49	-2,61*	-4,09*	-7,71*
Sweden	C	,081*	,085*	,025	-,019	-,002	,009
	T	3,84*	4,05*	0,97	-0,92	-0,08	0,43
Switzerland	C	,066*	,049*	-,021	-,031	-,049*	-,073*
	T	3,78*	2,82*	-1,19	-1,74	-2,82*	-4,19*
Trinidad & Tobago	C	,126*	,062*	,038*	,020	,020	-,038*
	T	6,57*	3,24*	2,00*	1,02	0,04	-1,96*
USA	C	,082*	,065*	,001	,007	-,031	-,034*
	T	4,93*	3,89*	0,05	0,40	-1,86	-2,04*
Venezuela	C	,090*	,084*	,019	,008	-,038*	-,002
	T	5,94*	5,51*	1,25	0,53	-2,46*	-0,15
Slovenia	C	,038*	,095*	,038*	,030	,098*	-,045*
	T	5,65*	5,50*	2,17*	1,70	2,13*	-2,59*

Positive values* in light shaded areas indicate female advantage, negative values* in dark shaded areas indicate male advantage at the 5%-level. The column number indicate the given order of passages, booklet 1 is denoted "A", booklet 2 is denoted "B".

Some passages appear to suit either gender almost invariantly across countries. In Table 4, the more female suited passages are placed to the left, the more gender neutral in the middle and the more male suited passages to the right. The "Table of contents" shows a significant female advantage in every country. "Maria's Timetable" also shows a similar pattern for almost every country. For "Island" and "Temperature" there is either no significant difference or a male advantage. The same is true for females on "Buses". Gender differences in both directions are found for "Bottles" and for "Temperature".

The largest differences are found in the "Content" passage, favoring females in Spain. The second largest difference is found in the "Island" passage, favoring males in Spain.

Countries which show very few gender differences are East Germany and Greece followed by the Netherlands. The most pronounced pattern of gender differences in general was observed in Denmark, Hong Kong, Spain and Slovenia. Female advantage was most pronounced in Denmark, Iceland and New Zealand, while male advantage was most pronounced in Spain, followed by Portugal, Switzerland and Hong Kong.

For Pop A, the pattern of gender differences thus appears to be both variant and invariant in terms of female-male direction over countries. Gender differences in performance on Documents appears in general to match previous findings of female advantage in the verbal domain. Nevertheless, two Document passages, Island and Temperature, seem to fit male students better than females in many countries.

Turning to Pop B, results of gender comparisons of the observed performance on the nine Document passages are presented in Table 5.

Table 5 Correlations (C) and T-values (T) for gender comparison of observed means on Document passages subtests in Pop B (14-year-olds).

Country		Trav card 1A	Direc 6B	Aspi-rol 9B	Job vac 3A	Bus 5B	Wea-ther 7B	Temp 8B	Resou 2A	Hare & Lynx 4A
Belgium F	C	-,008	,033	,049*	,003	,032	,047*	-,020	-,126*	-,079*
	T	-0,39	1,70	2,55*	0,14	1,66	2,43*	-1,04	-6,55*	-4,07*
Canada BC	C	,150*	,099*	,091*	,073*	,050*	,076*	,017	-,054*	-,066*
	T	9,61*	6,27*	5,78*	4,63*	3,20*	4,82*	1,11	-3,40*	-4,22*
Denmark	C	,127*	,053*	,057*	,042*	,063*	-,001	-,025	-,107*	-,104*
	T	6,44*	2,66*	2,88*	2,11*	3,17*	-0,07	-1,24	-5,39*	-5,25*
Finland	C	,040	,030	-,028	,092*	,051	,018	-,029	,017	-,023
	T	1,36	1,05	-0,97	3,16*	1,74	0,63	-0,99	0,57	-0,80
Germany East	C	,081*	,024	,014	,086*	,025	,048*	-,017	-,063*	-,051*
	T	3,47*	1,02	0,58	3,70*	1,06	2,04*	-0,71	-2,69*	-2,17*
Germany West	C	,068*	,045*	-,013	,021	-,008	,028	-,091*	-,105*	-,101*
	T	4,36*	2,87*	-0,83	1,35	-0,40	1,75	-5,81*	-6,69*	-6,44*
Greece	C	,041*	,057*	-,013	,028	,014	,007	-,128*	-,152*	-,104*
	T	2,48*	3,46*	-0,77	1,72	0,83	0,45	-7,79*	-9,34*	-6,34*
Hong Kong	C	,070*	,002	-,015	,052*	-,039*	-,019	-,022	-,135*	-,053*
	T	3,39*	0,12	-0,82	2,91*	-2,18*	-1,05	-1,26	-7,66*	-2,97*
Hungary	C	,149*	,019	,015	,025	,073*	,056*	-,029	,016	-,084*
	T	8,60*	1,09	0,85	1,44	4,20*	3,19*	-1,64	0,94	-4,78*
Iceland	C	,195*	,089*	,100*	,096*	,052*	,044	,025	-,077*	-,081*
	T	8,55*	3,85*	4,31*	4,14*	2,23*	1,91	1,06	-3,32*	-3,51*
Ireland	C	,111*	,067*	,080*	,025	,050*	,040*	,005	-,069*	-,030
	T	6,66*	4,01*	4,78*	1,46	3,00*	2,37*	0,32	-4,10*	-1,77
Italy	C	,073*	,045*	,020	,077*	-,010	-,026	-,059*	-,094*	-,077*
	T	4,03*	2,49*	1,11	4,29*	-0,55	-1,46	-3,24*	-5,19*	-2,24*
The Netherlands	C	,105*	,088*	,060*	,032	,044*	-,008	,033*	-,043*	-,067*
	T	6,23*	5,23*	3,54*	1,88	2,60*	-0,47	1,98*	-2,52*	-3,99*
New Zealand	C	-,004	,007	-,003	,028	,004	-,074*	-,067*	-,104*	-,061*
	T	-0,21	0,36	0,19	1,52	0,20	-4,05*	-3,67*	-5,76*	-3,37*
Norway	C	,072*	,014	,002	,026	,020	-,024	-,031	-,047*	-,095*
	T	3,40*	0,68	0,07	1,23	0,94	-1,12	-1,47	-2,22*	-4,49*
Portugal	C	,064*	-,028	-,079*	,014	-,169*	-,077*	-,131*	-,195*	-,180*
	T	3,56*	-1,56	-4,42*	0,79	-3,85*	-4,30*	-7,36*	-11,04*	-10,19*
Singapore	C	,133*	,030*	,006	,021	,005	,001	-,042*	-,126*	-,103*
	T	9,32*	2,05*	0,43	1,45	0,36	0,06	-2,94*	-8,77*	-7,18*
Spain	C	,068*	-,012	,056*	-,007	-,015	-,058*	-,084*	-,126*	-,119*
	T	6,22*	-1,12	5,12*	-0,67	-1,41	-5,32*	-7,72*	-12,43*	-11,02*
Sweden	C	,182*	,050*	,077*	,070*	,086*	,028	,009	-,008	-,060*
	T	10,90*	2,96*	4,54*	4,12*	5,11*	1,63	0,53	-0,48	-3,53*
Switzer-land	C	,075*	,011	,018	,005	,001	-,014	-,034*	-,082*	-,097*
	T	5,43*	0,82	1,26	0,35	0,04	-0,97	-2,43*	-5,94*	-6,99*
USA	C	,040*	,037*	,085*	,017	,077*	,007	,009	-,048*	-,104*
	T	2,31*	2,05*	4,90*	0,97	4,42*	0,41	0,51	-2,75*	-6,01*
Slovenia	C	,127*	,038*	-,024	,006	-,017	-,001	-,045*	-,048*	-,110*
	T	7,29*	2,14*	-1,36	0,35	-0,96	-0,05	-2,56*	-2,75*	-6,26*

Positive values* in light shaded areas indicate female advantage, negative values* in dark shaded areas indicate male advantage at the 5%-level. The column number indicate the given order of passages, booklet 1 is denoted "A", booklet 2 is denoted "B".

Again, certain document passages seem to appeal to the same gender, consistently, across countries. Females perform better in almost every country on "Traveller's card" and on "Directions". No bias or female advantage is the most common result for "Job Vacancies", "Bus" and "Aspirol". An almost invariant male advantage is present on " Resources" and "Hare & Lynx". A mixed pattern of gender differences over countries is found on "Global weather" and on the total Document score .

The largest differences favouring males are to bee seen in "Resources" and "Hare & Lynx" for Spain and Portugal. The largest female advantage is present on "Traveller's card" in Canada BC, Singapore and in Sweden.

Gender differences are most infrequent in Finland, followed by Norway. Most frequent are gender differences in Canada BC, followed by Denmark, Iceland, Netherlands and Portugal. Female advantage is most common in Canada BC, Denmark, Iceland, Ireland and Sweden. Male advantage is most common in Portugal, New Zealand and Spain.

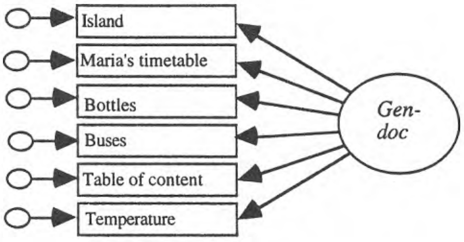
To sum up the univariate analysis, one may conclude that a female advantage on the total document score was more frequent in Pop A than in Pop B. This was true for the Document passages sub scores as well. In general gender differences on Document passages were more common among 14-years-olds than among 8-year-olds. Furthermore, the direction of the differences was more mixed in Pop B, but still more often to a female advantage. Both Pop A and Pop B thus show a pattern of gender differences that is both variant and invariant over countries. Some passages appear to suit one gender better than the other.

From the above analysis it is impossible to conclude whether the differences are derived from differences in Document reading proficiency in general, or from specific ingredients that may have to do with the content or actual topic in each passage. As will be demonstrated in the following, a multivariate approach is needed to clarify more, these questions.

Latent Variable Models for Documents

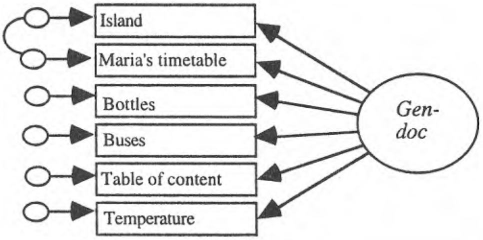
In the next step, measurement models were fitted to the data in order to uncover underlying latent dimensions in the total document raw score. All latent constructs are denoted in italics. As a starting point, separate one factor models for each country were fitted to the covariance matrices in both populations, using the LISREL 8 software (Jöreskog & Sörbom, 1993). They thus included one latent variable labeled *Gendoc*, with relations to all passages, and the remaining variation in each passage when *Gendoc* is partialled out is regarded as the sub-test or passage specific component, which also are latent constructs. Small adjustments and additional factors were then imposed on the models whenever the data so suggested. These adjustments are described below.

For Pop A, a one-factor model with 6 uncorrelated test specific components was fitted to the data (Model A). This model turned out to fit the data from three countries. The data for the rest of the countries suggested a correlation between some passage specific components. A very good model fit was obtained for most remaining countries when a correlation between *Maria* and *Island* (Model B) or a correlation between *Bottles* and *Content* (Model C) was allowed in the model. A visual presentation of the models is presented in Figure 1.



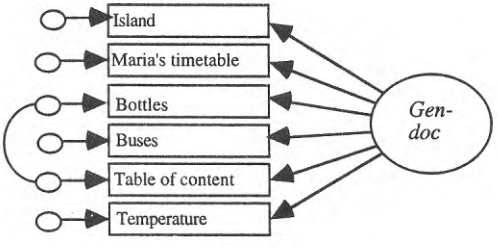
Model A

Model A is a one-factor model with *Gendoc* with no additional correlation between passage specific components.



Model B

Model B is a one factor model with *Gendoc* and a correlation between the passage specific components of "*Maria*" and "*Island*"



Model C

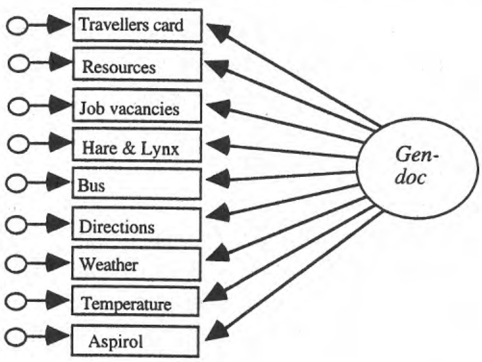
Model C is a one factor model with *Gendoc* and a correlation between the passage specific components of "Bottles" and "Table of content".

Figure 1 One factor models in Pop A.

Model A is applied to the covariance matrices of Hong Kong, the Netherlands and Trinidad & Tobago (Df = 9, χ^2 = 16 - 48, GFI = .99 - 1.00, NNFI = .97 - 1.00, RMSEA = 0.018 - 0.036). Model B is the most frequent model in Pop A, fitting the data of Belgium F, France, Greece, Denmark, East and West Germany, Hungary, Finland, Iceland, New Zealand, Norway, Portugal, Spain, Sweden, Switzerland Venezuela and Slovenia (Df = 8, χ^2 = 6 - 50, GFI = .99 - 1.00, NNFI = .98 - 1.00, RMSEA = 0.000 - 0.030). Model C suited Canada BC, Italy, Singapore and USA (Df = 8, χ^2 = 28 – 43, GFI = .99 - 1.00, NNFI = .97 - 99, RMSEA = 0.024 - 0.033)

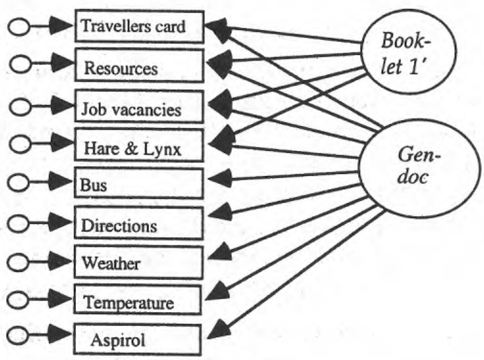
The fit indices for the models fitted to the data in each country indicates an excellent fit according to Cudeck & Brown (1993), RMSEA varying between .00 to about .04.

Figure 2 presents the models fitted to the covariance matrices in Pop B.



Model A

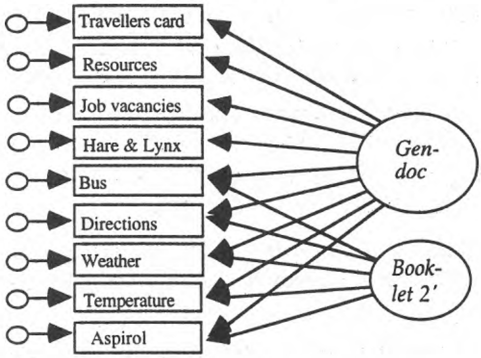
Model A is a one factor model with uncorrelated passage specific components.



Model B

Model B is a two-factor model with one general document factor *Gendoc*, a nested factor

Booklet1 of residual variance from all document passages in Booklet 1,and nine passage specific components.



Model C

Model C is a two-factor model with one general document factor *Gendoc*, a nested factor *Booklet2'* of residual variance from all document passages in Booklet 2, and nine passage specific components.

Figure 2 One factor models in Pop B.

Model A is the plain one factor model, with one *Gendoc* factor with relations to all passages, and nine passage specific components. This model fitted almost half of the countries very well, that is East and West Germany, Greece, Hong Kong, Italy, Portugal, Singapore, Spain and Slovenia (Df = 27, χ^2 = 68 - 154, GFI = .98 - 1.00, NNFI = .89 - .97, RMSEA = 0.018 - 0.043).

The data for the rest of the twelve countries in Pop B suggested an additional factor based on residual variance from all Document passages in either Booklet 1 (Model B) or Booklet 2 (Model C). These latent variables are labeled *Booklet1'* or *Booklet2'*. A very preliminary interpretation of the booklet-factors are that they may have something to do with the test administration.

Model B thus includes one *Gendoc*, and one residual factor *Booklet1'*, along with the nine test specific components. Model B shows satisfactory fit indices for Belgium F, Canada BC, Denmark, Finland, Hungary, Iceland, Sweden, Switzerland, and the US (Df = 23, χ^2 = 66 - 192, GFI = .98 - 1.00, NNFI = .91 - .97, RMSEA = 0.024 - 0.051). Model C includes one residual factor *Booklet2'*, along with *Gendoc*, and the nine test specific components, and is fitted to the data of Ireland, New Zealand, the Netherlands and Norway (Df = 22, χ^2 = 60 - 182, GFI = .99 - 1.00, NNFI = .93 - .97, RMSEA = 0.027 - 0.049).

Fit indices for all three types of models are quite sufficient in both Pop A and Pop B, and they continue to be so in the next step when gender is entered into the models.

Gender Differences in Latent Dimensions

In the final step gender differences were investigated in the models derived. The latent variable models offer the opportunity to investigate gender differences, both in the general dimension and in the passage specific components, that are left when general dimensions (*Gendoc* for Pop A, and *Gendoc* and sometimes *Booklet* for Pop B) are taken into account. Gender was inserted as a dummy variable (males coded 0 and females coded 1). The factors were then regressed upon the dummy variable and so were those passage specific components that revealed a significant relation.

The correct measure of gender differences in the latent means were not obtained until all significant relations were included in the model. The pattern of gender differences kept changing during the process of searching for significant relations, and did not stabilise until all significant relations were established. Just allowing for one relation between gender and *Gendoc* thus does not reflect the true difference between the genders.

Including gender differences in the model improved the model fit slightly for most countries in both populations. Gender may thus explain a significant part of the variance in the models. Allowing for a correlation between two passage specific components in Pop A did not affect the gender differences in any substantial way. Allowing an additional factor in Pop B, instead of allowing those passage specific components to correlate, turned out to affect estimates of gender differences on both *Gendoc* and on passage specific components.

The pattern of gender differences in the models of 9-year-olds in Pop A is presented in Table 6, and is summarised in Table 7 where the number of observed (manifest) differences can be compared with differences in latent dimensions

Table 6 Gender differences in latent constructs and passage specific components for 9-year-olds (Pop A), Standardised estimates (S) and T-values (T).

Passage / Country		*Content* 5B	*Maria* 2A	*Gen-doc*	*Buses* 4B	*Bottles* 3B	*Temp* 6B	*Island* 1A
Belgium F B)	S	0,08*	0,06*	-0,05*	0,06*			
	T	4,42*	3,35*	-1,97*	3,32*			
Cananda BC B)	S	0,09*		0,01			-0,05	
	T	5,20*		0,29			-2,69	
Denmark B)	S	0,05*		0,10*			-0,06	
	T	3,42*		4,53*			-3,83	
Finland B)	S	0,06*	0,07*	0,01				
	T	2,74*	3,14*	0,20				
France B)	S	0,11*		-0,03				-0,08*
	T	5,01*		-0,99				-3,48*
Germany East B)	S			0,05				
	T			1,84				
Germany West B)	S	0,12*	0,06*	-0,02				
	T	7,13*	3,29*	-1,01				
Greece B)	S	0,05*		0,01			-0,04*	
	T	3,28*		0,49			-2,72*	
Hong Kong A)	S			0,08*	-0,08*			-0,10*
	T			3,39*	-4,67*			-5,97*
Hungary B)	S	0,13*	0,08*	-0,04			0,05*	
	T	7,23*	4,68*	-1,70			2,89*	
Iceland B)	S	0,04*		0,14*			-0,10*	-0,06*
	T	2,14*		4,98*			-4,66*	-2,76*
Ireland B)	S	0,80*	0,08*	-0,05	0,08*			
	T	6,34*	4,49*	-1,79	4,02*			
Italy C)	S	0,09*	0,07*	-0,04			0,07*	
	T	4,58*	3,32*	-1,38			4,55*	
The Nether-lands A)	S	0,11*		-0,03				
	T	4,79*		-1,05				
New Zealand B)	S	0,09*	0,08*	0,06*				
	T	5,50*	4,78*	2,71*				
Norway B)	S	0,11*	0,09*	0,01	0,06*			
	T	6,19*	4,78*	0,51	3,44*			
Portugal B)	S	0,06*	0,05*	0,02		-0,06*		-0,04*
	T	3,06*	2,79*	0,83		-3,28*		-2,21*
Singapore C)	S			0,18*	-0,07*	-0,04*	-0,10*	-0,13*
	T			9,17*	-5,58*	-3,50*	-10,07*	-10,03*
Spain B)	S	0,11*	0,05*	-0,04	0,04*			-0,05*
	T	10,29*	4,98*	-2,71	5,00*			-4,57*
Sweden B)	S	0,10*	0,09*	-0,01				
	T	5,56*	4,72*	-0,48				
Switzerland B)	S	0,09*	0,07*	-0,05*				-0,05*
	T	5,39*	4,21*	-2,13*				-2,91*
Trinidad & Tobago A)	S	0,08*		0,08*			-0,05*	-0,06*
	T	4,38*		3,26*			-2,48*	-3,45*
USA C)	S	0,09*	0,06*	-0,03		0,06*		
	T	5,43*	3,59*	-1,27		3,30*		
Venezuela B)	S	0,06*	0,06*	0,03			-0,05*	
	T	3,68*	4,63*	1,16			-3,04*	
Slovenia B)	S	0,05*	0,07*	0,07*				-0,06*
	T	3,01*	3,93*	3,07*				-3,78*

A, B & C): Refers to model type. Positive values* in light shaded areas indicate female advantage, negative values* in dark shaded areas indicate male advantage at the 5%-level

Table 7 **Comparison of manifest and latent results. The number of countries showing significant gender relation in Pop A, n=25**

POP A	**Number of countries (n=25) showing either male or female significant advantage or balanced performances between the genders**		
Passage *Passage-factor*	**Female advantage**	**Balance**	**Male advantage**
Tot Doc Raw score	8	17	0
Tot Doc Rasch 1	3	21	1
Tot Doc Rasch 2	8	17	0
Gendoc	7	15	3
Island	1	13	11
Island'	0	16	9
Maria	20	5	0
Maria'	15	10	0
Bottles	3	20	2
Bottles'	1	22	2
Buses	9	15	1
Buses'	4	19	2
Content	25	0	0
Content'	22	3	0
Temperature	2	19	5
Temperature'	2	16	7

Again, the latent constructs which show females to have an advantage, are placed to the left in the table, and those appealing more to males are placed to the right.

The differences on the *Gendoc* factor among 9-year olds are partially different from the gender differences on the total Document score. About the same number of countries (10 out of 25) showed significant gender differences on the *Gendoc* as on the total score. However, the pattern of results was dramatically different in terms of which countries that showed a difference, and in which direction. The only country, for which the difference remained of the same magnitude, was in New Zealand (to the female advantage). In Denmark, Iceland, Singapore, Trinidad & Tobago and Slovenia the magnitude of the female advantage increased, compared to their advantage on the manifest score, as did the male advantage in Spain. Three countries showed significant gender differences on the latent score but not on the manifest (Hong Kong to the female advantage, Belgium and Switzerland to the male advantage), while the gender difference in Venezuela disappeared.

The pattern of gender differences in the passage specific components, after *Gendoc* has been taken into account, also turned out to be partially different from the pattern found in the observed Document sub scores.

The invariant female advantage on "Content" for all countries became more variant when the specific component *Content* was considered, and the magnitude of the relation changed in many countries.

The almost invariant female advantage on "Maria" is not so pronounced for the *Maria*-specific component. A total of 15 countries showed a significant female advantage compared to 20 in the univariate analysis.

The male advantage on "Island" in 11 out of 12 countries showing significant gender differences, held true for nine countries with a male advantage on the *Island* specific component.

An almost completely changed pattern of gender differences appeared on the passage specific component of *Temp* compared to the differences on the observed sub score "Temp" in terms of which country showing a significant difference.

On *Buses*, four countries showed a significant female advantage and two a significant male advantage, whereas eight showed female advantage and one showed male advantage on "Buses". Singapore reversed direction on its significant gender difference from female on the manifest to male on the latent dimension.

Most countries showed no significant gender difference on *Bottles*, but compared to the univariate analysis the differences decreased in number from five to three. *Bottles* turned out to favor males in two countries and to favor females in one country. Only in Portugal the difference remains the same when the two approaches are compared.

Only East Germany ended up with no gender differences at all in the multivariate analysis, followed by the Netherlands with only one difference. By showing differences in five latent constructs, Singapore and Spain revealed the highest number of gender differences.

To summarise the multivariate analysis in Pop A, it may be concluded that gender differences in any direction are less frequent compared to the univariate analysis. This is true for both *Gendoc* and for all six passage specific components. The pattern of gender differences on the *Gendoc* dimension was not so regular to the female advantage as in the univariate analysis of the total document score. The overall pattern of gender differences on the passage specific components reflected almost the same pattern as the univariate analysis although not necessarily for the same countries and also not with the same magnitude, and sometimes not in the same direction. *Island* and "Island" together with *Temperature* and "Temperature" favour males in many countries. Even more consistently are the advantages for females on *Content* and "Content", and on *Maria* and "Maria".

The results from the multivariate analysis for 14-year-olds are presented in Table 8, and Table 9 presents the frequency comparison between observed and latent differences.

Table 8 Gender differences in latent constructs and passage specific components for 14-year-olds (Pop B), Standardised estimates (S) and T-values (T).

Passage Country		*Trav Card* 1A	*Gen-doc*	*Aspi-rol* 9B	*Job vac* 3A	*Dire-tion* 6B	*Weat-her* 7B	*Bus* 5B	*Book let fact'*	*Hare & Lynx* 4A	*Reso-urcea* 2A	*Temp* 8B
Belgium B)	S	0,88*	0,06*	0,05*					-0,02	-0,07*	-0,13*	-0,06*
	T	7,17*	2,18*	2,78*					-0,64	-3,16*	-6,19*	-3,16*
Canada B)	S	0,21*	0,19*		0,10*		-0,04*	-0,04*	-0,22*			-0,10*
	T	10,84	8,02*		5,51*		-2,18*	-2,72*	-7,29*			-6,31*
Denmark B)	S	0,25*	0,10*		0,15*				-0,36*			-0,10*
	T	7,88*	3,77*		5,59*				-7,75*			-5,21*
Finland B)	S		0,01					0,06*	0,05			
	T		0,22					2,13*	1,54			
Germany East A)	S		0,13*						A)	-0,07*	-0,12*	-0,07*
	T		4,15*							-3,01*	-5,11*	-3,02*
Germany West A)	S	0,04*	0,07*					0,05*	A)	-0,12*	-0,13*	-0,11*
	T	2,74*	3,07*					-3,19*		-7,56*	-7,56*	-6,49*
Greece A)	S		0,14*		-0,05*			-0,07*	A)	-0,14*	-0,21*	-0,18*
	T		5,08*		-2,81*			-4,05*		-8,35*	-11,75*	-9,95*
Hong Kong A)	S	0,06*	-0,05	0,07*		0,04*	0,06*		A)		-0,09*	
	T	3,59*	-1,66	3,96*		2,19*	3,27*				-5,09*	
Hungary B)	S	0,11*	0,09*						-0,02	-0,10*		-0,08*
	T	4,94*	3,52*						-0,51	-5,87*		-4,36*
Iceland B)	S	0,27*	0,06	0,10*	0,18*	0,08*			-0,43*			
	T	10,55*	1,70	4,28*	6,34*	3,35*			-4,74*			
Ireland C)	S	0,08*	0,00	0,10*					0,13*		-0,04*	
	T	4,39*	-0,10	4,40*					4,46*		-2,41*	
Italy A)	S	0,13*	-0,18*	0,18*	0,17*	0,12*	0,07*	0,07*	A)			
	T	6,89*	-6,68*	9,21*	8,89*	6,44*	3,87*	3,47*				
The Netherlands C)	S		0,19*	0,05*	-0,04*	0,06*			-0,12*	*-0,12**	*-0,13**	
	T		5,60*	2,85*	-2,19*	3,17*			-3,15*	*-5,91**	*-6,32**	
New Zealand C)	S	-0,04*	0,07*	0,05*					-0,07*	-0,08*	-0,14*	-0,06*
	T	-2,30*	2,36*	2,53*					-2,15*	-4,38*	-6,87*	-3,16*
Norway C)	S	0,08*	-0,10*		0,07*				0,12*			
	T	3,72*	-2,95*		3,30*				2,76*			
Portugal A)	S	0,05*	-0,11*	0,07*	0,06*				A)	-0,12*	-0,16*	-0,08*
	T	2,57*	-3,72*	3,58*	3,02*					-6,45*	-8,98*	-4,37*
Singapore A)	S	0,04*	0,07*	0,04*					A)	-0,10*	-0,15*	-0,06*
	T	2,27*	2,29*	2,55*						-6,21*	-9,24*	-3,91*
Spain A)	S		-0,02	0,09*					A)	-0,11*	-0,10*	-0,07*
	T		-1,37	8,14*						-9,54*	-8,78*	-5,96*
Sweden B)	S	0,29*	0,09*		0,16*				-0,27*		0,08*	-0,04*
	T	8,90*	4,30*		5,32*				-5,12*		3,11*	-2,73*
Switzerland B)	S	0,16*	0,01	0,05*	0,09*				-0,20*			-0,05*
	T	8,99*	0,44	3,45*	5,26*				-5,77*			-3,44*
USA B)	S		0,06*	0,06*				0,05*	0,01	-0,10*	-0,07*	
	T		2,44*	3,62*				2,68*	0,60	-5,70*	-4,42*	
Slovenia A)	S	0,09*	0,09*					-0,05*	A)	-0,14*	-0,11*	-0,09*
	T	4,59*	2,67*					-2,19*		-7,56*	-5,68*	-8,82*

A, B & C) Refers to model type. Positive values* in light shaded areas indicate female advantage, negative values* in dark shaded areas indicate male advantage at the 5%-level.

Table 9 Comparison of manifest and latent results. The number of countries showing significant gender relation in Pop B, n=22

POP B	**Number of countries (n=22) showing either male or female significant advantage or balanced performances between the genders**		
Passage ***Passage-factor***	**Female advantage**	**Balance**	**Male advantage**
Tot Doc Raw score	6	11	5
Tot Doc Rasch 1	4	8	10
Tot Doc Rasch 2	5	7	10
Gendoc	13	6	3
Trav card	19	3	0
Trav card'	15	6	1
Resou	0	3	19
Resou'	1	8	13
Job vac	8	14	0
Job vac'	8	12	2
Hare & Lynx	0	1	21
Hare & Lynx'	0	10	12
Bus	8	12	2
Bus'	3	15	4
Direc	12	10	0
Direc'	4	18	0
Weather	5	14	3
Weather'	2	19	1
Temp	1	12	9
Temp'	0	8	14
Aspirol	9	12	1
Aspirol'	12	10	0

Again, the pattern of gender differences differs, to a certain degree, from the univariate analysis. The most dramatic difference concerns the general construct of document reading. A comparison between *Gendoc* and the total Doc score indicates more gender differences on the latent construct, and more to the female advantage. A male advantage was found in 3 countries, whereas a female advantage on the *Gendoc* factor was found in 13 countries. On the observed scores a female advantage was found in 6 countries at most (raw score), and a male advantage in 10 countries at most (Rasch 1&2).

The *Booklet1* factor identified in eight countries, favoured males in five and no significant differences in two. The *Booklet2* factor identified in four countries favoured females in two and males in two countries. These rather specific context factors do not have any matching manifest variable.

Differences in the passage specific components, after the above factors had been taken into account showed some similarities with the univariate analysis of passage sub scores. *Trav Card* almost consistently favoured females, whereas *Resource, Hare & Lynx* and *Temp* favoured males.

Some dissimilarities with the univariate analysis could also be seen. A consistent female advantage on "Direction" did not appear so obvious for *Direction*, whereas the female advantage on *Aspirol* became more pronounced compared to "Aspirol". *Job Vac* and *Bus* reflected to a higher degree a pattern of differences in both directions. The pattern of differences favouring males on *Resource, Hare & Lynx* were less striking compared to the observed scores, and finally, a somewhat more pronounced pattern of differences favouring males was found on *Temp* in terms of the number of countries compared to "Temp" in the univariate analysis.

Gender differences were most rare in Finland, showing differences on one passage specific component only, while gender differences were most common in Italy and The Netherlands showing differences on seven latent constructs.

To summarise the multivariate analysis in Pop B, it may be concluded that the pattern of gender differences, that is, direction and magnitude in each country, was quite different from the univariate analysis, although the frequency of gender differences were similar to the univariate analysis. This was true for both *Gendoc* and the passage specific components. However, the pattern of gender differences on the *Gendoc* dimension was more regular to the female advantage than the univariate analysis of the total document score indicated. Furthermore, for those countries where a second (residual) *Booklet*-factor was identified, a male advantage was the most common result.

Superficially, the overall pattern of gender differences on the passage specific components reflected almost the same pattern as the univariate analysis. A deeper look however, indicated that the differences were not necessarily for the same countries and not necessarily with the same magnitude, and sometimes not in the same direction. *Resource* and "Resource", *Hare & Lynx* and Hare & Lynx together with *Temp* and "Temp" favored males rather consistently. The same was true for females on *Trav card* and "Trav card", and on *Aspirol* and "Aspirol".

DISCUSSION AND CONCLUSIONS

Earlier, as has been specified, arguments for regarding Documents as requiring some additional skills compared to reading Expository or Narrative prose have been put forward both theoretically (Elley, 1994) and empirically (Balke, 1995; Gustafsson, 1995; Lundberg & Rosén, 1995). With this in mind, this study addressed two overall questions regarding gender differences in reading Documents. One concerned whether or not there exists any gender differences at all in Document reading proficiency. Related to this is the question of if there are any "global" gender differences, or if some differences vary with culture. The other question addressed the contribution of a multivariate approach to the previous questions. They will all be discussed under two subheadings; gender differences in Document reading skills, and gender differences in content specific dimensions of Documents.

Gender Differences in Document Reading Skills

The univariate analysis addresses the questions about differences in Document reading skills by looking at the total score, constructed either as raw scores or as Rasch scores. Gender differences were found in many countries in both populations. Females were more successful than males in eight countries among 9-year-olds, although most of the countries did not report any gender differences. A somewhat different pattern was found among the 14-olds, where six countries showed a male advantage and five a female advantage. Rasch scores favoured male performance more compared to raw scores in both populations, but particularly so in population B.

However, the above procedure assumes unidimensionality, which has proven to be too strong an assumption. In the multivariate analysis, the shared variance of all Document passages is partialled out into a *Gendoc* factor. For the 9-year-old sample, a well fitting model with one general dimension and six passage specific components was obtained. For the older sample, many countries reported an additional residual *Booklet* factor. The model in Pop B thus allowed for investigating gender differences in two latent variables along with 10 passage specific components.

One should not only interpret the *Gendoc* factor as a reflection of Document reading proficiency. There are good reasons to believe that *Gendoc* reflects a mixture involving other dimensions as well, like dimensions of general reading, reasoning , numerical, spatial and verbal abilities. The mixture may well be different in Pop A than in Pop B. Nevertheless, the *Gendoc* variable is much closer to a measure of Document reading skills than is any composite of observed variables.

The pattern of gender differences in the *Gendoc* factor came out very differently from the pattern obtained in the univariate analysis in both populations. In Pop A, about the same number of countries reported differences, but the pattern of results was quite different in terms of which countries showed a difference, and in which direction. Only one country showed the same differences regardless of approach.

In Pop B, the pattern was almost opposite compared to the univariate analysis, which showed male advantage more often than female. However, in 13 out of 16 countries with significant relations to the *Gendoc* factor, females were the favoured group. Considering that the females in Pop B reported more practice of document reading than the males (Elley, 1994), this latter pattern makes more sense than the univariate.

However, the female advantage in both populations may also be due to their previously reported advantage in more general reading and verbal skills. This pattern also indicates that a fair amount of the differences found in the univariate analysis are reflections of differences in other dimensions.

Gender Differences in Content Specific Dimensions of Document Tasks

In Pop B, 13 out of 22 countries reported an additional *Booklet* factor, derived from residual variance from all passages in either Booklet 1 or Booklet 2. The *Booklet* factor, in which seven countries reported a male advantage and two countries a female, is hard to interpret, and has no manifest equivalence. One hypothesis is that the *Booklet* factor is a reflection of the test situation, which many students experience as a competitive situation. Males, at least in Nordic countries, have been reported to benefit from such situations, whereas females have been reported to perform below their capacity in the same situation (Wernersson, 1989). However, the interpretation must be regarded as very preliminary. The factor is not well defined in the data of the present study, nor is it well specified theoretically. The factor is included in the model because it fits the data quite well, and because it may be possible to realise some sense from it. Also because it reflects some gender differences, and in a way helps to explain the observed (manifest) pattern of gender differences in some of the countries. The proportion of variance explained by the factor is fairly small, and although it varies across countries it should not be given too much attention at this stage. Further research is thus needed.

The univariate analysis of passage sub scores among 8-year-olds, reported an invariant or almost invariant female advantage on the sub scores of two passages ("Table of contents" and "Maria's timetable"). When keeping the relation between the genders to the latent factors under control, the signs of invariance decreased somewhat in the multivariate analysis of the passage specific components. However, the male advantage on "Island" and "Temperature" in many countries held true in the multivariate analysis too, as did the female advantage on "Maria" and "Content".

The same type of phenomenon was found in the older population. The multivariate analysis showed a rather consistent pattern of male advantage in three passages, after differences in *Gendoc* and *Booklet* had been taken into account ("Resources", "Hare & Lynx" and "Temperature"). Two passage specific components showed a consistent female advantage ("Traveller's card" and "Aspirol"). The rather consistent female advantage on "Direc" and "Job Vac" disappeared in the multivariate analysis. Along with the other passage specific components they reflected to a higher degree gender differences in both directions.

Another interesting feature of the multivariate analyses was that the magnitude and direction of gender differences in latent dimensions changed during the process of discovering significant relations. Not until all significant relations were included did the pattern of gender differences stabilise.

One may thus conclude that the univariate analysis disguises interesting patterns of gender differences. Furthermore, it may be concluded that the observed pattern of gender differences performance on Document reading tasks, is partly due to differences in the latent construct of Document reading proficiency and partly due to differences in passage specific components.

Final Conclusions

Gender differences in Document reading exist even though item and passage selection was conducted with gender in mind (Wagemaker, 1996). The hypothesis of an invariant global pattern of gender differences is not supported in the present analysis of Document reading, and thus not physiological explanations of the pattern either, as suggested by Taube & Munck (1996). Since the pattern of gender differences on the latent construct of Document reading proficiency turned out to vary across countries, the overall conclusion is that gender differences in Document reading skills is mainly a reflection of cultural influences. However, communalities in the pattern of differences among subgroups of countries are common.

The multidimensionality of performance on Document tasks has been demonstrated, which also explains why multivariate analysis is preferable in attempts to gain better understanding of the pattern of differences in performances on cognitive tasks. It may well be that most of the differences observed are not due to differences in the shared underlying dimension, but in more narrow dimensions similar to the passage specific components in this study.

Another conclusion that concerns the demonstrated mutivariate methodology is, that it is necessary to include all significant relations between latent constructs and the independent variable (as gender here) in order to obtain correct estimates for the differences.

Further analysis of gender differences in the document domain of reading literacy is of great interest. The description of the complex nature of Documents requires similar analysis of document task on the item level, or an item-parcel level. There may be more "going on" than a general dimension together with passage specific components. The document characteristics, for example, may be of importance, such as if the document information is numerical, spatial or verbal. Adding narrative and expository items or parcels to the analysis, will permit models including more general dimensions, and allow investigations if there is a variance left to be explained by one or more document dimensions. However, the cultural influence on the various parts of the IEA reading literacy test, probably makes it difficult to find such a detailed model that fits all the countries involved. Of course, separate country analysis or subgroups of countries will also contribute valuable information regarding gender differences.

The results in this paper may also serve as a starting point for analysing sociocultural influences on gender differences, an area rich of theoretical discourse and not equally rich in terms of empirical research. The multivariate technique demonstrated in this paper is also available for so called multi-level modeling (Muthén, 1994) where the variance from the country level can be separated from the individual level, and various effects of meso- and macro-level variables can be investigated. The possibilities for further research on this type of data material seem to endless.

REFERENCES

Anastasi, A. (1958). *Differential psychology*. New York: Macmillan.

Balke, G. (1995) Decomposition of reading comprehension - Analysis of the IEA literacy tests. Paper presented at the Annual Meeting of AERA, San Francisco.

Carroll, J, B. (1993) *Human cognitive abilities*. New York: Cambridge University Press.

Elley, W. B. (Ed.) (1994). *The IEA study of reading Literacy: Achievement and instruction in thirty-two school systems*. International studies in educational achievement. Great Britain, Exeter: Pergamon.

Emanuelsson, I., & Svensson, A. (1990). Changes in intelligence over a quarter of a century. *Scandinavian Journal of Educational Research, 34* (3), 171-187.

Gates, A. I. (1961). Sex differences in reading ability. *The elementary school Journal,* May, pp. 431-435.

Gould, S. J. (1981). *The mismeasure of man*. New York: W. W. Norton & Co.

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

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

Gustafsson, J-E. (1989). Broad and narrow abilities in research on learning and instruction. In R. Kanfer, P. L. Ackerman, & R. Cudeck (Eds.), *Abilities, motivation and methodology: The Minnesota Symposium in Learning and Individual Differences*. Hillsdale, N.J.: Erlbaum, pp. 203 - 237.

Gustafsson, J-E. (1992). The relevance of factor analysis for the study of group differences. *Multivariate Behavioral Research, 27* (2), 239 - 247.

Gustafsson, J-E. & Balke, G. (1993). General and specific abilities as predictors of school achievement. *Multivariate Behavioral Research, 28* (4), 407-434.

Gustafsson, J-E. (1994). Hierarchical models of intelligence and educational achievement. In A. Demetriou & A Etklides (Eds.) *Intelligence, Mind, and reasoning: Structure and development.* Elsevier Science B.V.

Gustafsson, J-E. (1995). Alternative hierarchical models of reading achievement. Paper presented at the Annual Meeting of AERA, San Francisco.

Gustafsson, J-E & Undheim, J. O (1996). Individual differences in cognitive functions. In D. Berliner & R. Calfee (Eds.) Handbook of Educational Psychology. New York: Macmillam

Guthrie, J. T., & Greaney, V. (1991). Literacy acts. In R. Barr, M. L. Kamill, P. Mosenthal, & P. D. Pearsson (Eds.) *Handbook of reading research* (Vol. 2). New York: Longman.

Halpern, D. K., (1992). *Sex differences in cognitive abilities.* 2nd edition. Hillsdale, N.J.: Erlbaum

Hyde, J, S. (1981). How large are cognitive gender differences? A meta-analysis using w2 and d. *American Psychologist, 36* (8), 892-901

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability: A meta-analysis. *Psychological Bulletin, 104* (1), 53-69.

Hyde, J, S,. Fennema, E., & Lamon, S, J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107* (2), 139-155.

Jacklin, C. N. (1989). Female and male. Issues of gender. *American Psychologist,* Vol. *44*, No. 2 , February, pp. 127-133.

Jöreskog, K. G. & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Hillsdale, N.J.: Erlbaum

Linn, M, C., & Petersen A, C. (1985). Emergence and characterization of differences in spatial ability: A Meta-Analysis. *Child development, 56,* 1479-1498.

Lohman, D. F. (1994). Component scores as residual variation (or why the intercepts correlates best). *Intelligence, 19,* (1), 1-11.

Lundberg, I. (1991). Cognitive aspects of reading. *International Journal Of Applied Linguistics*, 1, 151-163.

Lundberg, I., & Rosén, M. (1995). Structural modeling of reading achievement as a basis for evaluating teaching effects. Paper presented at the Annual Meeting of AERA, San Francisco, 1995.

Lundberg, I., & Høien, T. (Eds.). (1991). *Literacy in a world of change.* The Hauge: IEA.

Maccoby, E, D. (1966). *The development of sex differences.* Stanford University Press, Stanford California.

Maccoby, E, D. & Jacklin, C. (1974). *The psychology of sex differences.* Stanford University Press, Stanford California.

Mosenthal, P. B. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology, Vol. 88*, No. 2, 314-332.

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22* (3), 376-398.

Taube, K & Munck, I. M. (1996) Gender differences at the item level. In H. Wagemaker (Ed). *Are girls better readers? Gender differences in Reading Literacy in 32 countries.* The Hague: IEA

NAEP. (1982). National Assessment of Educational Progress. Reading Comprehension of American youth: Do they understand what they read? Results from the 1979-80 national assessment of reading and literature (Report No. NAEP - 11-R- 02) Princeton NJ

Rosén, M. (1995). Gender differences in structure, means and variances of hierarchically ordered ability dimensions. *Learning and Instruction,* 1, pp. 37-62.

Tavris, C. (1992). *The mismeasure of woman.* New York: Touchstone. Simon & Schulster Inc.

Thompson, B. G. (1975). Sex differences in reading attainment. *Educational Researcher,* vol. 18, pp. 16-23

Thorndike, R. L. (1973). *International studies in education (vol 3). Reading comprehension in 15 countries: an empirical study.* New York. Wiley

Tyler, L, E (1965). *The psychology of human differences*, 3d ed. New York: Appelton-Century-Crofts.

Wagemaker, H. (1996) (ed) *Are girls better readers? Gender differences in reading literacy in 32 countries.* The Hague: IEA

Wernersson, I. (1989). Olika kön samma skola? En kunskapsöversikt om hur elevernas könstillhörighet påverkar deras skolsituation. [Different sex same school? A research overwiev about how the students gender affect their school situation]*Skolöverstyrelsen, Vad säger forskningen F 89:1.*

GÖTEBORG STUDIES
IN EDUCATIONAL SCIENCES

ISSN 0436-1121

1. *Karl-Gustaf Stukát:* Lekskolans inverkan på barns utveckling. Sthlm 1966. Pp.148.
2. *Urban Dahllöf:* Skoldifferentiering och undervisningsförlopp. Sthlm 1967. Pp. 306.
3. *Erik Wallin:* Spelling. Factorial and experimental studies. Sthlm 1967. Pp.180.
4. *Bengt-Erik Andersson:* Studies in adolescent behaviour. Project Yg, Youth in Göteborg. Sthlm 1969. Pp. 400.
5. *Ference Marton:* Structural dynamics of learning. Sthlm 1970. Pp. 112.
6. *Allan Svensson:* Relative achievement. School performance in relation to intelligence, sex and home environment. Sthlm 1971. Pp. 176.
7. *Gunni Kärrby:* Child rearing and the development of moral structure. Sthlm 1971. Pp. 207.
8. *Ulf P. Lundgren:* Frame factors and the teaching process. A contribution to curriculum theory and theory on teaching. Sthlm 1972. Pp. 378.
9. *Lennart Levin:* Comparative studies in foreign-language teaching. Sthlm 1972. Pp. 258.
10. *Rodney Åsberg:* Primary education and national development. Sthlm 1973. Pp. 388.
11. *Björn Sandgren:* Kreativ utveckling. Sthlm 1974. Pp. 227.
12. *Christer Brusling:* Microteaching - A concept in development. Sthlm 1974. Pp. 196.
13. *Kjell Rubenson:* Rekrytering till vuxenutbildning. En studie av kortutbildade yngre män. Gbg 1975. Pp. 363.
14. *Roger Säljö:* Qualitative differences in learning as a function of the learner's conception of the task. Gbg 1975. Pp. 170.
15. *Lars Owe Dahlgren:* Qualitative differences in learning as a function of content-oriented guidance. Gbg 1975. Pp. 172.
16. *Marie Månsson:* Samarbete och samarbetsförmåga. En kritisk granskning. Lund 1975. Pp. 158.
17. *Jan-Eric Gustafsson:* Verbal and figural aptitudes in relation to instructional methods. Studies in aptitude - treatment interactions. Gbg 1976. Pp. 228.
18. *Mats Ekholm:* Social utveckling i skolan. Studier och diskussion. Gbg 1976. Pp. 198.
19. *Lennart Svensson:* Study skill and learning. Gbg 1976. Pp. 308.
20. *Björn Andersson:* Science teaching and the development of thinking. Gbg 1976. Pp. 180.
21. *Jan-Erik Perneman:* Medvetenhet genom utbildning. Gbg 1977. Pp. 300.

(cont.)

GÖTEBORG STUDIES
IN EDUCATIONAL SCIENCES

ISSN 0436-1121

(Continuation)

22. *Inga Wernersson:* Könsdifferentiering i grundskolan. Gbg 1977. Pp. 320.
23. *Bert Aggestedt and Ulla Tebelius:* Barns upplevelser av idrott. Gbg 1977. Pp. 440.
24. *Anders Fransson:* Att rädas prov och att vilja veta. Gbg 1978. Pp. 188.
25. *Roland Björkberg:* Föreställningar om arbete, utveckling och livsrytm. Gbg 1978. Pp. 252.
26. *Gunilla Svingby:* Läroplaner som styrmedel för svensk obligatorisk skola. Teoretisk analys och ett empiriskt bidrag. Gbg 1978. Pp. 269.
27. *Inga Andersson:* Tankestilar och hemmiljö. Gbg 1979. Pp. 288.
28. *Gunnar Stangvik:* Self-concept and school segregation. Gbg 1979. Pp. 528.
29. *Margareta Kristiansson:* Matematikkunskaper Lgr 62, Lgr 69. Gbg 1979. Pp. 160.
30. *Britt Johansson:* Kunskapsbehov i omvårdnadsarbete och kunskapskrav i vårdutbildning. Gbg 1979. Pp. 404.
31. *Göran Patriksson:* Socialisation och involvering i idrott. Gbg 1979. Pp. 236.
32. *Peter Gill:* Moral judgments of violence among Irish and Swedish adolescents. Gbg 1979. Pp. 213.
33. *Tage Ljungblad:* Förskola - grundskola i samverkan. Förutsättningar och hinder. Gbg 1980. Pp. 192.
34. *Berner Lindström:* Forms of representation, content and learning. Gbg 1980. Pp. 195.
35. *Claes-Göran Wenestam:* Qualitative differences in retention. Gbg 1980. Pp. 220.
36. *Britt Johansson:* Pedagogiska samtal i vårdutbildning. Innehåll och språkbruk. Gbg 1981. Pp. 194.
37. *Leif Lybeck:* Arkimedes i klassen. En ämnespedagogisk berättelse. Gbg 1981. Pp. 286.
38. *Biörn Hasselgren:* Ways of apprehending children at play. A study of pre-school student teachers' development. Gbg 1981. Pp. 107.
39. *Lennart Nilsson:* Yrkesutbildning i nutidshistoriskt perspektiv. Yrkesutbildningens utveckling från skråväsendets upphörande 1846 till 1980-talet samt tankar om framtida inriktning. Gbg 1981. Pp. 442.
40. *Gudrun Balke-Aurell:* Changes in ability as related to educational and occupational experience. Gbg 1982. Pp. 203.
41. *Roger Säljö:* Learning and understanding. A study of differences in constructing meaning from a text. Gbg 1982. Pp. 212.
42. *Ulla Marklund:* Droger och påverkan. Elevanalys som utgångspunkt för drogundervisning. Gbg 1983. Pp. 225.

(cont.)

GÖTEBORG STUDIES
IN EDUCATIONAL SCIENCES

ISSN 0436-1121

(Continuation)

43. *Sven Setterlind:* Avslappningsträning i skolan. Forskningsöversikt och empiriska studier. Gbg 1983. Pp. 467.
44. *Egil Andersson and Maria Lawenius:* Lärares uppfattning av undervisning. Gbg 1983. Pp. 348.
45. *Jan Theman:* Uppfattningar av politisk makt. Gbg 1983. Pp. 493.
46. *Ingrid Pramling:* The child's conception of learning. Gbg 1983. Pp. 196.
47. *Per Olof Thång:* Vuxenlärarens förhållningssätt till deltagarerfarenheter. En studie inom AMU. Gbg 1984. Pp. 307.
48. *Inge Johansson:* Fritidspedagog på fritidshem. En yrkesgrupps syn på sitt arbete. Gbg 1984. Pp. 312.
49. *Gunilla Svanberg:* Medansvar i undervisning. Metoder för observation och kvalitativ analys. Gbg 1984. Pp. 194.
50. *Sven-Eric Reuterberg:* Studiemedel och rekrytering till högskolan. Gbg 1984. Pp. 191.
51. *Gösta Dahlgren and Lars-Erik Olsson:* Läsning i barnperspektiv. Gbg 1985. Pp. 272.
52. *Christina Kärrqvist:* Kunskapsutveckling genom experimentcentrerade dialoger i ellära. Gbg 1985. Pp. 288.
53. *Claes Alexandersson:* Stabilitet och förändring. En empirisk studie av förhållandet mellan skolkunskap och vardagsvetande. Gbg 1985. Pp. 247.
54. *Lillemor Jernqvist:* Speech regulation of motor acts as used by cerebral palsied children. Observational and experimental studies of a key feature of conductive education. Gbg 1985. Pp. 146.
55. *Solveig Hägglund:* Sex-typing and development in an ecological perspective. Gbg 1986. Pp. 267.
56. *Ingrid Carlgren:* Lokalt utvecklingsarbete. Gbg 1986. Pp. 299.
57. *Larsson, Alexandersson, Helmstad and Thång:* Arbetsupplevelse och utbildningssyn hos icke facklärda. Gbg 1986. Pp. 165.
58. *Elvi Walldal:* Studerande vid gymnasieskolans vårdlinje. Förväntad yrkesposition, rollpåverkan, självuppfattning. Gbg 1986. Pp. 291.
59. *Eie Ericsson:* Foreign language teaching from the point of view of certain student activities. Gbg 1986. Pp. 275.
60. *Jan Holmer:* Högre utbildning för lågutbildade i industrin. Gbg 1987. Pp. 358.
61. *Anders Hill and Tullie Rabe:* Psykiskt utvecklingsstörda i kommunal förskola. Gbg 1987. Pp. 112.
62. *Dagmar Neuman:* The origin of arithmetic skills. A phenomenographic approach. Gbg 1987. Pp. 351.

(cont.)

GÖTEBORG STUDIES
IN EDUCATIONAL SCIENCES

ISSN 0436-1121

Editors:

Ingemar Emanuelsson, Jan-Eric Gustafsson and Ference Marton

(Continuation)

63. *Tomas Kroksmark:* Fenomenografisk didaktik. Gbg 1987. Pp. 373.
64. *Rolf Lander:* Utvärderingsforskning - till vilken nytta? Gbg 1987. Pp. 280.
65. *Torgny Ottosson:* Map-reading and wayfinding. Gbg 1987. Pp. 150.
66. *Mac Murray:* Utbildningsexpansion, jämlikhet och avlänkning. Gbg 1988. Pp. 230.
67. *Alberto Nagle Cajes:* Studievalet ur den väljandes perspektiv. Gbg 1988. Pp. 181.
68. *Göran Lassbo:* Mamma - (Pappa) - barn. En utvecklingsekologisk studie av socialisation i olika familjetyper. Gbg 1988. Pp. 203.
69. *Lena Renström:* Conceptions of matter. A phenomenographic approach. Gbg 1988. Pp. 268.
70. *Ingrid Pramling:* Att lära barn lära. Gbg 1988. Pp. 115.
71. *Lars Fredholm:* Praktik som bärare av undervisnings innehåll och form. En förklaringsmodell för uppkomst av undervisningshandlingar inom en totalförsvarsorganisation. Gbg 1988. Pp. 364.
72. *Olof F. Lundquist:* Studiestöd för vuxna. Utveckling, utnyttjande, utfall. Gbg 1989. Pp. 280.
73. *Bo Dahlin:* Religionen, själen och livets mening. En fenomenografisk och existensfilosofisk studie av religionsundervisningens villkor. Gbg 1989. Pp. 359.
74. *Susanne Björkdahl Ordell:* Socialarbetare. Bakgrund, utbildning och yrkesliv. Gbg 1990. Pp. 240.
75. *Eva Björck-Åkesson:* Measuring Sensation Seeking. Gbg 1990. Pp. 255.
76. *Ulla-Britt Bladini:* Från hjälpskolelärare till förändringsagent. Svensk speciallärarutbildning 1921-1981 relaterad till specialundervisningens utveckling och förändringar i speciallärarens yrkesuppgifter. Gbg 1990. Pp. 400.
77. *Elisabet Öhrn:* Könsmönster i klassrumsinteraktion. En observations- och intervjustudie av högstadieelevers lärarkontakter. Gbg 1991. Pp. 211, XXI.
78. *Tomas Kroksmark:* Pedagogikens vägar till dess första svenska professur. Gbg 1991. Pp. 285.
79. *Elvi Walldal:* Problembaserad inlärning. Utvärdering av påbyggnadslinjen Utbildning i öppen hälso- och sjukvård. Gbg 1991. Pp. 130.
80. *Ulla Axner:* Visuella perceptionssvårigheter i skolperspektiv. En longitudinell studie. Gbg 1991. Pp. 293.
81. *Birgitta Kullberg:* Learning to learn to read. Gbg 1991. Pp. 352.
82. *Claes Annerstedt:* Idrottslärarna och idrottsämnet. Utveckling, mål, kompetens - ett didaktiskt perspektiv. Gbg 1991. Pp. 286.

(cont.)

GÖTEBORG STUDIES
IN EDUCATIONAL SCIENCES

ISSN 0436-1121

(Continuation)

83. *Ewa Pilhammar Andersson:* Det är vi som är dom. Sjuksköterskestuderandes föreställningar och perspektiv under utbildningstiden. Gbg 1991. Pp. 313.

84. *Elsa Nordin:* Kunskaper och uppfattningar om maten och dess funktioner i kroppen. Kombinerad enkät- och intervjustudie i grundskolans årskurser 3, 6 och 9. Gbg 1992. Pp. 253.

85. *Valentin González:* On human attitudes. Root metaphors in theoretical conceptions. Gbg 1992. Pp. 238.

86. *Jan-Erik Johansson:* Metodikämnet i förskollärarutbildningen. Bidrag till en traditionsbestämning. Gbg 1992. Pp. 347.

87. *Ann Ahlberg:* Att möta matematiska problem. En belysning av barns lärande. Gbg 1992. Pp. 353.

88. *Ella Danielson:* Omvårdnad och dess psykosociala inslag. Sjuksköterskestuderandes uppfattningar av centrala termer och reaktioner inför en omvårdnadssituation. Gbg 1992. Pp. 301.

89. *Shirley Booth:* Learning to program. A phenomenographic perspective. Gbg 1992. Pp. 308.

90. *Eva Björck-Åkeson:* Samspel mellan små barn med rörelsehinder och talhandikapp och deras föräldrar - en longitudinell studie. Gbg 1992. Pp. 345.

91. *Karin Dahlberg:* Helhetssyn i vården. En uppgift för sjuksköterskeutbildningen. 1992. Pp. 201.

92. *Rigmor Eriksson:* Teaching Language Learning. In-service training for communicative teaching and self directed learning in English as a foreign language. 1993. Pp. 218.

93. *Kjell Härenstam:* Skolboks-islam. Analys av bilden av islam i läroböcker i religionskunskap. Gbg 1993. Pp. 312.

94. *Ingrid Pramling:* Kunnandets grunder. Prövning av en fenomenografisk ansats till att utveckla barns sätt att uppfatta sin omvärld. Gbg 1994. Pp. 236.

95. *Marianne Hansson Scherman:* Att vägra vara sjuk. En longitudinell studie av förhållningssätt till astma/allergi. Gbg 1994. Pp. 236.

96. *Mikael Alexandersson:* Metod och medvetande. Gbg 1994. Pp. 281.

97. *Gun Unenge:* Pappor i föräldrakooperativa daghem. En deskriptiv studie av pappors medverkan. Gbg 1994. Pp. 249, [33].

98. *Björn Sjöström:* Assessing acute postoperative pain. Assessment strategies and quality in relation to clinical experience and professional role. Gbg 1995. Pp. 159.

99. *Maj Arvidsson:* Lärares orsaks- och åtgärdstankar om elever med svårigheter. Gbg 1995. Pp. 212.

(cont.)

GÖTEBORG STUDIES
IN EDUCATIONAL SCIENCES

ISSN 0436-1121

(Continuation)

100. *Dennis Beach:* Making sense of the problems of change: An ethnographic study of a teacher education reform. Gbg 1995. Pp. 385.
101. *Wolmar Christensson:* Subjektiv bedömning - som besluts och handlingsunderlag. Gbg 1995. Pp. 211.
102. *Sonja Kihlström:* Att vara förskollärare. Om yrkets pedagogiska innebörder. Gbg 1995. Pp. 214.
103. *Marita Lindahl:* Inlärning och erfarande. Ettåringars möte med förskolans värld. Gbg. 1996. Pp. 203.
104. *Göran Folkestad:* Computer Based Creative Music Making - Young Peoples´ Music in the Digital Age. Gbg 1996. Pp. 237.
105. *Eva Ekeblad:* Children • Learning • Numbers. A phenomenographic excursion into first-grade children's arithmetic. Gbg 1996. Pp. 370.
106. *Helge Strömdahl:* On *mole* and *amount of substance*. A study of the dynamics of concept formation and concept attainment. Gbg 1996. Pp. 278.
107. *Margareta Hammarström:* Varför inte högskola? En longitudinell studie av olika faktorers betydelse för studiebegåvade ungdomars utbildningskarriär. Gbg 1996. Pp. 263.
108. *Björn Mårdén:* Rektorers tänkande. En kritisk betraktelse av skolledarskap. Gbg 1996. Pp. 219.
109. *Gloria Dall'Alba and Biörn Hasselgren (Eds.).* Reflections on Phenomenography - Toward a Methodology? Gbg 1996. Pp 202.
110. *Elisabeth Hesslefors Arktoft:* I ord och handling. Innebörder av "att anknyta till elevers erfarenheter", uttryckta av lärare. Gbg 1996. Pp. 251.
111. *Barbro Strömberg*: Professionellt förhållningssätt hos läkare och sjuksköterskor. En studie av uppfattningar. Gbg 1997. Pp 241.
112. *Harriet Axelsson*: Våga lära. Om lärare som förändrar sin miljöundervisning. Gbg 1997. Pp 326.
113. *Ann Ahlberg*: Children's ways of handling and experiencing numbers. Gbg 1997. Pp 115.
114. *Hugo Wikström*: Att förstå förändring. Modellbyggande, simulering och gymnasieelevers lärande. Gbg 1997. Pp 305.
115. *Doris Axelsen*: Listening to recorded music. Habits and motivation among high-school students. Gbg 1997. Pp 226.
116. *Ewa Pilhammar Andersson:* Handledning av sjuksköterskestuderande i klinisk praktik. Gbg 1997. Pp 166.
117. *Owe Stråhlman:* Elitidrott, karriär och avslutning. Gbg 1997. Pp 350.
118. *Aina Tullberg*: Teaching the 'mole'. A phenomenographic inquiry into the didactics of chemistry. Gbg 1997. Pp 200.

(cont.)

GÖTEBORG STUDIES
IN EDUCATIONAL SCIENCES

ISSN 0436-1121

119. *Dennis Beach*: Symbolic Control and Power Relay: Learning in Higher Professional Education. Gbg 1997. Pp 259.
120. *Hans-Åke Scherp*: Utmanande eller utmanat ledarskap. Rektor, organisationen och förändrat undervisningsmönster i gymnasieskolan. Gbg 1998. Pp 228.
121. *Staffan Stukát*: Lärares planering under och efter utbildningen. Gbg 1998. Pp 249.
122. *Birgit Lendahls Rosendahl*: Examensarbetets innebörder. En studie av blivande lärares utsagor. Gbg 1998. Pp 222.
123. *Ann Ahlberg:* Meeting Mathematics. Educational studies with young children. Gbg 1998. Pp 236.
124. *Monica Rosén*: Gender Differences in Patterns of Knowledge. Gbg 1998. Pp 210.

Subscriptions to the series and orders for single volumes should be addressed to:
ACTA UNIVERSITATIS GOTHOBURGENSIS, Box 222, SE-405 30 Göteborg, Sweden

ISBN 91-7346-333-7