Immigrant Groups and Cognitive Tests

GOTHENBURG STUDIES IN EDUCATIONAL SCIENCES 333

Immigrant Groups and Cognitive Tests

Validity Issues in Relation to Vocational Training

Ann Valentin Kvist

UNIVERSITY OF GOTHENBURG
ACTA UNIVERSITATIS GOTHOBURGENSIS

ISBN 978-91-7346-741-4
ISSN 0436-1121

Thesis in Education at the Department of Education and Special Education.
The thesis is also available in full text at:
http://hdl.handle.net/2077/32549

Subscriptions to the series and orders for individual copies sent to:
Acta Universitatis Gothoburgensis, PO Box 222, SE 405 30 Göteborg, or to
acta@ub.gu.se

Cover photo: Karin Kvist

Print: Ineko AB, Kållered 2013

Abstract

Title: Immigrant Groups and Cognitive Tests - Validity Issues in Relation to Vocational Training
Language: English with a Swedish summary
Keywords: Cognitive test use, immigrant groups, validity, vocational education.
ISBN: 978-91-7346-741-4

Psychologists at the Employment Service use cognitive tests to assess suitable work or training areas. The tests are standardized on a Swedish population and possible differences in cognitive constructs, prognostic properties and social consequences, when they are used for individuals with non-Swedish background, are not well known. The aim of this thesis is to investigate questions of validity in relation to cognitive test use with immigrant groups, in a setting of assessment for vocational training.

Validity aspects are probed in three studies. Data are supplied from the Public Employment Service and Statistics Sweden. The methods used belong to the family of Structural Equation Modeling.

In Study I aspects of the structure of cognitive functions are investigated and interpreted in relation to Cattell's (1987) Investment Theory. The results give support to Cattell's theory only when individuals have similar cultural backgrounds. Study I also indicates that the major difference between groups in mean levels of cognitive achievement is to be found in areas with strong cultural influence, such as language, while fluid intelligence shows small difference. Finally Study I shows that the pattern of influence from different cognitive factors on specific tests is similar over groups.

Study II has focus on the way the psychologists interpret test results, and integrate them into a summary evaluation of suitability for an area of study. Psychologists are lenient in assessing suitability for lower levels of cognitive functioning in the verbal area, but fail to give full credit to fluid intelligence for the immigrant groups.

In Study III cognitive test results, the psychologists' assessments, and having been granted vocational education are investigated in relations to employment rates. The substantially lower employment rates of the immigrant groups can only to a very small part be explained by results on cognitive functions.

The results from the studies are summed up in a discussion, which largely supports the validity of the use of cognitive tests for immigrant groups. Some suggestions for policy and for future research are given.

Table of contents

Acknowledgements

This work has been in progress for more than a decade and is the result of the inspiration, decisions, and contributions from many people. All deserve thanks and should be mentioned, but some may be unintentionally omitted. I start by apologizing for this!

The base for this work was laid through the work of the psychologists at the Employment Agency. More than 3600 assessments were carried out and over the years this involved the work of a dozen psychologists, as well as administrative personnel. Special thanks go to Håkan Freidlitz, who contributed both with a sizable part of the material and inspiring discussions and suggestions. Foresighted decisions made by managers at different levels made it possible to turn this into a research project. Ulla-Britt Selander approved the initial methods project, including consultations and work with Bertil Mårdberg. This put me on the track to the family of Structural Equations Models. Staffan Nilsson approved taking the project into a formal doctoral dissertation. Caterina Holmquist eased the way when the voyage was rough.

At Gothenburg University it was a privilege to be introduced to the FUR group, and be part of this special learning environment. It was also a distinctive privilege to have Jan-Eric Gustafsson, outstanding in his methodological knowledge and in the field of SEM, as a mentor. Funding from IFAU contributed financial support. At the last leg of this trip Christina Cliffordson turned out to be an encourager and a doer. Thank you, all!

Dissertation work demands some self-obsession and social seclusion. Family and friends have been more than patient with me. Special thanks to Lena and Henry Kvist, who generously supplied room and board as well as inspiring discussions when I was in Gothenburg. And to my husband, Lars, and children Karin, Peter, Anna, and Erik: You have been patient, encouraging, and supporting, more than anyone could ask. You made sure that life included joyous breaks and relaxing moments where work and research took their proper proportions. My heartfelt thanks!

1 Introduction

An increasing number of individuals move over cultural and language barriers, and try to establish a working life on a labor market different from that of their original culture. Demands on cognitive abilities in Swedish working life are high, and probably increasing. Test that assess cognitive resources have proved to be a very useful tool when predictions of success in work or studies are concerned. Thus, cognitive tests are used abundantly in the Western world, as a tool for guidance, but also as a basis for selection of individuals into work or work training. The now more than century long use of cognitive tests for these purposes has created considerable benefits (e.g. Cook, 1993, Schmidt & Hunter, 1998) but also caused great debate (e.g. Gould, 1996; Neisser et al. 1996).

Advocates of test use often emphasize the efficiency aspects. In the field of industrial and organizational psychology the use of aptitude or intelligence tests has been shown to be an efficient method to match the resources of the individual to different work opportunities. Although work samples tests show the highest single validity in predicting overall job performance (Schmidt & Hunter, 1998) these are only possible to use when applicants are experienced in the relevant area. When for this reason work samples tests cannot be used cognitive tests measuring general mental ability provide the highest predictive validity for job performance and job training (Salgado & Anderson, 2003; Salgado, Anderson, Moscosco, Bertua, & de Fruyt, 2003; Schmidt & Hunter, 1998, 2004; Schmidt, Shaffer & Oh, 2008). The possibility of wide application is only one advantage of cognitive test use. Compared to other selection procedures the use of cognitive tests is usually more cost efficient. Also, the advantage of selecting applicants on the basis of cognitive test results increases as job content becomes more complex (Hunter, 1986; Salgado, Anderson, Moscosco, Bertua, de Fruit & Rollande, 2003). According to Salgado and Anderson (2002) the use of cognitive ability tests is more common in Europe than in the United States.

Adversaries of test use are often concerned with equity aspects. Most cognitive tests have been developed and their value assessed in a specific cultural context. Test use needs to be scrutinized to meet the challenges of a more multi-cultural working life. Specific test based procedures may produce disadvantages for individuals or groups. When test scores are used as a basis for selection even

fairly small differences in group means will produce different selection rates for different groups. While test use is more or less accepted for the majority population questions appear when the subject belongs to a minority group. Are the processes and structures captured by the tests the same in different cultural groups? Are the psychometric properties the same across groups? Byrne, Leong, Hambleton, Oakland, van der Vijver, and Cheung (2009) discuss these aspects, and how they should be considered when tests are translated into a new culture or language, under the labels of structural and measurement equivalence. However, translation and adaption of tests, to secure that the measurement properties meet requirements, is only one aspect of the complexities that may arise when tests are used in diverse groups. Other important questions concern how possible group mean differences could be handled in order to secure, as far as possible, both performance and ethnic diversity in working life (e.g. Sackett, Schmitt, Ellington, & Kabin, 2001). When test use brings different consequences for different groups this poses questions that should be considered in relation to consequential validity. If consequential validity, as defined by Messick (1989) is fully accepted as an aspect of validity, this will imply that validation projects must be extended from being scientific, empirical enterprises to include socio-political processes as well (Crocker, 1997).

Test use in connection with minority groups thus demands some consideration. There is a risk that test results may be difficult to evaluate or produce biased outcomes. The quandaries in connection with test use for minority groups may make psychologists decide against using cognitive tests when the person being tested does not belong to the majority population. It is also possible that psychologists will use tests but be more reluctant to draw conclusions from test results. While it may be prudent with such caution it can also have adverse effects. The person who has been excluded from testing may also be excluded from the possibility of being selected into a desirable job or training position. Cognitive test results may reveal specific or general cognitive difficulties and the knowledge of the magnitude and nature of the difficulties makes it possible to choose suitable tasks, to adapt a task, to provide special support or to compensate for the troubles. When test results indicate signs of illness a cure may be provided. Thus, practices of test use, including the lack of test use, may produce differential outcomes for individuals with minority background. This makes the area of test use in multi-cultural settings pressing to investigate.

Many countries, including Sweden, have a steady flow of immigrants into the work force. Immigrants tend to have a cultural background that is increasingly

diverse. Also, for each society the groups to be considered are different. Salgado and Anderson (2002) list the many substantial minority group communities in Europe, either native to different countries or who have immigrated as migrant workers over more recent years. They lament the fact that the area of subgroup differences related to cognitive test use has been largely neglected in Europe. They mention the lack of harmonization of legislative structures across Europe, and thus the low risk of legal action, as one reason for the lack of research into areas such as subgroup differences, or adverse impact. In comparison, implications and consequences of test use for minority groups is an area thoroughly scrutinized and debated in the United States; in the public, legal and scientific societies. Although American knowledge and conclusions can bring valuable contributions to other societies, it is also necessary to consider the particular context of each country. As Salgado and Anderson (2003) point out European societies show great variation in testing practices, values, laws, and not least languages. Thus, in each country different research projects are necessary.

If a testing procedure is to produce meaningful results, the basis for interpretation and prediction must be well grounded. Although the body of knowledge concerning cultural influence on cognitive functions as reflected in test results has grown some in recent years (e.g. Salgado, Anderson, Moscosco, Bertua & Fruyt, 2003; te Nijenhuis & van der Flier, 1997, 1999, 2000) there are still many unanswered questions. Salgado and Anderson (2002) find it an important and potentially fruitful area of research. They state that Sweden is ranked among the countries with rather frequent use of testing in selection situations. Still, effects of test use with minority groups have not been intensely researched in this country. In evaluating test use it is important to analyze and discuss both efficiency and equity aspects. A key factor for both efficiency and equity aspects is validity. With poor validity there is no efficiency. With bias in testing procedure or results there are threats to equity.

Questions that have been highlighted in relation to test use with immigrant groups can all be grouped under the heading of validity. Validity in test use encompasses a multitude of aspects. The theoretical framework provides the basis for test construction and meaningful interpretations of test results. Other aspects are the use of test score information as a base for decisions, and social and economic effects of test use for the individual. The different aspects form a chain, where every link contributes to the total strength of the evaluation.

1.1 Aim and outline of the thesis

This dissertation has its focus on validity aspects of cognitive test use when immigrant groups in Sweden are concerned. Specifically the use of tests as a basis for selection into vocational training and the impact of test based selection and training, as practised at the Swedish Public Employment Service[1], will be investigated. This involves investigating construct validity of measures of cognitive functions for the groups, i.e. questions about the underlying cognitive functions and how they may differ between cultural groups. It also concerns the process of test interpretation, and investigation of the prognostic properties of test results, as well as consequences of use of test scores. In sum the aim is to evaluate efficiency and equity of test use for vocational purposes in a multi-cultural setting in Sweden.

The more detailed aims are made explicit in the three studies. These will form links in the attempt to investigate the whole process of test use.

The layout of the thesis is as follows:

First, the theme of the thesis is introduced, leading up to the aims. Then some contextual background will be presented: immigration to Sweden over time; the mission and work of the Employment Service; and the public regulation of test use. Next comes a theoretical part that introduces theories of validity, and theories of intelligence and cognitive functions. These theories form the framework for the research questions. The design to explore them, with perspectives and questions of method, follows. A summary of the results from the three studies comes next. Finally, from the perspective of the research questions the outcomes of the studies are discussed and evaluated, and the limitations and need for further studies is outlined.

[1] Arbetsmarknadsverket (AMV), with the national Labour Market Board and around 20 County Labour Boards, was changed into the new authority Public Employment Service (PES) January 1, 2008. For clarity, the latter term, or its short form Employment Service, will be used in this thesis also for periods before 2008.

2 Contextual background

This section will provide a picture of the immigrant's establishment in Sweden, with focus on entrance into the work force. The role of the Swedish Employment Service will be described, and vocational education will be presented. This section will also describe test use practices, in relation to vocational guidance and selection, at the Employment Service in Sweden.

2.1 Migration into Sweden, public policies, and a new working life

In recent decades Sweden has received immigrants and refugees to an extent that has had considerable impact on society. According to Statistics Sweden 12.2 percent of the population were foreign born in 2005. This share had risen to 14.9 percent by the end of 2011. With these numbers Sweden is on a level with USA and Germany, has a higher proportion of immigrants compared to the Netherlands, France and the United Kingdom, and a considerably higher proportion compared to the Nordic neighbors Norway, Denmark and Finland (Integration Report, 2005).

During the first decades after the Second World War the immigrants were mainly job seekers, who came from southern and Eastern Europe. There was also a substantial inflow of job seekers from Finland. The official Swedish policy focused on immigration, rather than integration (Integrationsverket, 2007). Sweden needed workers, and a majority of the immigrants quickly entered the work force. At this time the employment rate for foreign born was higher than that for Swedish born.

A start for a more articulated Swedish government policy can be said to be found in 1975 (Prop. 1975:26), when guidelines for Swedish official policy were expressed in three goals; equality, freedom of choice, and cooperation. Equality implied that immigrants should be citizens fully equal to the majority. Freedom of choice referred to the right for the immigrant to keep and nourish a cultural identity of the original culture or the Swedish. Co-operation, finally, referred to the Swedish model of negotiating solutions based on compromise. The state encouraged the immigrants to form their own organizations, partly with the purpose of creating a negotiating partner. The Canadian model of multi-culturalism was the inspiration behind these goals (Integrationsverket, 2007). There was an expectation that questions of culture would be handled in the

traditional Swedish negotiation model, where individuals were represented by organizations. At this time instruction in Swedish for immigrants was established as a right.

In the last decades of the 20th century the immigration patterns changed into an influx of immigrants consisting mainly of refugees and their families (Lemaître, 2007). Due to the unrest in the Balkan in the 1990s more refugees came from this area, and in the last decades there has been a substantial inflow of refugees from the Middle East, Northern Africa and Afghanistan. Thus, both the cultural distance and the reasons for immigration have changed over the last decades. In the same time period the demand at the labor market fell. In response to the new situation Swedish official policy was reframed (Integrationsverket, 2007). The Government adopted a new rule (Prop. 1997/98:16) which framed the questions of immigration in a context which concerned all citizens and society at large. The basic values were expressed as equal rights and opportunities, a society based on diversity, and mutual respect and tolerance in which all, regardless of ethnic or cultural background, are participants and responsible. The political work should especially focus on support for individual self-sustenance, participation in society, promotion of democratic values, equal rights for men and women, and prevention of discrimination and racism.

2.1.1 Immigration and employment rates

Although a wealth of official documents emphasize the importance of employment as a road to integration, employment rates for foreign born have not developed in the desired direction. In the middle of the 1970s the gap between native and foreign born started to increase. The gap remained both during the economic boom of the 1980s and the economic recession of the early 1990s. In the recovery years of late 1990s the gap again closed somewhat. Figure 1 illustrates the development over the years 1987-2003. These years roughly cover the time when the empirical material for this dissertation was collected.

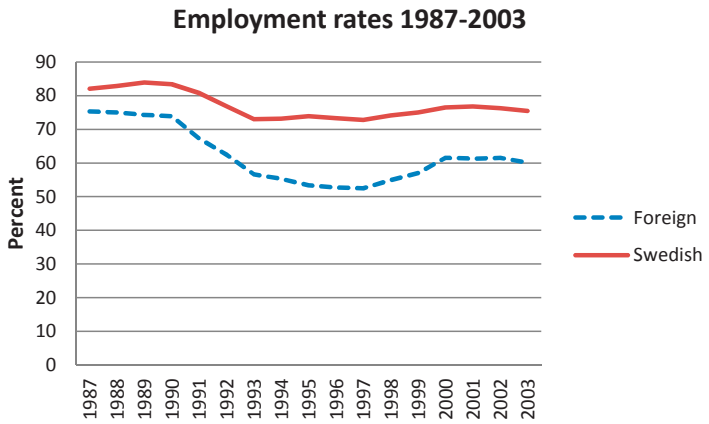

Figure 1. Employment rates for Swedish and foreign born in the years 1987-2003. Adapted and translated into English by the author. Source: Integrationsverket, 2004, p. 238.

Reasons for changes in employment rates for immigrant groups are difficult to pinpoint since many factors interact. However, it seems that one factor has a more prominent relation to changes in labor demand, and this is time of residence in Sweden. The groups with short time in Sweden are most vulnerable to economic cycles (Integrationsverket, 2003). The same source also informs that during all cycles immigrants born in Africa or Asia have lower employment rates than people with other background. Pocket facts – Statistics on Integration (2006) confirm that level of employment among people born in countries in Africa, Asia or Europe outside EU15 is lower than for those born in other regions, and that these differences cannot be explained by differences in age, education or civil status.

OECD (2004) notes that a higher share of humanitarian immigration compared to other forms of migration is related to slower labor market integration in all countries. It could be hypothesized that refugees and their families have a more difficult immigration process and a higher degree of traumatization. Possibly there could also be a greater cultural distance. Countries with closer geographical distances will usually show greater resemblances in social patterns and structures in working life, which can facilitate integration. Sjögren and Zenou (2007) discuss aspects of depreciation (such as cultural distance, reasons for migration, physical and psychological health, and loss of social networks) and new investment of human capital as aspects that affect entering the work force. Segendorf and Teljusto (2011) mention three reasons that give immigrants difficulties at the labor market: devaluation of human capital (specifically cultural knowledge, languages), thresholds of the market (e.g.

high wages for simple jobs, discrimination), and lack of network. Eriksson (2011) highlights the possibility that some immigrant groups have a lower degree of human capital in terms of education, labor market experience, and language skills. He also mentions the importance of network and job seeking patterns, and discrimination of different kinds, as factors that may influence employment rates.

Difficulties in work force integration thus may be due to factors residing in the receiving society or in the immigrant groups, and be different for the same groups in different societies, and vice versa. It has been shown that although immigrant women have higher absolute working rates in Sweden than immigrant women elsewhere, the gap in joblessness between immigrant and native-born men is larger in Sweden than in comparable countries such as Great Britain and Germany (Kesler, 2006). Sweden has also been known to have higher demands concerning command of the Swedish language than other nations of comparable development (Myrberg, 2001). It has been hypothesized that this is due to the relative flat hierarchical structures in Swedish working life, where instructions, manuals and problem shooting processes are handled by the single worker, rather than by a supervisor.

Research on the situation of the immigration groups at the labor market has a focus on group level explanations. However, it is important to note that immigrants differ in their characteristics, both between and within groups. In the single case the success in achieving a place in the working life of the receiving country will depend on the circumstances for the specific individual, among them his or her adaptability and other resources.

2.1.2 Learning the language – Swedish for immigrants

Newly arrived adult immigrants will meet an institution known as Swedish for immigrants, usually referred to by its acronym SFI. Adult immigrants who are registered with a permanent residence in Sweden are entitled to take part in instruction, free of charge. SFI provides instruction in the Swedish language and knowledge about Swedish society. According to the governing document (Skolverket, 2009) the purpose is to provide the students with the linguistic tools that enable them to communicate and be active participants in daily life, the life of the community, and working life.

In middle of the 1960s SFI was offered by independent educational associations free of charge for the students. There were no official guidelines and the quality varied considerably (Mårtensson, 2009). In 1984 the Government introduced a bill (Prop. 1983/84:199) which established the national

responsibility and organization of SFI. The bill also established that employed participants had the right to take time off from work to take part in the instruction and that they should be economically compensated. In 1994, a new bill gave additional instructions for the organization and content, such as examinations and grades (Förordning, 1994). The regulation from 1994 is still largely in force, but has been updated in details.

A report from the Swedish Schools Inspectorate (Skolinspektionen, 2010) describes the special circumstances of the instruction at SFI; students enter and finish their studies continually, studies can be combined with work or practice, students are often traumatized, interruptions in studies due to childbirth are common. Still, mastering the Swedish language is an important prerequisite for integration and entering the work force. Participation in SFI has continually increased in numbers. In 1997 around 38 000 persons took part. This number had risen to more than 84 000 in 2008, and reached 102 400 by 2011 (Skolverket, 2011). The number for 2011 represents an increase by seven percent from the year before. The most common native language among the students is Arabic, which is spoken by one fifth of the participants.

2.2 The Employment Service

According to the Instruction for the Employment Service (Svensk Författningssamling, 2007) its basic task is to improve the function of the labor market by efficient matching of employers and job seekers, while prioritizing those who have a weak position in the labor market. The Employment Service should also strive to contribute to long term increase in employment rates. A specific goal (3§ 8) is to increase diversity and to counteract discrimination. Another specified task (6§) is the responsibility to offer new immigrants services that promote a speedy and efficient establishment in the labor market. The Employment Service has a coordinating, supporting and pushing role in relation to other institutions in this field.

Up until 1985 the Employment Service had the main responsibility to receive new immigrants in Sweden. At that time the task was transferred to the local communities. In 2010 the responsibility for the reception of refugees and their families was again reverted to the Employment Service. Integrationsverket (2007) notes that the emphasis on services from the Employment Service in integrating new immigrants in Swedish working life has been given continually stronger emphasis in the yearly instructions from the Government.

Immigrants have access to all ordinary programs and measures at the Employment Service. In addition special offers are made to immigrants. These

are listed at the website of the Employment Service. At present the following opportunities are available:

- Re-entry Jobs [Nystartsjobb]. Employers are exempt from payroll taxes and social security contributions for the same period that the individual has been absent from working life, up to a maximum of five years.
- Work place introduction [Arbetsplatsintroduktion] which entitles employer and employee to individual support and introduction from a case worker over a period of maximum six months. The scheme is designed to compensate for lack of contacts and network.
- Entry recruitment incentive [Instegsjobb]. Entry recruitment incentives entitle the employer to economic support for a maximum of 24 months. Work must be combined with studies in Swedish (SFI).
- Practice [Praktik]. Different types of practice are offered, depending on individual qualifications. All involve economic support to the job seeker.
- Vocational training [Arbetsmarknadsutbildning], which offers the job seeker subsidized training in work areas with employment demand.
- Support to start business [Stöd för start av näringsverksamhet]. Different types of support, including professional advice and economic support is offered for individuals who plan to start their own business.

Finally, since 2010 the Employment Service is responsible for Introduction activities [Etableringsinsatser] for new adult refugee immigrants and their relatives (Arbetsförmedlingen, 2010). The Employment Service coordinates full time activities with the minimum requirement that they shall encompass SFI, preparatory work activities, and orientation concerning Swedish society. The activities also include the right to choose an introduction pilot. The Introduction activities are based on a law that came into effect in December 2010 (Svensk författningssamling, 2010:197). The law gives the Employment Service the major role in integrating refugees and their families into Swedish society and working life, coordinating the activities of other public services, such as those from local communities and health care systems.

However, not only new but all immigrants are prioritized at the Employment Service. For a number of reasons they are to a larger extent than other groups dependent upon these services. Immigrants often lack the personal network that is a key factor in finding jobs, and the Employment Service should provide such a network. Immigrants may have diplomas and merits that are difficult to assess on a Swedish market, and they could be given help to validate and if necessary complete their merits. Factors outside the job seeking process, such as housing and economy, could be problematic. Limitations and need for support due to

functional disabilities in relation to work may not have been clarified. In this case the Employment Service should offer services of different kinds, including the assistance of specialists.

The specialists at the Employment Service include occupational therapists, physiotherapists, social welfare supervisors and licensed psychologists. The occupational therapists and physiotherapists assess physical functions that may be limited in a work situation and suggest choice of suitable tasks and/or compensatory measures. In this function they can assist in suggesting and implementing suitable adaptations in the work places. The social welfare supervisors give support in social dilemmas, such as economic problems, or drug abuse, and often work in cooperation with social authorities. The psychologists assist the employment officers in assessing individual suitability for different types of work, where the individual will be described by a cognitive and emotional resource profile. Sometimes the psychologists assess specific difficulties in the cognitive or personality area, and may also suggest suitable choices of work and/or adaptations to limit the effect of the difficulties. When studies or vocational training is considered, the psychologists may conduct a more systematic assessment of strengths and limitations for a particular work and training area. The psychologists often use cognitive tests in their work. As immigrant groups have grown in size, and the focus on quick establishment on the labor market has been continuously more emphasized, questions concerning test use with individuals from other cultures and with limited experience in Swedish have grown.

2.2.1 Vocational Training

Sweden has long been known for its active labor market policy. Among the programs offered is vocational training, denoted by its Swedish acronym AMU (Arbetsförmedlingen, 2011). Richardson and van den Berg (2008) describe it as the most prestigious and most expensive active labor market program offered by the Employment Service. AMU has been described as a flagship by de Luna, Forslund, and Liljeberg (2008). According to the Government Ordinance that regulates the program (Förordning 2000:634 om arbetsmarknadspolitiska program) the purpose of AMU is to provide vocational education that facilitates for the individual to obtain or retain employment and counteracts labor shortages. AMU targets persons who are unemployed or at risk of unemployment and the purpose is to promote a better match between the output of the work force and the employers' demands for competencies. AMU had a peak enrolment during the early and mid-1990s. Since then the numbers in

AMU have successively decreased, but in year 2000 a monthly average of 30 000 individuals still took part in the programs for longer or shorter periods at a yearly cost of 5.5 billion SEK (Martinson & Lundin, 2003). By 2011 the monthly average was just below 10 000, which amounted to around 44 500 individual participants in the course of the year (Statskontoret, 2012).

Over the years AMU has been debated. It is the labor market program that has been most thoroughly evaluated (de Luna, Forslund, & Liljeberg, 2008). Some of the issues have been the effectiveness of the program in leading to employment and the displacement effects on the market. Evaluating AMU as well as other programs, it has been concluded that "the programme component of the Swedish active labour market system is at best a costly and ineffective approach" (Adda, Costa Dias, Meghir, & Sianesi, 2007, p. 50). However, this study included several programs in addition to AMU, and also covered times when it was possible to renew unemployment benefits by taking part in a program.

In the last decade the use of AMU has been more severely regulated. Since 1999 the government has set a goal that 70 percent or more of the participants should be employed 90 days after finishing a program. Martinson and Lundin (2003) investigated the effects of close cooperation with potential employers as a way to reach this goal. Richardson and van den Berg (2008) found positive effects of AMU on employment, especially the weeks after having left the program. In a paper evaluating the effect of AMU for the years 2002-2004 de Luna et al. found it to have a distinctly positive effect on employment, especially for groups with lower levels of education and immigrant groups. As part of their conclusion they added that "…it seems that AMU can be made more effective by stronger pre-screening and selection of potential candidates…" (de Luna, Forslund, & Liljeberg, 2008, p. 42). A comprehensive description of the AMU program and the enrollment process is found in their article.

Sohlman (2006) has compiled and evaluated studies that focus on the efficiency of different labor market policy measures on integration. She presents a summary where some evaluations show positive effect of AMU on future employment rates and wages, while others find that it has a negative effect compared to being unemployed. The latter find is often explained by a "lock-in" effect; the individual will be less efficient in job-seeking activities when enrolled in AMU. All evaluations that compare AMU with subsidized employment show better results for the latter. However, she points to some difficulties in the evaluations; some of the early studies do not control for the fact that characteristics differ between participants and non-participants, and the content

of AMU courses also differs considerably. Some courses have a content that is geared towards orientation and preparation. Only few outcome studies consider course content. When this aspect is included, courses with a clear vocational focus show better results than general, shorter courses. She concludes that non-Nordic citizens are underrepresented in courses with better outcomes. She also comments that the better results that are achieved by AMU after the goal of 70 percent in employment after 90 days was introduced could be caused by a stricter selection of participants. She sums up by pointing to the need for more evaluations.

2.2.2 Testing Practices

The Employment Service has a long tradition of using psychological testing for the purposes of vocational guidance. In the era after World War Two and up to 1980 these services were provided by external consultants. Major actors in this field were the Psychotechnical Institute, a subdivision of Stockholm University; the Psychotechnical Institute of Gothenburg, a subdivision of University of Gothenburg; and the Occupational Psychological Institute, a subdivision of the Council for Personnel Administration. At these institutes a great number of tests were developed. Some were based on foreign (usually Anglo-Saxon) tests, and adapted for Swedish purposes; others were developed on tests from the Swedish Enlistment Service, or originally developed tests for specific purposes. Many of the tests were designed to measure the seven Primary Mental Abilities of Thurstone (1938), but tests with a more applied purpose, such as the so called "apprentice tests" [Lärlingsprov 50], were also developed.

In the late 1970s the National Labor Board set up a committee with the purpose of coordinating the work of the institutes and creating common norms for the larger educational groups (Bergquist, personal communication, 2012-05-16). In 1987 many of the tests that were used at the Employment Service were re-standardized. At this time norms also were upgraded (Haglund, 1987).

In 1980 Labor Market Institutes [Arbetsmarknadsinstitut] were established to serve the Employment Service with vocational rehabilitation services and psychological consultation and assessments. Many of the psychologists who had worked as consultants found employment at the institutes. At this time the Council for Personnel Administration was restructured, and the documentations concerning their tests, and the license rights, were eventually sold. Psykologiförlaget, a publishing house, started making the tests from the Council for Personnel Administration more available in the 1990s. Today, however, these tests are no longer sold, and some of the documentation is no longer available.

The Labor Market Institutes were organized as a part of independent County Labor Markets Boards. There was considerable local freedom to choose methods, which resulted in some diversity in test practices and test choices over the years 1980-2008. When the Employment Service was organized on a national basis in 2008 this initiated extensive efforts in methods and practices, such as the establishment of a common manual for psychological testing.

Today the Employment Service employs over 380 licensed psychologists. Their work spans a broad area of guidance and assessment. The assessment procedures are regulated through several policy documents, such as a handbook (APU-handboken, 2009) and a manual. The handbook and manual provide guidance for test use, test interpretation, and predictions based on tests. The main theoretical orientation concerning cognitive functions in these texts is that of Thurstone (1938), and the Primary Mental Abilities. This framework has been used by psychologists at the Employment Service since it was introduced by the psychological consultants of the 1960s, and has guided selection of tests and interpretation of tests results. Cognitive job demands have been analyzed in terms of the structure of this model. With a larger influx the last decade of job seekers with functional disabilities this perspective has been supplemented with tests that focus on specific functions such as short term or working memory, cognitive flexibility, emotional status, aspects of the perception process, distractibility, etc.

In the last decade there have been projects focusing on hierarchical intelligence theory, creating hierarchical models based on the existing tests (Gagnerud & Haglund, 2005). Although these projects so far have had only marginal practical impact the present discussion is on an update of theories concerning cognitive functions, as well as theories of personality aspects, and conative functions (Järnefors, 2012).

The manual describes the general characteristics of a psychological investigation and specifies eight different types of assessment procedures with somewhat differing foci. One of eight assessment procedures is Directed Aptitude Testing [Riktad Arbetspsykologisk Utredning] (Arbetsförmedlingen, 2008). Other test based assessments are investigations of general suitability for different work areas, at times combined with analysis of specific limitations due to functional disabilities. The aim of all types of test based investigations is to find a good job match for the individual, but the focus can be more or less specific, depending on the situation of the individual job seeker.

Directed Aptitude Testing is a highly structured procedure used when a vocational training course has high demands, receive many applicants or for

other reasons need a more structured selection procedure. Directed Aptitude Testing involves creating a requirement profile for a particular training program which includes the requirements of the work area, for which the program prepares. Core competences in the profile are assessed by psychological tests, domain knowledge tests, and a structured interview (Valentin Kvist, 1992). The structure of the requirement profile and choice of actual tests is usually framed in the theoretical terms of Thurstone and his idea of cognitive functions organized as a set of Primary Mental Abilities (Thurstone, 1938). This is in line with how the psychological assessment work has been described in the policy documents (APU-handboken, 2009). It implies use of a test battery covering the different ability aspects that are important in the area, and results in an aptitude profile for each candidate.

The standardized testing procedure is applied to all or some of the applicants for a specific program. The achievement profile of an individual is matched against the requirement profile of the desired training area. The psychologists use this information as a base for an individual evaluation of suitability. Sometimes the evaluation includes suggestions for individual preparatory steps (such as rehearsing areas of mathematics), or adaptations needed in the studies (such as specific adaptations necessary for dyslectics). The purpose of the selection procedure is not to pick the most qualified, but to secure that the individual applicant possesses the necessary base qualifications. The purpose is also to assure equal opportunity regardless of sex or ethnic origin. Measures to secure optimal test performance could include choice of tests that have content that is not unduly burdened with old fashioned vocabulary or sex specific examples. Individuals with little testing experience could benefit from the inclusion of extra emphasis on instructions and tutorials. In order to avoid unnecessary difficulties for individuals from other cultures tests with excessive verbal content or excessively short time limits are only used when clearly motivated by the requirement profile.

The Directed Aptitude Testing approach has been evaluated by the psychologists involved in the procedure (Borén, 1995, 1999; Freidlitz & Willquist-Gustavsson, 1996, 1997, 2000a, 2000b, 2001; Valentin Kvist, 1995a, b; Valentin Kvist et al., 1995), with focus on prognostic properties of test results in relation to different requirement profiles. However, aspects of group equity have not been scrutinized.

2.3 Public regulation of test use in relation to minority groups

The practice of using results on cognitive tests in high stake decision situations, such as selecting students for an education or a person for a job, has prompted calls for some public regulation. The nature of these regulations differs between societies, from voluntary guidelines upheld by professional societies to detailed regulation by law. The general trend is for more thorough regulation. While test practice influences the legal or professional demands on these activities the regulations in turn influence testing practices, which include ethical and practical concerns such as in test interpretation.

Thus, frequency of test use, research in the test use area, and public regulation of test use interact. Salgado and Anderson (2002) describe how the greater bulk of research into the validity of general mental abilities tests in personnel selection and assessment is conducted in the United States, while frequency of test use for these purposes is actually higher in Europe. They suggest as a possible explanation "that fear over claims of adverse impact may be suppressing ability tests' use in the U.S. compared to Europe" (p. 82). Although they describe attempts at coordinating regulation in Europe they also state that compared to the U. S. the European legislation is far less stringent and only sporadically enforced.

In the U.S. the connection between legal demands and validation research is strong. Kane (2008), for example, states that the need for validation derives from legal, scientific, and social expectations. Including legal expectations together with scientific and social expectations has become self-evident for contemporary American scientists and practitioners within the field of psychological testing, and it has been implied (e.g. Popham, 1997) that legal demands and increased test-related litigations have been of a source of influence on Messick's (1989) insistence on including value aspects in the global validity concept.

2.3.1 A European outlook

In most European countries the regulation of test practices is exercised through professional societies, where membership is voluntary. Salgado and Anderson (2002) have surveyed the area of ability test use in 16 European countries. They found a more frequent test use in Belgium, Britain, The Netherlands, Portugal and Spain, compared to France, Germany, Greece, Ireland, and Italy. They also found a considerable national variation in standards for test use, user qualification, and test construction. They identified a pattern where individualistic cultures, mainly in north-western Europe, showed a longer history

of test use and more established standards of test use, while collectivistic cultures of southern Europe showed less regulation.

Several sources (e.g. Salgado & Anderson, 2002) highlight the work of the British Psychological Society (BPS), which has created and implemented detailed and comprehensive standards for certifying test users and reviewing tests. The work of BPS has been a model for the European Federation of Professional Psychologists' Associations, aiming towards a cross-national harmonization of testing standards. However, Salgado and Anderson conclude that it will be problematic to reach a complete harmonization, considering the historical and cultural differences between countries.

In 1978 the International Test Commission was formally established. The members are national professional psychological associations that cover North America and many European countries, but also some countries in the Middle East, South America and Africa. In 2000 International Guidelines for Test Use (Council of the International Test Commission, 2000) were published. These guidelines are now translated into a dozen languages, including Swedish. It is a 31 page document, listing many aspects of testing. However, less than one page is devoted to issues of fairness in testing individuals from different groups, such as groups differing in terms of gender, cultural background, education, ethnic origin, or age. The Guidelines advice (p. 18) that the test used should be "unbiased and appropriate for the various groups that will be tested", i.e. the advice is set in very general terms. The European Federation of Professional Psychologists Associations' Task Force on Tests and Testing has endorsed the guidelines.

In addition to the professional regulations concerning test use many countries have legislation that covers the general area of discrimination. In Great Britain, the Race Relations Act of 1976 issued a code of practice in 1984. Other anti-discrimination laws followed, and in 2010 these were all gathered under the Equality Act (2010), which concerns equal opportunities in the workplace and in wider society. Cook (2009), commenting on selection practices in Britain, concludes that these documents do not give detailed instructions, and thus have had limited impact on test use.

According to Higuera (2001) the European Court of Human Rights has set down a doctrine on the "discrimination by results". This doctrine has prompted the United Kingdom, Italy, Ireland, and the Netherlands to incorporate the concept into their legislation. However, there is no obvious practical application of this doctrine reflected in governing documents.

In 2011 an ISO Standard on Procedures and Methods to assess people in work and organizational settings was launched. More than a dozen European countries support this initiative in which the complete assessment process is covered, not just test use. However, certain topics, such as equity or fairness in test use, have little coverage. In a few lines it is stated that methods and procedures should be fair, results should be interpreted with due regard for equity issues, and consideration should be given to available evidence of the technical properties of the assessment method for the particular group. The Standards leave the user with no practical advice as to how these goals can be achieved.

2.3.2 Regulation in Sweden

In the last decade of the 20th century a number of laws with the purpose of regulating discrimination issues were introduced in Sweden. In 1991 a law was established that prohibits discrimination in working life due to sex. In 1999 three more laws were introduced; the laws prohibiting employment discrimination due to ethnic background, sexual preference, and disability. The law on measures against discrimination in working life due to ethnic group or religion makes no references to cognitive tests. The preparatory work introducing the law cites several legal cases but makes no references to cognitive tests. The issue was not discussed at the time the law was framed (Personal communication Department of Justice, March, 2012). Thus, the connection between group differences on cognitive tests and their impact on the employment opportunities of different groups has no specific legal coverage in Sweden.

As in most European countries, testing practices in Sweden are not primarily a legal concern but regulated by professional bodies, such as the Swedish Psychological Society. Through its authorization of psychologists the National Swedish Board of Health and Welfare also has a role in the test policy area. The Board awards its license to psychologists who have graduated with a master's degree in psychology and completed a year of supervised practice. These demands should guarantee competent use and interpretation of tests; however, the emphasis and time spent on education in test use has varied considerably over time and over universities, especially in the years after 1968. It has been possible for a psychologist to graduate and become licensed with only rudimentary training in the area of psychometrics and cognitive testing.

The Swedish Psychological Society presents ethical and legal standpoints at its website. Here the role of the psychologist is discussed in general terms. Questions of test use are referred to the Foundation for Applied Psychology, a

non-profit organization founded by the Psychological Society. The Foundation offers Swedish versions of the International Guidelines on Test Use, mentioned above, as well as suggestions for Test Policy in Organizations, Companies, and Authorities (Sveriges Psykologförbund, 2000). This is a fourteen page document that offers definitions and policy suggestions, but does not specifically comment test use with diverse groups. At the home page of the foundation it is stated that policy suggestions are based on the work made in the British Psychological Society and the guidelines provided by the International Test Commission. The foundation has no authority to enforce its guidelines, so the option to abide by its suggestions is voluntary. Membership in the Psychological Society is also voluntary.

A restriction in access to cognitive tests is enforced by some test publishers. There are only a handful publishing houses that provide cognitive tests in the I/O area in Sweden, and they have slightly different policies concerning the accessibility of cognitive tests. Some sell to licensed psychologists only, while some sell to users who have completed a certification. In sum, the use of cognitive tests in Sweden is regulated by professional and ethical commitment and has its focus on the individual level.

2.3.3 Regulation in the United States

The United States stands out as a country with detailed and exceptional legal demands in the area of test use, with special emphasis on consequences concerning minority groups. Since 1964, the laws enforced by the Equal Employment Opportunity Commission, especially the Title VII of the Civil Rights Act (Civil Rights Act, 1964) have had a strong impact on test use in relation to selection to employment or promotion. This act prohibits employment discrimination based on race, color, religion, sex or national origin. It prohibits the use of discriminatory employment tests and selection procedures. Restrictions are also imposed on the scoring of tests. When using employment-related tests the employer is not permitted to treat the results in any way that could relate to race, color, religion, sex, or national origin (U. S. Equal Employment Opportunity Commission, 2007). Practices that are not permitted include adjustment of scores and the use of different cut-off scores for different groups.

The law prohibits both "disparate treatment" and "disparate impact". The "disparate treatment" refers to employers using different procedures for different groups, such as subjecting only one group, and not others, to a certain procedure. The "disparate impact" refers to employers using tests or selection

procedures that have the effect of disproportionately excluding persons on the basis of race, color, religion, sex, or national origin. Tests or procedures that result in different acceptance rates for different groups can only be used when it can be shown that they are directly job-related and consistent with business necessity. Thus, despite their general predictive properties, it is not possible to use general measures of cognitive skills, if these produce adverse impact for a protected group. Tests or selection procedures of skills that are shown to be related to the particular job in question are permissible, as long as a less discriminatory alternative is not available. If such a test or procedure exists, it should be preferred.

In order to determine if a test or other selection procedure has disparate impact a statistical analysis is usually required. If the selection rate for any of the protected groups mentioned in the legal document is less than four-fifths of the rate for the group with the highest rate this will be regarded by the federal enforcement agencies as evidence of adverse impact.

The requirements of the Title VII have created the ground for a large number of litigations. In 1978, the Equal Employment Opportunity Committee adopted the Uniform Guidelines on Employment Selection in order to provide guidance for test users (Code of Federal Regulations, 1978). The purpose of the Guidelines is to help users determine if the tests or other selection procedures are used according to requirements, and to encourage use of valid procedures. The Guidelines require that selection procedures are subjected to validity studies, unless it is clear that they do not produce adverse impacts. The Guidelines indicate criterion-related validity studies, content validity studies or construct validity studies as acceptable. The guidelines include detailed requirements as to how validity studies should be conducted. In the Technical Standards, section 14, construct validity studies are singled out from the other two types of validity studies by a section on the "Appropriateness of construct validity studies." Here it is warned that construct validity is a complex strategy. It is also described as a procedure that is fairly new, with a lack of substantial literature that links the concept to employment practices. The effort to obtain empirical support for construct validity is described as extensive and arduous. Thus, the Guidelines do not explicitly support the construct validity concept, but promote the more narrow criterion related or content validity concepts.

The obligation to prove that a test procedure does not discriminate against a minority group has led to extensive research concerning measurement properties of tests as well as many attempts at reducing adverse impact. As examples of measures aimed at reducing adverse impact Cook (2009) mentions

administration on computer rather than administration on paper; video presentations; requiring test takers to construct their own answer, rather than multiple choice formats; and using items that do not require a specific vocabulary or knowledge.

2.4 Summary

Swedish public policy emphasizes speedy integration of immigrants into the work force. One tool to achieve this aim is to offer instruction in Swedish. The Employment Service has been given a prominent role in advancing immigrant employability, and offers programs such as vocational training. Psychologists at the Employment Service assess individual suitability for different job and training areas, using cognitive tests.

The public regulation of test use with immigrant or minority groups is largely a legal matter in the United States, while European societies – including Sweden – rely on professional and ethical commitment. The strict American policies have resulted in substantial efforts from the scientific society concerning questions of measurement properties and validity aspects in relation to minority groups. In Europe there is a higher frequency of test use, but less research into these questions.

3 Validity

Specific questions concerning test use with immigrant groups can constructively be viewed from a validity perspective. This chapter will present some validation perspectives, and their implications in relation to the research area will be described.

Test can be used for a multitude of purposes, among them prediction and selection. In industrial/organizational psychology an individual's results on cognitive tests are accumulated into a cognitive profile which can be used to guide the individual towards suitable choices of study and work. It can also form the basis of critical decisions, such as who will be admitted to a desirable education or be chosen for a position. "Validity refers to the soundness of those interpretations, decisions, or actions" (Moss, Girard, & Haniford, 2006, p. 109). Cronbach (1971) has stated that the purpose of test use is to reduce the number of incorrect predictions and decisions that will be regretted. Kane (2006) has a similar line of reasoning. He states that the need for validation derives from the scientific and social requirement that public claims and decisions be justified. Thus, the whole decision-making process must be validated, which entails determining the extent to which conclusions and decisions based on test results are well-founded, defensible, and legitimate. Since test use is composed of many steps, from the decision to use one or several cognitive tests, via administration and evaluation to a decision, validity cannot be summed up in a single statement and even less so in a number. The "payoff" (Cronbach, 1971, p. 448) for the proposed method must be compared to that of decisions made without test data.

In addition to the individual perspective Moss et al. (2006) point to the importance of understanding how assessment functions not only in the single case, but also as a part of complex activity systems, which evolve over time. In these systems individuals and their decisions influence the system, and vice versa. Thus they conclude that a robust validity theory must consider the situated nature of interpretations, decisions, and actions. Their suggestion to meet this complexity is to draw on the theoretical discourse from three sources; hermeneutics, sociocultural studies, and educational measurement. Each theoretical perspective illuminates some aspects of social phenomena in more detail while it leaves others in the background.

3.1 A hermeneutical perspective

In the hermeneutics approach interpretation is formed from a juxtaposition of different perspectives, and from an interpretation that focuses from detail to whole and back again in a dialectic process. It is a useful approach when trying to understand and validate a procedure that can be apprehended differently by the actors involved. This could be the case when an assessment procedure is viewed in an intercultural perspective. Different aspects of the assessment would probably be apprehended in different ways depending on the cultural perspective of the individual. Examples in this work are the implications of concepts such as "intelligence" and "time" in different cultures. Another example of a hermeneutical stance is the practical interpretation and integration of diverse test information that is performed in each individual assessment. In this process the psychologist starts with the diverse test information, forms a theory based on hypothetical conclusion that includes as much of the data as possible, tests this conclusion against all information available, and possibly revises or modifies the conclusion. In this way the diverse information on each applicant is processed into a meaningful and coherent picture. Thus, there are aspects of a hermeneutical approach that can be meaningful in understanding and working with a single case.

3.2 A sociocultural perspective

The sociocultural framework focuses on the situated nature of an assessment. It takes place within a particular activity system, a community of practice or a learning environment. In the sociocultural framework the focus is on analyzing the environment in which a procedure is conducted, and how the different parts of an activity system mutually influence one another. This approach could contribute useful information regarding questions such as how the selection of candidates based on an assessment of cognitive functions interplays with the practice and content of teaching in the vocational courses. Questions could be asked about what is taught and how it is taught. For instance, a strict selection procedure could make it possible to introduce qualified material and put higher demands on an independent study process. Such aspects of vocational training in turn could be related to the actual practice in a certain profession. Another question which this perspective highlights is the question of alternatives to the present assessment practice. In validation of an assessment procedure the alternative to be considered could be that of not assessing at all, of assessing other elements, or with a different procedure.

3.3 An educational measurement perspective

Both the hermeneutical and the sociocultural perspectives can thus bring important contributions to validation of test practices, depending on the type of questions that are being asked. However, the research questions of this thesis focus on the specific and systematic contributions and complexities that a multicultural background brings to the outcome of the assessment procedure. The procedure per se is not questioned here. Neither is the focus on individual outcomes of an assessment process. Thus a third theoretical framework, educational testing, is more apt for the questions asked.

The third validation framework is named educational measurement in the terminology of Moss et al. (2006). Educational measurement indicates a perspective that has its focus on general, law-like explanations and predictions, and evaluates intended interpretations and uses of test scores. The focus is on standardized forms of assessment where much of the rationale supporting the interpretation of test results is developed before the test is put to use, and where the standardized test is intended to be used for specified groups in order for interpretation and validity to hold. However, it is the responsibility of the test user to identify cases where this cannot be assumed and then initiate specific validation procedures. Such a case is the use of standardized Swedish tests on groups with non-Swedish background.

Over the decades, validation efforts in the educational measurement tradition have evolved from rather straight-forward calculations of the relation of the tests scores to one or several criteria (Cureton, 1951) into successively more complex endeavours aiming to capture relevant aspects. An important step was the publication of Testing Standards 1954/1955, (American Psychological Association, 1954) where four types of validity indices were distinguished; content validity, predictive and concurrent criterion-related validity, and construct validity. Each of the different types of validity investigation was related to a specific aim. These concepts were largely kept through the three first editions of the Standards.

Content validity concerns the extent to which the items or questions in the test are representative samples of the area or function that the test aims to measure. Content validity should be established deductively, by defining a universe of items and sampling systematically within this universe (Cronbach & Meehl, 1955). In content validation, the essential focus is on acceptance of the universe of content that defines the variable to be measured. The relevance of a particular behavioral domain and the representativeness of the test items are central. Given these premises the observed test performance can be taken as an

estimate of overall performance in the domain (Kane, 2006). It is a model that has frequently been applied to measures of academic achievement.

In investigating criterion-related validity the aim is to establish the connection between the tests and the external criteria they aim to measure. If the criteria are measured at a time after the test, the focus is on predictive validity. A typical example would be a test developed with the purpose to predict educational achievement. Validity is supported if there is a positive and significant correlation between test outcome and grades. If the purpose of a test is to sample and measure a criterion already present, the focus is on concurrent validity. Concurrent validity is supported if the test correlates positively and significantly with other tests measuring the same criterion. In both predictive and concurrent validity studies defining the criteria can be problematic. They should be neither too broad nor too narrow, and they should be possible to measure with a reasonable degree of precision. Criterion related validity studies are often used in admission, placement and employment testing.

In 1955 Cronbach and Meehl proposed the necessity of investigating construct validity. The aim in construct validation is to establish the extent to which a test is related to a construct or concept supposedly responsible for the performance on the test. A necessary quality of a construct defined this way is that some aspects of it are observable and thus open for scientific acceptance or adaption. Cronbach and Meehl define a construct as "some postulated attribute of people, assumed to be reflected in test performance" (p. 283). In establishing construct validity questions are posed that relate the construct to performance on the test. When attempting to establish construct validity the entire body of evidence, and assertions about this evidence, should be considered. This includes correlations with criteria, both positive (relating to the construct), and negative (delineating the construct from non-relevant aspects), postulated group-differences, correlation matrices and factor analysis, studies of change over occasions, and studies of process. All these sources of evidence are used to create a nomological network, defining the theoretical construct. Construct validation requires that substantially the same nomological net is accepted by several users. This, in turn, implies that construct validation cannot be an entirely quantitative process. As Cronbach and Meehl suggest, it resembles the general scientific procedures for developing and confirming theories.

The 1985 edition of the Standards (American Education Research Association et al., 1985) articulated a unified concept of validity that would draw on multiple types of evidence. The three traditional categories were renamed

construct-, content-, and criterion *evidence*, emphasizing that different types of evidence spanning all categories were needed in validation.

3.3.1 Messick's integrated validity model

Over time the emphasis in the validation process has shifted from specific measures and methods to a more comprehensive understanding, including aspects of test use. Messick (1989) argued a unified conception of validity as scientific inquiry into score meaning. He has proposed a definition of validity that has become widely accepted:

> Validity is an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment. (Messick, 1989, p. 13)

Including not only test interpretation, but also test use, as an aspect of validity Messick (1989) stated that these were interconnected facets of validity. Thus, he argued for validity as a unitary concept, incorporating aspects of interpretability, relevance and utility of scores. His definition expanded the validation process from considering test scores per se into considering test scores used in a social context. When test scores are used as a foundation for decisions the effect of the decisions must also be scrutinized from a validity perspective, including both value implications of scores and social consequences of their use. While questions of reliability and generalizability can be determined mainly on scientific merits, the question of validity thus has both scientific and social implications. This point is supported by Kane (2006). The reason is that validity questions tie into values, concepts and uses that have practical and ethical implications in people's lives. Messick (1989) emphasized the need for validation practice to address questions of testing consequences, not only in the single case, but also the often subtle systemic effects on institutional or societal functioning caused by recurrent testing. This is in line with the socio-cultural framework presented earlier.

Messick has presented the different validation approaches, that he names "facets", in a model (Figure 2).

	Test Interpretation	Test Use
Evidential basis	Construct validity	Construct validity + Relevance/Utility
Consequential basis	Construct validity + Value implications	Construct validity + Relevance/Utility + Value Implications + Social Consequences

Figure 2. Facets of validity according to Messick (1989).

While the evidential facets have to do with hypothetical thinking concerning how to understand the concept measured and the means to achieve a desired outcome, the consequential facets concern values and results. Although the model is presented with boxes, Messick (1989) emphasized their interconnectivity and interdependence. Validation is seen as a progressive process, where new meaning is added in each step. The different aspects are not only interlinked but overlapping and exert mutual influence. For instance, values connected to test use both derive from and contribute to the meaning of test scores. He has emphasized this interrelated quality between different components in the validity process by presenting a more complex model, with flow arrows, both single and double, that create a closed system (Messick, 1980). The gist of presenting his model of validity in this way is that no conclusion is ever final, but needs to be continually evaluated in the light of changing circumstances.

Although there was general consensus on the importance of considering value issues in connection to test use, at the time of Messick's article there was – and still is – difference of opinion as to the necessity and appropriateness to include consequential validity as an aspect in the general validity concept. An adversary was Popham (1997), who, although acknowledging the laudable intent, feared that the inclusion of social consequences in the validity concept would lead to confusion and blurring the concept. Shepard (1997) on the whole agreed with Messick, clarifying: "In a validity investigation, we don't just express a personal preference for consequences that we like or dislike. Consequences are evaluated in terms of the intended construct meaning." (p. 8). However, she also voiced some criticism since she found that Messick's presentation of the different facets of validity in a matrix leads the thoughts to the traditional fragmented concept. Moss (1992) supported the inclusion of value aspects and in addition brought forth the importance of including methods of evaluating validity aspects in performance assessment.

Kane (2008) acknowledged that Messick's broad conceptual framework provided useful structure for thinking about fundamental issues in validity theory. It was formulated within an abstract, philosophical framework and consequently won the support from philosophically oriented researchers. However, he found it less useful for the practical purpose of planning a validation effort or support in evaluating test use. There was need for more guidance into actual validation research strategies. Kane (2006) aimed at contributing a remedy by rather explicit descriptions on the procedures required to validate different areas of uncertainty. Thus he gave a structure to many of the

technical and practical problems. Concerning validation of decisions based on outcomes of tests he stated that in principle this was no different from other validation approaches. The intended interpretations and uses should be specified and evaluated by examining their coherence and the plausibility of their inferences and assumptions.

3.4 Evaluating validation models

Influenced by the work of Cronbach (1971), Messick (1989) and Kane (2008) validation has evolved from validating a test to validating an interpretation of test scores and from conducting an empirical study on the validity of a test to conducting a research program, including considering competing interpretations of test score use. This move from evaluating a specific property of a test towards a process of evaluating a test based procedure has been criticised by Borsboom, Mellenbergh, and van Heerden (2004). They claim that over time the focus has been shifted from validity, as a property of a test, to validation, as the process of evaluating that property. They find this perspective "strangely divorced" (p. 1061) from the concept that the working researcher has in mind, and propose a shift from epistemology, meaning, and correlation to ontology, reference, and causality. They argue that while the epistemological validation process, with its emphasis on correlations, can disqualify a test as invalid for the purposes it is claimed to serve it can never fully support validity. This is because validity has to do with whether the construct a test claims to measure actually has a causal influence, not whether the correlations to other measurements is of a certain amplitude. A valid test is a test that conveys the effects of variation in the attribute one intends to measure. They thus suggest that validation research must be directed at the processes that convey the effects of the measured attributes on the test scores. This, in turn, requires the formulation of a theory that is the base for hypotheses concerning the causal processes that lie between variations in the attribute and differences in test score. In a terminology borrowed from Embretson (1983), they state that validation should be concerned primarily with construct representation and only secondarily with nomothetic span. While nomothetic span refers to the efforts to establish a nomological net from relationships between test scores and theoretically related variables, the efforts regarding construct representation should concern efforts to identify the theoretical mechanisms that underlie item response.

The contentious tone of Borsboom, Mellenbergh, and van Heerden (2004) seems to indicate that Messick (1989) had lost focus of the need to examine construct validity as construct representation altogether. This is hardly a just

perspective. When considering the entire body of evidence concerning validity the focus is wider than on correlational evidence only. In a paper originally presented as a keynote address Messick explicitly stated in relation to correlational evidence:

> More illuminating are studies of expected differences over time, across groups and settings, and in response to experimental treatments and manipulations. Most illuminating of all is direct probes and modelling, e.g. via thinking aloud.(Messick, 1993, p. 11)

However, the insistence on including consequential aspects as a facet of validity may have put the other aspects of Messick's reasoning somewhat in the shade. Borsboom, Mellenbergh, and van Heerden (2004) have highlighted construct validity as a focus that must be present in all validity efforts.

Although less contentious, a turn in the direction of validation as an investigation using experimental, rather than correlational, methods is also suggested by Bornstein (2011). He states that it is impossible to validate test scores rigorously without the use of experimental procedures as part of the overall validation strategy. He thus suggests a process-focused model, where validity is conceptualized as the degree to which respondents can be shown to engage in a predictable set of psychological processes during assessment, with those processes dictated a priori by the nature of the instrument(s) used, and context in which testing takes place. This is a conception with empirical emphasis that suggests experimental methods to manipulate variables that moderate the relationship between test score and criterion. The purpose is to enable the researcher to draw more definite conclusions regarding the impact of underlying processes. By illuminating the processes that lead to differential performance Bornstein claims an increased possibility to understand test bias and test score misuse. He thus suggests that test bias should be defined as empirically demonstrable differences in the psychological processes engaged by different groups of respondents. When such differences in processing can be identified it is possible to experiment to find strategies for reducing their influence in an applied setting.

Bornstein (2011) has reviewed the methods used in practical assessments of validity. The vast majority of validity studies published in leading journals used correlational methods, often in combination with self-report outcome measures. Less than 10 percent used experimental procedures. The conclusion is that there is very slow progress in using new and more powerful methods. The use of structural equation modelling and confirmatory factor analyses could contribute to a more precise understanding of causation and underlying processes, but these were only used in around 24 percent of the studies. A process focus on

validation need not be limited to true experiments. Quasiexperiments, where groups are selected on basis of their presumed process difference, can be equally informative. Bornstein suggests that a process-focused model of validation could be helpful in establishing links to other areas of psychology, such as cognitive or developmental psychology. This, in turn, would shed light on validity questions. In sum, he calls for the use of methods that not only enhances the understanding of test score validity, but also helps integrate disparate subfields of psychology.

Despite some of the criticisms voiced above the validation approach presented by Messick (1989) has the most comprehensive approach. It covers all the aspects that questions of cognitive test use with immigrant groups raise. Questions concerning what aspects of cognitive functioning tests capture in relation to individuals with different cultural backgrounds are discussed under the facet of construct validity. Test based decisions have to do with perceived criterion-related validity which is discussed in the relevance/utility facet. Finally the outcomes of the testing procedure for the individual in terms of acceptance or rejection to vocational training are evaluated under the heading of social consequences. Value implications may influence all steps, as well as be created in each step. Although the validation procedures will rely heavily on correlational evidence the model is certainly open for other types of input also. Thus, the validation questions that have been asked in relation to cognitive test use with immigrant groups will be explored with the help of Messick's integrative model.

3.5 Summary

In the validity chapter different approaches to validity questions are described. The emphasis is on validity theory for educational measurement, which focuses on general, law-like explanations and predictions, and evaluates intended interpretations and uses of test scores. Here, Messick's integrated validity model is singled out for its usefulness and comprehensiveness. It is a model that can capture the relevant aspects when questions of cognitive test use with immigrant groups are concerned.

4 Theories of intelligence and cognitive functions

Cognitive functions are often colloquially equated with "intelligence" and tests are often presumed to measure "intelligence". A more exact definition of intelligence and a concomitant theory of intelligence is necessary both for theoretical and practical reasons. Without a theoretical basis research on intelligence is impossible to evaluate and concepts such as construct validity lack abutment. In practical work a theory is necessary in order to interpret tests results and use test outcomes as a valid base for decisions. The scientific definition of this construct has been the subject of controversy for more than a century. All theories address the question of how intelligence manifests itself in behavior but the behavioral manifestations that are considered to reflect intelligence differs considerably. They range from neural measures; such as inspection and reaction time, to complex behaviors; such as getting along in life. Another watershed is the methodological approach. Most theories of intelligence are based on statistical methods that form the psychometric approach, but other forms of methodological approaches can also be found. This chapter presents some major conceptions of intelligence in the context of a short historical overview. Some present questions will be discussed. The focus is on aspects where individuals differ in intelligence.

4.1 Broad behavior-based definitions of intelligence

The phrase "general cognitive ability" is used abundantly in the industrial and personnel literature concerning test use, when describing an individual property that has a relation to work performance. Many times it is not clearly defined. Hunter, however, has provided a definition:

> General cognitive ability in the industrial psychology literature means cognitive ability as it has been developed in adult workers or job applicants. General cognitive ability is usually measured by summing across tests of several specific aptitudes, usually verbal aptitude, quantitative aptitude, and sometimes technical aptitude. A typical measure would sum across a vocabulary test, an arithmetic reasoning test, and a test of 3-dimensional spatial patterns. (Hunter, 1986, p. 342)

Another frequently used term is "general mental ability". Salgado et al., following Schmidt (2002) define this as "any measure that combines two, three, or more specific aptitudes, or any measure that includes a variety of items measuring

specific abilities (e.g. verbal, numerical, spatial)" (2003, p. 1086). Implicitly they refer to adult individuals. Thus, whether the term "cognitive" or "mental" is used the practical measurement implication is the same. Both concepts are here defined only in relation to the measurement technique.

Wechsler's definition of intelligence emphasizes content and behavioral components rather than the measurement process. In his words intelligence is "the aggregate or global capacity of the individual to act purposefully, to think rationally and to deal effectively with the environment" (Wechsler, 1958, p.7).

Gottfredson attempted a comprehensive definition of intelligence in colloquial language:

> Intelligence is a very general mental capability that, among other things, involves the ability to reason, plan, solve problems, think abstractly, comprehend complex ideas, learn quickly and learn from experience. It is not merely book learning, a narrow academic skill, or test-taking smarts. Rather, it reflects a broader and deeper capability for comprehending our surroundings -"catching on," "making sense" of things, or "figuring out" what to do. (Gottfredson, 1997, p. 13)

As these examples illustrate the emphasis on certain aspects of cognitive behavior can differ, which may make the broad summary evaluations difficult to interpret. Still, the attempts to define intelligence broadly often focus on capturing aspects related to valued outcomes in social life, which make them interesting to consider, also from a cross-cultural perspective. The definitions and their underlying values reflect standards in the society where they evolve, and possibly influence the way intelligence is shaped in that culture. This theme is given some exposition in the section Intelligence in non-western cultures – values and patterns. However, for purposes of defining the internal structure of cognitive functions, measurement models, and prognostic properties in relation to education and work, the dominant framework is that of psychometrics.

4.2 Psychometrically based theories of intelligence

Already in early studies of intelligence it was empirically established that measures of cognitive functions showed positive manifold, i.e. were positively related. This formed the basis for use of statistical measures to determine the number of factors involved, their internal relations, and their relations to outcomes in behavior characteristics involved in studies and work. A major tool in this tradition is the use of factor analysis. Factor analytic approaches use a correlation matrix for different test outcomes as a starting point. Test relations expressed as correlations form the basis for mathematical procedures that are used to identify areas or clusters with particularly strong relations. Such clusters

are called factors and are often assumed to reflect an underlying latent variable. Although latent variables should not be confounded with concrete entities, Carroll (1997) states that it is scientifically appropriate to accept the existence and functioning of the postulated factors, when correlational data fit well to a factor-analytic model.

This implies that an individual performance in a test situation can indicate performance in a variety of situations in real-life.

The methodological advances of factor analysis and the advances in formulating theories of intelligence have had a mutual influence and they show parallel trajectories. A starting point can be found in the work of Spearman.

4.2.1 Spearman's two-factor theory

Spearman's two-factor theory of intelligence (1904) can be seen as a starting point for theories of intelligence based on factor analytic procedures. Spearman noted that the result of different measures of cognitive performance all had positive correlations. By an early version of factor analysis he could show that the correlations between tests could be explained by a general factor - "*g*" - , which influenced all tests, and a host of narrow factors, "s", that were specific for each test. While he was well aware that the general factor would differ slightly with the composition of the specific test battery he later (1923) set down general principles for the cognitive processes assumed to be responsible for intelligent reasoning and eventually the emergence of correlations in cognitive tasks. Most central in these processes were the perception of relations and the eduction of correlates; i.e. an ability to form abstract principles from the apprehension of facts, data or experiences, and an ability to apply principles of relations in new data or new situations. In the Manual for Raven's Standard Progressive Matrices the eductive component of *g* is described as

> … the ability to forge new insights, the ability to discern meaning in confusion, the ability to perceive, and the ability to identify relationships. Since perception is primarily a conceptual process, the essential feature of eductive ability is the ability to generate new, largely non-verbal, concepts which make it possible to think clearly (Raven, Raven & Court, 2000, p. SPMI).

Spearman described the general factor in energy-like terms, but was concerned that this could cause needless controversy, so later he suggested instead the term "power". This would open for the possibility of speaking of "mind power" (Deary, Lawn, & Bartholomew, 2008).

In addition to the *g* factor a host of specific factors, *s*, determined the cognitive content and demands of each task. The latter, Spearman speculated, could be more like neural organizations of the brain. These two types of factors

he arranged hierarchically, with the higher level placement of *g* indicating the wider extent of its influence. However, the level of influence did not indicate the actual strength of influence, which would depend on the complexity and content of a specific task.

Although his two-factor model was soon shown to be too simple to match the complexity in data from cognitive testing, the suggested concept of *g* and the endeavor to organize cognitive factors in a hierarchical fashion, with higher level functions having a wider influence, proved very fruitful. His work to establish his model by the help of factor analysis was also pioneering. In the British tradition great efforts were applied in using factor analysis to find a restricted set of general or broad factors that could explain wide types of cognitive functions and achievements.

4.2.2 Primary Mental Abilities

In the United States a different methodological approach developed. The "American tradition" was also based on a factor analytic approach, but instead of aiming for broad or general factors the focus was on finding a restricted but sufficiently large number of latent variables to give a meaningful explanation to the correlations between test results. Thus the analysis was conducted from bottom-up, i.e. starting with the correlations of specific tests explorative factor analysis was used to find a restricted number of factors. A leading researcher in this tradition was Thurstone (1938), who introduced the theory of Primary Mental Abilities. He developed an extension of the factor analytic procedure which allowed for multiple factors and set up the guiding principle of simple structure for his model. The idea of simple structure should ensure that the factors identified should be as independent as possible and that each test should relate chiefly to one factor, with only small loadings on other factors. In technical terms this meant rotating factors to positions where very large and very small loadings were maximized, while maintaining orthogonality of all the factors (Jensen, 1998).

Based on this method Thurstone originally proposed nine factors, which eventually were reduced to seven. He did not set up any preconceived, theoretically based guidelines for these factors, but let the content of the tests that had the highest loadings on each factor determine its description and name. His factors are Verbal Comprehension (V), which largely reflects vocabulary and other verbal skills such as reading comprehension; Word Fluency (W), which captures the ability to find the right word or right expression when needed; Reasoning (R), which indicates the ability to solve logical problems of some

complexity; Spatial ability (S), which includes two- and three-dimensional visualization and assessment; Perceptual Speed (P), which is the ability to visually scan words, numbers or figures with speed and accuracy, Numerical ability (N), which denotes the ability to work with over-learned standardized problem solving such as numbers in simple arithmetic problems; and finally a Memory (M) factor.

Thurstone (1938) claimed to find no *g* in his analysis, but his model was criticized for using a computational method which did not allow for a general factor (Spearman, 1939). Using a different computational method Thurstone (1947) could in fact later identify and acknowledge a second level general factor. Another criticism, which was implied by the establishment of a second order factor, was that the different Primary Abilities were not completely orthogonal; some of them correlated quite substantially. Cattell (1987, p. 21) stated that even though the statistician preferred the orthogonal solution, in fact it was unlikely that data, or an underlying cognitive function, should adhere to that pattern: "Being in one universe, they interact". Thurstone's model has also been challenged for not supporting the prognostic validity implied in the ability profiles (Thorndike, 1985). Despite these rather serious criticisms, the Thurstone model has had wide and long lasting impact on industrial and organizational psychology. The concept of ability profiles has been seen as fruitful in conducting analysis of the requirements of different areas of work and in matching individuals to different types of jobs.

4.2.3 The British School

Vernon (1950), working in the British tradition, championed a model with a general factor at the apex in the Spearman tradition. After extracting a general factor he named the remaining broad factors v:ed for its verbal and educational content, and k:m for a broad factor with visual, kinaesthetic and mechanical content. He assumed that the v:ed factor was largely shaped from experiences in formal schooling, while the k:m factor was shaped in activities dealing with the outside world. His model has been the base for recent attempts to find an optimal model for the structure of cognitive functions (Johnson & Bouchard, 2005).

A great influence on the development of a theory of intelligence has been exerted from Cattell's model of fluid and crystallized intelligence. Cattell (1963) suggested that Spearman's general factor were in fact two general factors, both involving problem solving, but with different characteristics. He named these G*f* and G*c*. General fluid ability – G*f* – is used in solving problems of a new and

unstructured nature, often involving the capacity to find patterns and relations in an intuitive fashion, relying on working memory for keeping different aspects in focus. Not being an explicitly taught skill G*f* was seen as more genetically determined. Cattell describes it as

> ... a single, relation-perceiving ability connected with the total, associational, neuron development of the cortex. This general power is applicable to any sensory or motor area and any process of selective retrieval from storage. Because it is not tied to any specific habits or sensory, motor, or memory area, we have called it fluid ability, g_f (Cattell, 1987, p. 138).

In line with Spearman's speculation of *g* as based on neurological qualities of the brain he describes the capacity G*f* as "based on the magnitude of a neurologically efficient cell mass, and appearing as an existing energy in any current behaviour" (Cattell, 1987, p. 142).

According to Cattell (1963) the general crystallized factor – G*c* – is a product of environmentally varying, experientially determined investments of G*f*. Thus, G*c* is formed and developed through investing G*f* in specific tasks, eventually developing what Cattell describes as "aids" or "generalized habits", which are efficient in attacking problems of a more well-known nature. Without specifying the process or mechanism he states that "Learning begets learning capacity", thereby indicating a facilitation of new learning from former learning, After it has been established, G*c* is related to problem solving using sequential reasoning and learned skills, often of a verbal nature. However, in Cattell's definition G*c* is a much wider concept than vocabulary, which is illustrated by the fact that in his experimental work it loads in seven primary factors; among them verbal, mechanical, numerical and social skills primaries. In being related to social skills, or personality factors, G*c* differs from G*f*, which has no such connection. Since G*c* reflects skills learned or developed in a certain society it is obviously influenced by the learning environment of the individual and efficient in solving problems of a known nature. How G*c* appears in a factor analysis will also partly depend on the content of the included test batteries.

Cattell (1963) further motivated the separation of the general ability into two aspects since these show different developmental trajectories. Fluid ability peaks in early adulthood, while crystallized ability continues to develop well into adulthood as long as it is applied in general use. Other separating qualities were the difference in day-to day and month to month variation (larger in G*c*) and the impact of general brain damage which was expected to be more pronounced in G*f*. Cattell pointed out that at school age there was a distinct difference in the magnitude of the standard deviation of the two factors. Since school curriculum and training would boost the pupils with the weakest performances, and put a

ceiling on the performance of the brightest, the variance in G*c* would be smaller than the variance in G*f*.

The theory of crystallized and fluid abilities was further developed and expanded in Carroll's cooperation with Horn (Horn & Cattell, 1966). Although they inspired each other's work they were not always in agreement on theoretical issues. One such aspect was the nature of a third order general problem solving factor, *g*. While Cattell was open to the reality of such a factor Horn was reluctant. He maintained that the correlation between G*c* and G*f* was not necessarily due to a single higher order g factor, but could have other statistical explanations, such as positive self-image, and/or influence from G*v* and G*s*, or sampling bias. Cattell (1987), however, noted that the second order factor that correlated most strongly with the third order *g* factor was G*f*, and proposed the Investment theory to explain this.

4.2.4 The Investment Theory

According to the Investment theory there is initially, in the individual's development, a single, relations-perceiving ability, which is based on the neurological maturity of the brain. This is the G*f* factor. As the child is subjected to different learning experiences this relations-perceiving ability will be invested in different areas, and result in skills and expertise of different kinds, e.g. G*c*. However, since the rate of learning will depend on the degree of acumen provided by G*f*, the different areas of learning will correlate. "…the child high in one manifestation will be high in another…" (Cattell, 1987, p. 139). In a factor analysis these large and positive correlations will yield a general factor. Since the general factor is caused by earlier investments of the G*f* factor, Cattell named it "historical G*f*" (p. 138).

According to the Investment theory the early investment of G*f* into different cognitive areas is the cause of the general factor. This will explain the high relationship between the general factor and G*f*. However, if the circumstances and areas where G*f* can be invested differ, such as for cultural reasons, the relation will not be uniform over cultural groups. Thus, mixed cultural testing groups should result in a relation between the general factor and G*f* that is lower than for culturally homogenous groups. When the influence of different cultural influence is separated, the high relation should appear.

By the Investment theory Cattell has provided a definition of the *g* factor, and provided an explanation of its mechanisms, which makes it possible to put the concept under scientific scrutiny.

4.2.5 The *g* factor in the work of Jensen

Jensen has spent much of a lifelong work in prolific cognitive research focusing on the *g* factor; its history, definitions, biological correlates, heritability, predictive validity, and – most controversial – expression in population or race differences. His book "The *g* factor. The Science of Mental Ability" (Jensen, 1998) gives a detailed account of his and other researchers' work on the subject. From a critical analysis of suggested definitions of intelligence he has concluded that no scientifically acceptable alternative is at hand. He acknowledges and accepts a definition of intelligence only as a general description of the processes in the nervous system that mediates an organism's (animal or human) adaptation to its environment. He does not find it useful when exploring or describing intraspecies' differences in *Homo sapiens*, i.e. individual differences in human intelligence. He offers no alternative definition, but suggests science should use factor analysis as the method to establish "certain objective realities" (Jensen, 1998, p. 49).

Jensen (1998) noted, as have many before him, that scores on mental ability tests were positively correlated. He applied a principal factor analysis by which, working top-down, he extracted a first factor that maximized the sum of the squared loadings on this factor. The mathematical solution was set up so that this first principal factor accounted for more of the total variance than did any other single factor. Thus, the model obtained by this method was based on maximizing mathematical loadings only, and free of any preconceived theoretical assumptions. From the pattern of loadings, though, Jensen concluded that the first principal factor extracted this way was the general factor, *g*. Although noting that the size and influence of *g* is underestimated in most test batteries, mainly due to restriction of range in the populations tested, he emphasized the ubiquitous influence of *g*. Under the heading "The Practical Validity of *g*" he claimed that the *g* factor "shows a more far-reaching and universal practical validity than any other coherent psychological construct yet discovered" (p. 270). It correlates not only with scholastic performance, but also job training and job performance, broad social outcomes, and biological variables.

Jensen has maintained his view that *g* cannot be described in behavioral or psychological terms. He has described it as a "property of the brain itself" (Jensen, 2008) which he proposed eventually will be understood by its physical correlates, and advances in knowledge of the physical properties of the brain. However, abstaining from a psychological definition, he still has often introduced questions of heritability of human intelligence, or group differences in measures of intelligence, based on psychometric measures, into the scientific

debate. Jensen (1986) has acknowledged that a first principal component can be slightly contaminated due to psychometric sampling, and that it thus has a disadvantage in relation to a hierarchical method. However, he sees this as a minor problem. In most cases he finds that hierarchical *g* and the first principal factor correlate "almost perfectly" (p. 308). In addition, the way Jensen apprehends a general factor might refer to "the Galtonian notion of general ability as a biological reality" (p. 302), rather than the more restricted Spearmanian psychometric conception. However, the different ways to apprehend *g* are not trivial and can be problematic in relation to the conclusions that can be supported.

4.2.6 Carroll's synthesis

By the end of the 20th century extensive research on the structure of cognitive abilities had amassed, but there was still little consensus as to one model's superiority to others. At this point Carroll (1993) contributed a uniquely comprehensive work. In his meta-analysis he evaluated research in the area for a period of more than five decades, starting with 1500 references. In more than 460 of these he reanalysed the correlation matrices, using Schmid-Leiman factor analysis. He finally arrived at a model where he organized the factors in a hierarchy with three levels. He named the levels Stratum I, II, and III. On the first level he could identify around 60 factors, where 35 could be found in at least 10 studies. On the second level the exact number of factors has varied somewhat. In the model suggested 1993 he identified seven broad factors. These were Fluid Intelligence (G*f*), Crystallized Intelligence (G*c*), General Memory and Learning (G*y*), Broad Visual Perception (G*v*), Broad Auditory Perception (G*a*), Broad Retrieval Ability (G*r*), and Broad Cognitive Speediness (G*s,* G*t,* G*p*).

The first two factors have been presented as they were suggested by Cattell (1963). They have since been subject to additional extensive theorizing and research. G*y* denotes short term learning, a memory factor that Carroll (1993) suggested was in need of further research. G*v*, a second order factor denoting the capacity to identify visual patterns, and manipulate two- and three-dimensional figures of some complexity was first identified as a primary (spatial ability) by Thurstone (1938), but was expanded into a second order factor in Horn's and Cattell's (1966) research. G*a*, the auditory factor, is yet another factor that according to Carroll needs more research to be firmly established. It denotes the capacity to discriminate auditory stimuli and control the perception of auditory input. The G*r* factor denotes the capacity to produce ideas, concepts

and connotations from memory in a rapid flow. It has a predecessor in Horn's suggestion of a factor of general fluency (1966). The cognitive speediness factor is composed from several narrower aspects; speed and accuracy in automated work (G*s*), reaction time (G*t*) and psychomotor speed (G*p*). This factor too was suggested by Horn.

On the third level, Stratum III, Carroll tentatively placed a general factor in accordance with the British tradition. Although he abstained from making a definite conclusion, he described the evidence for some kind of general factor as overwhelming (Carroll, 1997). However, he also acknowledged that the nature of the general factor could vary somewhat, depending on the sample of test used and the representativeness of the test population. Thus, more research is needed to establish the characteristics of a truly general factor.

The model created on the basis of Carroll's research has become known as the CHC model (McGrew, 2005) from the initials of the main contributors. It synthesizes many of the elements in earlier research into a comprehensive structure. At Stratum I many of Thurstone's Primary Mental Abilities can be identified. At Stratum II the nature of Vernon's v:ed and k:m factors seems to agree rather well with the G*c* and G*v* factors, although Vernon's factors focus more on content, while the Cattell and Horn factors of G*c* and G*v* involve process and familiarity to a greater extent. Recent research on some of the second level factors (Ackerman & Lohman, 2006) has contributed to their definition and their relation to Stratum I variables. When estimating the predictive properties concerning the learning process in adults they found it important to keep the G*c* factor separate from G*kn* (General domain specific knowledge) and G*rw* (Declarative and procedural reading and writing skills and knowledge). Also, it seems pressing to define and research the G*v* factor and the importance of visuo-spatial abilities as predictors of motivation and learning in the technical area. The Visual Processing Factor (G*v*) has been the object of "a resurgence of interest" (p. 140). It has been shown that this factor contributes to the prediction of educational and vocational preferences and performances of academically gifted students independently of other factors. There has also been continuing research on the relation between different primary spatial factors and different cognitive processes.

Carroll (1997) never claimed his model to be a final solution, but saw it as a map to guide further research on all three strata. There has been wide consensus in supporting the CHC model; however, the consensus is not unanimous. Critics of the main stream definition are e.g. Johnson and Bouchard (2005). They have conducted a detailed comparison between the Cattell-Horn G*f*-G*c* model,

Vernon's v:ed and k:m model and Carroll's model by reanalyzing a large dataset based on the three models. They claim that the Vernon model has the best fit of the three, but go a step further and transform his model in several steps in order to find a statistically optimal definition of the second order factors. It can be concluded that much research is still needed around the second level broad factors.

The *g* factor

For theoretical reasons, but also considering the ubiquitous influence of the *g* factor; its relation to educational, professional, social and biological outcomes, controversies around the existence and nature of the *g* factor are perhaps even more in the focus of scientific interest.

Jensen (1986) claims that evidence such as biological correlates supports the conclusion that *g* is not a methodological artefact. However, it is not yet possible to scientifically establish its causal nature. A line of research that has been fruitful has concentrated on the properties of G*f* in relation to *g*. The description of the G*f* factor in the Cattell-Horn model is in essential aspects similar to that of Spearman's *g*. This is also in line with Cattell's Investment theory (Cattell, 1987). If *g* can be shown to be identical to G*f* the third order factor will be invariably defined. Several studies have supported this identity of *g* and G*f* (Gustafsson 1984, 1988, 1994, 2002; Keith 2005; Reinolds & Keith 2007; Undheim, 1981; Undheim & Gustafsson, 1987). However, other studies have failed to do so.

The neuropsychological perspective, which Jensen has used as a support for *g*, has also been used to question the identity of *g* and G*f*. Blair (2006), using a neuropsychological perspective, but also reflecting on the general rise in test scores especially in the fluid reasoning realm ("the Flynn effect"), claims a differentiation between fluid skills and general intelligence. Starting with the find of Gustafsson (1988) of the unity between *g* and G*f* he argues, somewhat surprisingly, for studies examining a fluid factor purged of the *g* factor variance. He acknowledges that defining such a residual factor in the light of the "near unity" of *g* and G*f* would make such an endeavour "problematic" (p. 122), but argues that some aspects of working memory and executive function could be found. The need for a clear understanding of the "g" factor is illustrated by theoretical quandaries, such as these.

4.3 Brunswik symmetry

Although the emphasis in recent research on higher order factors such as *g*, G*f*, and G*c* as predictors of learning and work outcomes is abundant and

considerable criticism has been raised against Thurstone's theory of Primary Mental Abilities, there has been a somewhat surprising fidelity in Swedish industrial and organizational practice to the Thurstonian model. This is apparent in the documents that support the psychological practice at the Employment Service. There is also evidence that this is a general approach in the area of assessment and selection outside the Employment Service. A recent publication (Aronsson, Hellgren, Isaksson, Sverke, & Torbiörn, 2012) states that models based on the Primary Mental Abilities prevail over hierarchical models or use of the general factor of intelligence in large parts of Europe, and the Nordic countries.

The apparent controversy between the research based consensus concerning the important predictive properties of *g* and the continued use of Thurstone's Primary Mental Abilities or other specific measures in the applied area can possibly be resolved by the application of a Brunswik Symmetry perspective, as proposed by Wittman and Süß (1999). They applied the Gestalt principles of symmetry to create models of prediction, where area and level of outcome were matched with predictors of the same level of generality and with the corresponding content area. Since measures of cognitive capacity were hierarchically organized the outcome measures also had to be hierarchically structured, and the different levels of generality matched. Applying this principle they found increased correlations.

There are thus implications that a detailed and correct work requirement profile can be successfully matched with equally detailed test measures. It can be presumed that evaluation of individual abilities in a selection or recruitment situation often are geared towards rather specific areas of work, and that narrow tests that match the content of those areas could hold good predictive properties. The presumption, of course, is that both requirements and a matching selection of tests must be correctly defined in content and in level of generalization, for the predictive match to occur. In order to be able to apply the Brunswik Symmetry Principle systematically it is necessary to find clear definitions of all cognitive factors; for the general factor as well as factors at Stratum II and I in the hierarchical model.

4.4 Summary

Theoretical approaches concerning human cognitive capacities vary in approach and focus of interest. Theories of variations in human cognitive performance have, to a large part, developed in a psychometric perspective, where correlational methods form the basis for factor analyses and model building.

After around a century of psychometrically oriented research a hierarchical model of intelligence, with three strata of increasing generality, has achieved reasonable consensus. All strata in this model need additional research, but especially challenging are questions concerning the general factor, *g*. Since the general factor of intelligence is the focus not only on research on prognostic properties, but also on causes for group differences, and on causes for change over time, it needs to be unambiguously defined.

In industrial and organizational psychology models with lower order factors have often been used, despite lack of scientific support for criterion-related validity. It is suggested that this apparent paradox can be resolved, if predictive variables, such as test outcomes, are systematically matched to outcome variables in working life that are defined on the same level of generality.

5 Learning, intelligence, and testing outcomes in non-Western cultures

The psychometrically oriented contributors to the theoretical models of intelligence included culture as a source of influence on intelligence and intelligence test results. However, they mainly restricted their understanding of culture to the variations in social and home environment, scholarly instruction, and other sources of influence within their given society. A broader perspective on culture and the influence it has on cognitive functions was introduced with scholars in the last decades of the 20th century. It can be interesting to give some attention to these, since test scores for groups with diverse cultural backgrounds without this perspective tell us less about the underlying processes.

Group level outcome of test scores could reflect how well the practice of testing fits into the particular cultural understanding; it could reflect differences in level or structure in the processes that are being measured, or a combination of these. The purpose of this section is not to give a complete view of intelligence in diverse cultures – this would be a perspective far too wide for this thesis. But hopefully some short glimpses of the effect different cultural contexts can have on cognitive functions and eventually their influence on outcomes of testing practices can contribute to a somewhat deeper understanding. Thus a short digression from the psychometric theme is introduced here.

5.1 Frameworks and circumstances in the test situation

Addressing the framework of standardized testing, Nell (2000) has written about the problems and possibilities of conducting neuropsychological assessments with non-English speaking groups of very diverse backgrounds in South Africa. He speaks of "educating the executive" (p. 232-234) as a necessary first step. By this he refers to a number of explanations and practice sessions that serve to bridge the gap between the expectations and mind-frames of the subject exposed to a standardized testing session and those underlying the test situation. Many aspects that are self-explanatory for a person brought up and educated in the Western world, and hardly given a conscious thought in the test situation, need to be presented and even practiced. One such aspect is the understanding of and attitude to time. Standardized tests often have a time limit, and even when they do not there is an expectation from both the test administrator and the test taker

that he or she will work efficiently, i.e. at a fast pace. Thus, Nell (2000) introduces the concept of "working fast and accurate". This is a contradiction in terms in many cultures were a considerable amount of time must be spent on a task in order to indicate a serious effort. Experiencing time as a circumstance that has a limit and can be used up is alien to for instance many African cultures.

Kapuściński (2003), a Polish author and journalist with extensive experience of post-colonial Africa, has described and contrasted the European and African comprehension of time. He states that the European experiences time as objective and linear, and subjugates to its demands. In contrast, the African conception of time is subjective, elastic and open. Time is something shaped by man, because its existence is dependent upon occurrences initiated by man. Time appears as a result of man's actions, and if no action is taken time disappears. It is a substance which exists by the energy of man.

While this extreme difference in conception of time most likely has become more rare with the world wide introduction of Western technology and increased communication and education, the difficulty of introducing the concept "fast and accurate" into a testing situation is probably familiar to all who have worked with tests in a cross-cultural context. Explanation of expectations and practice can possibly bridge some of the gap, although deeply rooted attitudes are not easily influenced. Test results, in such a situation, may have good prognostic properties, since emphasis on speed in the test situation may be similar to demands in work. However, test results may not reflect maximum performance of the individual.

5.2 Contrasting the Occident and the Orient

Investigating cognitive functions in a cross-cultural perspective Nisbett, Choi, Peng, and Norenzayan (2001) started by contrasting Western views, as represented in ancient Greek society, with East Asian thoughts, as represented in ancient Chinese society. These cultures were chosen for their geographical distance, which makes it unlikely that they had any extensive influence on each other, and for their substantial influence on the modern world, as represented by the European and post-Columbian American civilization, and that of East Asia. Initially they state that these two cultures represented markedly different social systems. The self-sustained Greek landowner is contrasted with the Chinese villager, who is embedded in reciprocal social obligations. While the ancient Greek was socialized into a strong sense of personal freedom and agency the Chinese put emphasis on the individual as a member of a close knit collectivity, with obligations to maintain the harmony in that group. On basis of such

descriptions they argue that members of the different cultures are socialized from birth into different world views and habits of thought. Social organization directs focus to some aspects of the field of perception at the expense of others. What is attended to influences metaphysics i.e. beliefs about the nature of the world and about causality. Social practices can influence the plausibility of metaphysical assumptions, such as whether causality should be regarded as residing in the field or in the object. Metaphysics thus guide tacit ontology and epistemology, i.e. beliefs about knowledge and how it can be obtained. Eventually this affects the development of cognitive functions. They mention as an example the different emphasis on dialectical versus logical cognitive processes.

Nisbett, Choi, Peng, and Norenzayan (2001) hypothesized that patterns formed in the different cultures are still strong in present society. From this they derived a number of experimental hypotheses that they tested on subjects from the different societies. They found that cultural background influenced the perceptual process and cognitive processing along with several other aspects of cognitive functioning. They found East Asians to be holistic, attending to the entire field and assigning causality to it, making relatively little use of categories and formal logic, and relying on "dialectical" reasoning. Westerners, in contrast, they found more analytic. Western perceptions of reality pay attention primarily to the object and the categories to which it belongs. The perceptual process of individuals with an East Asian background was more strongly focused on wholes and relations, while that of individuals with Western background was focused on parts. In addition, Easterners showed stronger field dependence. In the cognitive processing the Easterners were showed to be more oriented to perceive relationships and similarities, relying on experiential knowledge, while Westerners used formal logic, applied rules and formed categories. Another finding was a stronger tolerance for contradiction in the Eastern group. In sum both cognitive content and cognitive processing showed considerable cultural influence. Their results thus indicated that basic cognitive processes can be shaped very differently in different cultures, both in perception, processing and output. Although single individuals can change in some aspects when exposed to a different culture or social system, these processes are not easily influenced.

Maruyama (1999) has put forth similar points, describing the epistemological types of different societies and times. He has shown how the mind-frame guides what kind of information that is attended to, what kind of principles that guide both cognitive and aesthetic processing, and also what kind of questions it is

possible to ask, given a certain epistemological type. However, he has not supported his finds with standardized testing.

5.3 Diverse values and standards

Sternberg and Kaufman (1998) have explored how aspects of intelligence are related to valued outcomes in social life. These values reflect standards in the society where they evolve, and possibly also the way intelligence is shaped in that culture. Sternberg and Kaufman point out that the extent to which parents and teachers shared the teachers' conception of intelligence perfectly predicted the rank order of the child's performance. Looking at definitions of intelligence in a cross-cultural perspective, they found that African definitions of intelligence in addition to cognitive skills emphasize skills that help facilitate and maintain harmonious and stable intergroup relations. Western values such as speed of processing, independence in problem solving, and self-assertion were not valued, and sometimes looked upon with suspicion. Many Asian cultures, as well as Latino cultures, also emphasize social aspects of intelligence.

Sternberg and Kaufman (1998) propose, after discussing the many different ways to define and understand intelligence in different cultural contexts, that cultures designate as "intelligent" the cognitive, social and behavioral attributes that they value as adaptive to the requirements of living in those cultures. However, they add that although cognitive skills have different manifestations across cultures, there is probably a common core of cognitive skills that underlies intelligence in all cultures. Not surprisingly, Sternberg (2004) has stated that it is not possible to understand intelligence outside its cultural context. He points to the usefulness of his theory of successful intelligence, which specifies the information-processing components as universal (the common core), while the specific culture defines the context in which these components are enacted.

Sternberg (2004) has proposed a taxonomy of four basic models describing key aspects concerning the design of instruments used to measure intelligence versus dimensions of mental processing, and how they are influenced by culture. In Model I, both the nature of intelligence and measurement instruments (apart from appropriate translations of texts) are considered to be the same across cultures. He gives as an example Jensen (1998) who works from the position that general intelligence is the same (with variations in levels) across time and place. Model II represents a difference in the nature of intelligence but no difference in the instruments used to measure it. The outcomes obtained from using the same measure in different cultures are structurally different as a function of the culture being investigated. Sternberg mentions the work of Nisbett as an example. Work

of Nisbett, Choi, Peng, and Norenzayan (2001) is presented above. In Model III, the dimensions of intelligence are the same, but the instruments of measurement are not. This position implies that measurement tools (tests) must be developed in the context of a specific culture in order for it to be meaningful for the test taker, and thus produce meaningful results. If such an instrument is used in a different culture the meaning of the score will differ. This is the position taken by Sternberg (2004), who has developed assessment procedures in a number of diverse cultures, but claim that they measure the same basic cognitive processes. Model IV, finally, represents the case where both the instruments and the ensuing dimensions of intelligence are seen as different as a function of culture. In this position intelligence can be understood and measured only as an indigenous construct and nothing can be generalized to other cultures. Examples of this position given by Sternberg are works based on a radical cultural-relativist position.

The model Sternberg (2004) presents seems clear; however, offering only two categories for each dimension makes simplification necessary. For instance, in order for Sternberg to identify his work with Model III it is necessary to reduce dimensions of intelligence to his three categories only. If dimensions of intelligence are presented with greater complexity, including aspects of content and process as well as aspects of generality, it seems difficult to claim that these dimensions can be completely universal, and only differ in the way they are measured. Complexity in dimensions of intelligence and how they are formed in a cultural context makes the position of Model I equally problematic. The position of Model IV makes cross-cultural research into measurement of intelligence impossible as a scientific project. Thus, for cross-cultural research the given position seems to be that of Model II.

5.4 Including personality, spirituality, and special skills

Gardner (1983), who is critical of the traditional definitions of intelligence and psychometrically based methods, has offered ways to look at intelligence that are based on an understanding from neurology, evolutionary biology, and anthropology. He describes in some detail how different cultures value, educate and nurture different aspects of intelligence. He highlights three main forms of education; acquirement of special skills in non-literate cultures, the focus on reading skills in traditional religious schools, and the learning of scientific principles in modern secular education. The societal and individual goals of these diverse training practices obviously differ. Like the perspective taken on intelligence and intelligent behavior in many non-Western cultures, Gardner

includes aspects of inter- and intrapersonal talents into his concept. He has also suggested (Gardner, 1998) the inclusion of spiritual intelligence.

Spiritual understanding is a central factor in an Islamic concept of education, as described by Halstead (2004). In an orthodox philosophical tradition this implies that rational faculties can only reach certainty with the help of revealed knowledge and spiritual understanding. Thus, rationality is valid only within the boundaries defined by religion. Halstead states that the Islamic word for education translates into three meanings which include emphasis on knowledge, growth to maturity and the development of good manners. The practice of education in an Islamic context implies that the Western focus on critical and independent thinking, which is a core value in intelligent behavior, is not necessarily encouraged. Also, knowledge in itself is not valued, but becomes valuable only when it is put to proper use in supporting the purposes of God's creation. However, skills that reflect and express a socially responsible and mature attitude are valued. This could include skills that integrate faith and daily life, such as being able to cite and use applicable parts of the Quran. Halstead's description of an Islamic concept of education could be typical for the type of learning that is encouraged in traditional religious education.

Gardner (1983) brings an example illustrating acquirement of special skills in non-literate cultures. He describes the development of special visuo-spatial skills with the Pulawat people of the Pacific. The Pulawat navigate the open sea over great distances. Being able to sail a canoe with speed and accuracy from an island to another without the guide of any of modern society's technical devices is important to the Pulawat society and consequently navigational skills are held in high esteem. Chosen Pulawat children are systematically trained from an early age. The training includes a great amount of memorizing, such as the positions and trajectories of stars, but also skills and techniques to interpret information from waves, clouds, and winds. Training of this or similar kind is not uncommon in non-literate cultures and could be expected to contribute to development of spatial skills to a level that is rare or non-existing in Western culture.

5.5 Neuropsychology and the effects of alphabetization

A European example (Petersson, Reis & Ingvar, 2001) shows how the cultural circumstances can influence not only cognitive processing, but also the functional organization of the brain. The study investigates cognitive processing in literate and illiterate subjects in Portugal, who for specific socio-cultural reasons had not had the opportunity to acquire basic reading and writing skills.

They found that learning an alphabetic written language modulated the audio-verbal language system in a non-trivial way, and thus support for the hypothesis that the functional architecture of the brain was modulated by literacy. They also found that the effects were not limited to verbal skills, but found that formal schooling also influenced the 2D and 3D visual naming skills.

Ardila et al. (2010), summing up a literature review also state that learning to read impacts many spheres of cognitive functioning. It is thus no surprise that illiterate people underachieve on many cognitive tests.

5.6 Summary

The chapter brings some snap-shots of how questions of intelligence, learning, and testing may be perceived in non-Western cultures. As a result of the differing values and different circumstances in life conditions cognitive processes are nurtured and developed in different ways. The immigrant groups in Sweden can be expected to represent many of these varieties. In the individual case it is important to make an effort to understand the strengths and limitations of each individual, given each person's background. However, it is practically impossible to give a standardized assessment in relation to each individual's sending culture. This would probably also be less useful, when the purpose of the testing is to make predictions. Since the purpose of the testing is to evaluate resources for studies and work in Swedish society, the meaningful perspective must be that of the opportunities and demands of the receiving culture. Working within standardized assessments the psychometrically based Western theories thus form the framework for interpretation and evaluation.

6 Group differences in cognitive functions; patterns, causes, and consequences

In this chapter questions concerning group difference in cognitive functions will be reviewed. A descriptive overview will introduce the theme. Possible sources of group differences will be discussed along with measures to diminish their effect, such as intervention studies, and technical measurement analyses.

6.1 Studies in U.S.A and Europe

Research in the area of group differences in intelligence can be said to have originated and developed in the United States. Salgado and Anderson (2002, p. 85) state "that there has been far less research into subgroup differences and adverse impact on all types of tests, including GMA tests, in Europe compared with the United States." A century ago the focus was on new arrivals, mostly from Southern and Eastern Europe, seeking a new life in America. Goddard, who conducted testing with immigrant groups at Ellis Island, concluded that "One can hardly escape the conviction that the intelligence of the average 'third class'[2] immigrant is low, perhaps of moron grade" (Goddard, 1917, p. 243). Brigham, who had served as a military psychologist during World War One, voiced a similar opinion:" Immigration should not only be restrictive but highly selective…. If all immigration were stopped now, the decline of American intelligence would still be inevitable" (Brigham, 1923, p. 210). Tenopyr (1996) indicates that Brigham´s work has been claimed to have led to the restrictive American immigration laws of the 1920s. Working from the accepted knowledge of the time Goddard and Brigham were confident in their views, even if Goddard later changed his position. In the century that has passed a wealth of knowledge has been amassed, most of it disclaiming their fears. These examples could serve to show how social policies and test practices often are intertwined. They also illustrate how the political climate of an era can influence areas of study and possibly the conclusions arrived at, which points to the importance of Messick's (1989) stance; to include social consequences of testing in validation efforts.

[2] Only steerage passengers went through the medical examination at Ellis Island. First and second class passengers were examined on board.

Another focus in early research was on the difference between Americans of African descent and Americans of European descent. Spearman (1932) cites some early American studies investigating this difference. Starting in the 1960s the Civil Rights Movement gave national attention to "the role of measurement as a gate keeper to the rewards of American society" (Tenopyr, 1996, p. 348). The practice of basing selection into work and higher education on test outcomes resulted in lower representation of minority groups. Much of the American debate has concerned the opportunities of the substantial African American minority, and the measurement qualities in relation to this group. In later decades also the growing Hispanic group has become the center of attention (Schmidt, Pearlman & Hunter, 1980). Comprehensive research has reached consensus that there is a group mean difference in general mental ability of about one standard deviation difference for African Americans in relation to the White majority and a somewhat smaller difference for Hispanics (e.g. Carneiro, Heckman, & Masterov, 2005; Herrnstein & Murray, 1994; Roth, Bevier, Bobko, Switzer III and Tyler, 2001; Rushton & Jensen, 2005; Schmidt & Hunter, 1998).

During the last decades research interest has broadened towards groups in Africa and Asia. Lynn and Vahanen (2002) provide "National IQ" scores for 81 nations. Rushton and Skuy have investigated Black, Indian, and White groups in South Africa (Rushton & Skuy, 2000). Guenole, Englert and Taylor (2003) found lower mean scores for Maori, compared to European applicants, on measures of verbal reasoning and numerical business analysis, but not on general numeric reasoning.

In Europe lower mean scores have been found for Surinamese, Netherland Antilles, Turkish and North African immigrant groups in the Netherlands (te Nijenhuis & van der Flier, 1997, 1999, 2000, 2003; te Nijenhuis , Jong, Evers, & van der Flier, 2004). Thus, differences in group mean scores between majority and minority groups seem to be the rule. In most cases the minority groups score lower than the majority. An exception for the lower means of minority group is the Asian American group in the United States.

6.2 Heredity, environment, and their interaction

The causal mechanisms behind achievement differences for different groups have also been subject of research and debate for more than a century (Rushton & Jensen, 2005). The scientific debate has at times been quite heated (see e.g. Neisser et al., 1996), which may not be surprising considering the social and practical implications of the different views.

Much research has focused on the so called Spearman hypothesis[3], which states that group differences in performance on cognitive tests are a function of the *g*-loading of the tests (e.g. Jensen 1998, 2000; Lynn & Owen, 1994; Rushton, 2004; Rushton & Jensen 2003, 2005; Rushton, Skuy & Bons, 2004; te Nijenhuis & van der Flier 1997, 2003). When it can be established that differences are caused by the general factor it follows, that all areas that are known to be influenced by *g* also will show group differences. However, often the inference is also made that these differences are more or less inevitable, and related to race (e.g. Gottfredson 1997, 2000, 2005; Herrnstein & Murray 1994; Rushton, 1998; Jensen 1998, 2000). The logic is based on the substantial heritability that has been shown for general intelligence. At the individual level the amount of influence from hereditary sources has been shown to be quite considerable, especially in middle and late adulthood (Plomin, Pedersen Lichtenstein, & McClearn, 1994). In the next step it is assumed, or implied, that group differences have similar hereditary basis and are more or less resistant to environmental influences. In their comprehensive article Rushton and Jensen (2005) have gathered evidence from thirty years of research which focus on "race differences" in cognitive ability. They cite a great number of studies on subjects with great variation; from transnational comparisons and transracial adoptions to biological markers such as brain size, and reaction time. Overall they find conclusive support for a hereditarian view. In defining the hereditarian view they have suggested a number of categories: Initially they define it as assigning 50 percent of the causes for group differences to genetic origins, and 50 percent to environmental factors. They have also proposed a dichotomy between "the *culture-only* (0 percent genetic–100 percent environmental) and the *hereditarian* (50 percent genetic–50 percent environmental) models of the causes of mean Black–White differences" (Rushton & Jensen, 2005, p. 235). However, as they sum up their conclusions, they suggest that the model may need to be revised to "perhaps to 80 percent genes-20 percent environment" (Rushton & Jensen, 2005, p. 273, also p. 279). Thus, they emphasize the hereditarian position.

The arbitrary and either-or method in the definition of the culture versus the genetic position has been pointed out as a problem (Suzuki & Aronson, 2005; Sternberg, 2005). While most research professes that both heredity and environment play important roles the assignment of percent numbers constitutes

[3] Rushton (1998) proposed that when a positive correlation occurs between g-loadedness and variable X, the result be termed a "Jensen effect. The use of the term "Spearman's hypothesis" may be restricted to research in the United States with Black and White groups and can be seen as a special case of the general Jensen effect.

a misleading dichotomy. It ignores the interaction found between genetic factors, anatomical structures, culture, and environment. Thus, when referring to neurological and physiological studies Rushton and Jensen do not take into account the approximate only relationship between these measures and the psychological constructs, such as intelligence.

In response to the Rushton and Jensen (2005) article Nisbett (2005) accuses the authors of ignoring or misinterpreting most of the evidence of greatest relevance to the question of heritability of the Black–White IQ gap. Certainly many questions may be raised. There is no substantial scientific support for the view that group differences are mainly based on hereditary mechanisms, although a hereditary influence on individual differences is generally accepted. The amount of influence attributed to hereditary mechanisms depends in part on the measures of environmental influence that are available. On the individual level this can be estimated, for instance through twin studies. In group studies such measures are more elusive. Suzuki and Aronson (2005) point out that most of the research has been performed on the individual level.

There is even less support for the view that areas with considerable influence from hereditary sources are resistant to environmental influence. Neisser et al. state point-blank: "A common error is to assume that because something is heritable it is necessarily unchangeable. This is wrong. Heritability does not imply immutability" (Neisser et al., 1996, p. 86). A recent article by Nisbett, Aronson, Blair, Dickens, Flynn, Halpern, and Turkheimer (2012) sums up some of the present positions on intelligence and heritability. They report the established fact that heritability varies with the age of the individual. They also report several studies that show considerable variation in heritability between social groups. A tentative conclusion is that the heritability estimates of cognitive ability are attenuated among impoverished children and young adults in the United States. The results were more mixed in European studies, and in studies involving adults. They offer as their interpretation of the findings that children in poverty do not get to develop their full genetic potential. While a poor environment will set limits to the possible variation an optimal environment will produce variations caused mainly by hereditary mechanisms. This would imply that children in lower socioeconomic circumstances could benefit from interventions. The rise in IQ scores for children adopted into a higher socioeconomic group is one such example. They also point out that individuals from low socioeconomic groups often are underrepresented in studies of heritability. That may have caused the overall estimates for heritability to be inflated.

Yet another point made by Nisbett et al. (2012) is the effect of education on intelligence. The substantial effects of schooling on intellectual performance have been shown in a number of studies, such as Cliffordson and Gustafsson (2008). Using enlistment scores they could show that performance was increased by schooling, and that the pattern of differential schooling curriculum matched effects on specific tests. Nisbett et al. point out that the seasonal drop in IQ over summer holidays is much greater for children from lower socioeconomic groups, possibly because children in better circumstances receive training of different kinds over the summer. Considering the variation in educational opportunities not only between different socioeconomic groups, but also due to cultural reasons, the possibility that environmental influences are considerable sources of group differences must be considered.

6.2.1 Group differences over time - The "Flynn Effect"

The Flynn effect refers to the considerable and steady rise in cognitive test scores, especially measures of fluid intelligence, which has been noted since systematic testing began. The name derives from James Flynn, who has described the phenomenon systematically (Flynn 1984, 1987). The average gain is about three IQ points per decade. The Flynn effect does not seem to be limited to the industrialized world. Nisbett et al. (2012) cite sources that indicate extremely high rates of gain in countries that have recently begun to modernize. Results for some countries that began to modernize more than a century ago, such as the Scandinavian nations, may be reaching asymptotic levels (e.g. Emanuelsson, Reuterberg & Svensson, 1993).

The causes for the large increases are difficult to understand. Hereditary mechanisms cannot operate on such short terms. Nisbett et al. (2012) state that it is easier to eliminate causes than to provide a convincing causal scenario. Increased quality and availability of schooling has been suggested as a cause, but the school related areas of verbal and numerical abilities are much less affected than changes on tests that were designed to be culture free, such as Raven's Standard Progressive Matrices. Lynn (1990) has suggested better nutrition as a cause for the increase. However, this would be expected to influence lower classes more than higher, and show a similar influence on physical measures such as length. According to Nisbett et al. the outcomes are not consistent in this respect. They conclude that the changes are related to the industrial revolution and its cognitive demands, "that modern societies somehow rose to meet" (Nisbett et al., 2012, p. 141). This reasoning is in line with a model presented by Dickens and Flynn (2001). The authors do not challenge the

traditional heritability estimates, but proceed to explain the massive amount of influence from the environment through a reciprocal causation between phenotypic IQ and environment. They suggest this is an iterative process, which causes a multiplier effect. While the model was developed to accommodate the finds of large gains in IQ measures over generations, they claim that the same model could be applied to group differences such as the Black-White difference in the United States. In fact, they claim that "any trait that has a tendency to match itself to an environment that reinforces that trait will behave in the fashion our model describes" (Dickens & Flynn, 2001, p. 366), as long as the effects of environment do not accumulate over time.

6.2.2 Are the differences mainly in *g* or *c*?

In main stream research the focus has been on the so called Spearman hypothesis, which states that group differences in performance on cognitive tests are a function of the *g*-loading of the tests (e.g. Jensen 1998, 2000; Lynn & Owen, 1994; Rushton, 2004; Rushton & Jensen 2003, 2005; Rushton, Skuy & Bons, 2004, te Nijenhuis & van der Flier 1997, 2003). Te Nijenhuis and van der Flier (1999) specifically criticized methods that used outdated intelligence taxonomies, such as Guilford's Structure of Intellect model (Guilford & Hoepfner, 1971) and Thurstone's (1938) Primary Abilities, which they claimed cause bias in the analyses. Instead they emphasized the importance of using hierarchical models. However, a hierarchical model typically includes latent variables at different levels of generality (not just *g*) and unless a hierarchical model is correctly specified it too can bring biased results. Based on studies of the Spearman hypothesis (Jensen, 1998) Rushton and Jensen (2005) argue that group differences largely can be understood as genetically caused differences in *g*. They base their claims on twin, adoption and family studies, where the *g* component of test measures is extracted as a principal component. Group differences are then computed by the method of correlated vectors. However, the principal component method produces a factor different from that produced using confirmatory factor analysis, and thus requires a different interpretation. Suzuki and Aronson (2005) raised the question concerning the difference between psychometric test measures of IQ, or *g* (general intelligence). They pointed out that Rushton and Jensen (2005) discuss *g* as a unitary construct, despite the fact that standardized IQ tests have different loadings on *g*. Ashton and Lee (2005) have shown that a larger proportion of the systematic variance in the G*c*-tests is turned into common variance. This effect causes a tendency for the first principal factor, or a measure of IQ, to be biased in favor of G*c*-tests,

i.e. verbally and culturally loaded tests. The implication is that the presumed *g*-factor in the earlier studies could in fact be interpreted to include variance attributable to a G*c*-factor.

Dolan (2000), using multi-group confirmatory factor analysis, investigated the tenability that group differences were mainly caused by *g*. His approach allowed for comparisons, based on fit indices, of several models, with and without *g*. Although the models supported the conclusion of factorial invariance over groups, models with or without *g* showed little differences in explaining group differences. Thus, it was not specifically supported that group differences should be caused by *g*. In conclusion, Dolan suggests that Jensen's test should be abandoned in investigating group differences in cognitive ability test scores in favor of multi-group confirmatory factor analysis.

Dolan, Roorda and Wicherts (2004), although in agreement on the usefulness of a multi-group confirmatory factor procedure, did not support Dolan's (2000) conclusion of factorial invariance. Thus, they raised the issue if the same construct was in fact measured in the different groups. The possible absence of factorial invariance complicates the comparison of test scores of different groups, and could make an assessment of Spearman's hypothesis invalid.

Helms-Lorenz, Van de Vijver, and Poortinga (2003) challenged the principal vector procedure with a procedure designed to capture different aspects of complexity. They enlarged the analysis to focus not only on cognitive complexity, but also on cultural complexity. While cognitive complexity was defined according to Carroll (1993) and designated as *g*, cultural complexity, or cultural loading (c), was rated by a group of psychology students. In a factor analysis they found two virtually unrelated factors, representing cognitive (*g*) and cultural (c) complexity. They also found tentative evidence that cultural complexity (c) was at least as important as cognitive complexity (*g*) in the explanation of performance differences of majority-group and migrant children. In an evaluation of commonly used intelligence batteries they found that subtests that require extensive verbal processing are often also cognitively more complex. This will cause a correlation between cognitive complexity and verbal processing, which complicates the interpretation of observed *g* loadings. A test of analogies, for example, can be considered to test abstract reasoning skills and be loading mainly on a cognitive complexity factor, but it obviously also has verbal – and thus culturally loaded – components. With these finds they challenge the assumption that group differences in scores necessarily are tests of the Spearman hypothesis.

Based on their earlier finds Fagan and Holland (2007) also challenged the claim that IQ differences between racial groups are due to the *g* factor. They claimed that the search for racial differences in IQ should be directed toward differences in exposure to information. They conducted a series of experiments involving African Americans, White Americans, and foreign students, who were not native English speakers. Based on the assumption that intelligence is information processing ability, differences in intelligence can be interpreted as differences in information processing ability, or in differences in access to information, on which tests are based. Fagan and Holland designed problems which were typical of standard tests of intelligence, where parts of the problems were solvable on the information generally available to either race and/or newly learned, and parts of the problems were solvable on the basis of specific previous knowledge. They could show that specific previous knowledge varied with race, and that solution rates of problems based on such knowledge varied with race. However, problems based on information equally available to the two groups showed no differences in outcome, or outcomes that favored the African American group somewhat. They conclude that cultural differences in provision of information account for racial differences in IQ.

In sum, there seems to be serious grounds to question unconditional acceptance of group differences as a result of differences in *g*, and thus also the notion that established group differences are examples of Spearman's hypothesis.

6.3 Test bias and differential prediction

Differences in mean test scores could indicate actual group differences in the functions the test aims to measure or be caused by some biasing mechanism in the test. Differential prediction could be caused by differences either in the predictors or in the criteria. In relation to test outcomes for diverse groups te Nijenhuis and van der Flier (1999) made a distinction between internal and external bias. Internal bias indicates the degree to which a test might refer differently to a theoretical construct in different groups. External bias indicates the degree to which a test might have different predictive strength in relation to a criterion. Internal bias is often investigated by checking if items function differently, and/or if the factor structure is the same for all groups. External bias is established if a test over- or underpredicts with respect to a specific external criterion for a group. Differential prediction is established if there are significant differences between the regression equations for two groups as indicated by differences in the slopes, intercepts, or both (Sackett, Laczo & Lippe, 2003).

Sometimes also error variances are investigated. Regression lines illustrate the relationship between test and predictor for different groups. A difference in intercept indicates group mean differences in the predictor, while significant differences in the slopes of the regression lines indicate test bias. However, Sackett et al. warn that the model investigating the relationship between predictor and criterion must be correctly specified; otherwise spurious differential prediction may be found.

In investigating selection effects focus has often been on scrutinizing measures for prediction. Equally important is the critical assessment of outcome variables. Outcome measures are often based on global assessments of supervisors, and these measures may not have undergone the careful scrutiny of prediction measures. Schmidt, Shaffer, and Oh (2008) point out that criterion reliability is usually higher when outcome results, whether work performance or job training, are evaluated with standardized procedures rather than ratings. Since ratings are more common in European studies, these often show a lower value for predictive validity.

Rotundo and Sackett (1999) investigated possible racial bias in supervisors' assessments of work performance, but found no evidence of predictive bias against Black employees based on supervisors' race. Frisby (1999b) has also reviewed the literature on race of examiner effect, which states that African American test takers will show depressed scores when the examiner is Caucasian. He concluded that there was no strong evidence in favor of a consistent race-of examiner effect. In investigating outcome variables Stauffer and Buckley (2005) claim to have identified racial bias in supervisory rating of job performance. Roth, Huffcutt, and Bobko (2003), however, found that objective measures showed larger standardized group differences than that of supervisory ratings, but theorized that raters may experience some pressure to minimize ethnic group differences.

Research in the United States has dominated this area, with focus on the predictive validity for English-speaking minority groups born and raised in the country. Although questions have been raised concerning subcultures with possible differential vocabularies, to a large extent there is a common culture. In this context there is considerable agreement between researchers that the predictive validity is the same, or similar enough for practical purposes, over groups born and raised in the United States. Frisby states that "the equivalent predictive validity of individually administered intelligence tests across different U.S. racial or ethnic groups …. is well established" (Frisby, 1999a, p. 287). Schmidt states that "any given GCA test score has essentially the same

implications for future job performance for applicants regardless of group membership" (Schmidt, 2002, p. 203). Sackett et al. state that bias in tests is "unequivocally rejected within mainstream psychology" (Sackett, Borneman, & Connely, 2008, p. 222).

There is much more cultural diversity in Europe. The countries have diverse histories, cultures, and languages, as well as very diverse minority groups. A much larger share of the minority groups is foreign born. This makes it necessary to verify the American results in intercultural studies. In a meta-study Salgado, Anderson, Moscosco, Bertua, de Fruit and Rollande (2003) have shown that there is international validity generalization for GMA measures to predict work performance and work training. A meta-study from 2007 (Hülsheger, Maier, & Stumpp) addressed the question whether findings of American and European meta-analyses were possible to generalize to Germany. They found that overall German operational validities were comparable with findings in the United States or other European countries, with the exception of a slightly lower estimate for the training area. Hülsheger, Maier and Stumpp (2007) proposed that the early and strict tracking practiced in Germany leads to restriction in intelligence measures and educational achievement in job-specific applicant groups. With less variation in the predictor scores the association with training success is somewhat reduced in the German studies, compared to other countries. However, the predictive validity for the majority groups in Europe is tentatively established.

Reports on differential prediction for minority groups in Europe are scarce. However, there are some Dutch studies. Te Nijenhuis and van der Flier (1999) made a review of fifteen years of research on test bias in the Netherlands regarding immigrants from Surinam, The Netherlands Antilles, Morocco, and Turkey. They examined studies with immigrant children and job applicants for item bias and comparability of factor structures, and found only few examples, with limited effect, of item bias. They concluded that factor structures were highly comparable over groups. Te Nijenhuis and van der Flier (2000) also compared the predictive properties for native and immigrant groups in a vocational training situation. Using a step-down hierarchical regression procedure they found group differences in the slopes of some of the regression equations. The results from the cognitive part of the assessment related to the examination results in the same way for both groups, but there were differential predictions for combinations of predictors and criteria when these were less cognitive and less objective. The immigrants were differently assessed in relation to personality variables, but the relations were weak and non-significant.

De Meijer, Born, Telouw, and van der Molen (2008) investigated the criterion-related validity of cognitive ability for majority and minority applicants in relation to Dutch police officer selection. They found, in line with previous research, that cognitive measures of cognitive ability showed very little predictive power for this area of work. However, for the minority group cognitive ability showed predictive power in relation to training performance.

Thus, on the whole, the hypothesis of differential prediction for immigrant groups in Europe has found little or weak support. However, this conclusion is based on limited research, and further research from different countries, on different groups, and for different criteria (such as education, training, work areas) needs to be conducted.

6.4 Environmental sources of underachievement

Environment could influence group means through several channels; among the more obvious of these are differences in quality of home environment and schooling. These factors are usually controlled in studies of group differences. More indirect environmental effects can operate since people exist in socio-political contexts that have profound impact on their experiences and worldview (Suzuki & Aronson, 2005). This could cause differences in motivation and attitudes in the testing situation. Questions have been asked whether groups perform on a level below or above their potential for social reasons, such as a higher or lower perceived value of the area of cognitive achievement, or finding the area more or less suited to the self-image of the group.

Frisby (1999a, p. 281) has reviewed the literature on the impact of culture "circumscribed by race, ethnicity, country of origin, language, and/or social class", and test session behavior on cognitive tests. He discusses the aspect of "test wiseness", which includes time management strategies, knowledge of "rules of thumb", as well as reasoning strategies, and concludes that few studies have examined test wiseness in relation to race or ethnic groups. The studies with direct focus on this question have yielded inconclusive results; have found no group differences, or showed that scores on test wiseness measures do not predict test scores. Again, though, it must be observed that most of this research is limited to the United States.

An area of research connected to test wiseness is training studies. In this line of research different interventions are attempted to compensate for supposed lack in such areas as test motivation, assertiveness, and strategic skills. Frisby (1999b) concluded that the type of training varied greatly between studies. The evaluations of the effects also varied, both in outcome measures and in time

period elapsed after training. Finally the groups receiving training differed, which altogether makes it hard to evaluate the training results. There is the additional question of what gained scores might indicate. They might be a temporary situation bound effect or actual substantive gains in accessing potential cognitive functions both in the testing situation and future achievements. Skuy et al, (2002), working with African and non-African student groups in Africa, started with a pre-test consisting of Raven's Standard Progressive Matrices that showed the usual substantial difference in favour of the non-African group. Using a mediated learning experience, they obtained post-test results that showed significant positive effects from the learning experience, and non-significant results for race. They concluded that the mediation was effective in improving the performance of both student groups, but that the intervention was more effective in the African group. The intervention results, however, did not transfer to a similar measure, and there was no significant correlation between Raven SPM scores (measured either pre- or post-intervention) and course grades, which raised questions concerning what function the test – or the course grades – actually measured in these groups. It was also concluded that the mediated learning experience did not claim to produce cognitive changes, but could be useful as an assessment of the individual's potential for change.

Roth, Buster and Bobko (2011) cited sources that indicate that both majority and minority groups benefit from training, which is in line with the results from Skuy et al., (2002). However, they claimed that individuals with higher level of ability often learned or gained more, which means that group differences were not reduced.

An area which has received much attention is stereotype threat theory. The theory assumes that individuals can be threatened in their performance when negative stereotypes are activated. The resulting stress may cause underperformance. The threat is activated when the individual has a strong sense of identification with an academic domain, but feel pressure that his or her performance may be judged in terms of negative stereotypes. Stereotype threat has been proposed to be the cause of depressed test performance for African Americans, and for depressed test performance in the mathematical domain for women (Steele, 1997).

Reviewing the evidence, Frisby (1999b) found inconclusive evidence, possibly because the influence of stereotype threat is difficult to separate from general test anxiety in relation to different achievement levels. Seibt and Förster (2004) tried to identify the different mechanisms that affected performance. They proposed a model where negative as well as positive stereotypes could affect

performance strategies by introducing different regulatory foci. In a series of experiments they found that when negative stereotypes were induced this led to a vigilant processing style, with emphasis on avoiding mistakes. The result was higher performance accuracy, diminished creativity, and enhanced analytic thinking. When positive stereotypes were induced this resulted in an approach strategy with focus on producing as many hits as possible. This led to an explorative processing style with enhanced speed and creativity but diminished analytic thinking. Since a test situation often requires the test taker to go beyond the given information and generate many solutions in a short time, the performance will suffer from activation of a negative stereotype threat. However, attempts to induce positive stereotypes could also lead to avoidance behavior, when risks of failing were perceived to be high. Thus, the mechanisms of stereotype threat are complex.

Suzuki and Aronson (2005) found the evidence for stereotype threat to be abundant. Woodcock, Hernandez, Estrada and Schultz (2012) investigated the long term effects of stereotype threat. Based on theory, chronic stereotype threat was hypothesized to lead to domain dis-identification (meaning that members of negatively stereotyped groups progressively place less importance on their performance in the stereotyped domain) and eventual domain abandonment (meaning a complete dis-identification with no ego dependence on the domain and possible resistance to any encouragement to develop one). Stereotype threat has been shown to have the greatest impact on individuals with high ability who are highly identified with the domain in question. This creates what the authors describe as a psychological bind. While domain identification is important for academic performance and persistence, highly domain-identified individuals may, as a result of chronic exposure to stereotype-threatening situations, enter into recursive patterns of defensive disengagement that spiral downward. Eventually this may lead to their dis-identification with the domain. However, they also hypothesize that this effect may be moderated by the way stigmatized individuals disengage in the face of negative experiences. They tested their hypotheses with a panel of minority students of African Americans and Hispanic/Latino(a)s in the scientific track. An initial result showed that both student groups experienced stereotype threat, with African Americans reporting significantly higher levels of threat than Hispanic/Latino(a)s. However, the results also demonstrated different experiences and different disengagement strategies for the African American group versus the Hispanic/Latino(a)s. Thus, different regression equations were estimated for the two groups. The final result showed that the direct effect of stereotype threat on scientific identity was negative for

both groups, but statistically significant only for Hispanic/Latino(a)s. There was also an indirect, negative effect on intention to pursue a scientific career for the Hispanic/Latino(a)s. In the absence of a direct effect, no indirect effect could be established for the African Americans. Several explanations were offered for this find, among them that African American students often were enrolled at institutions were they were in the majority. The authors also discuss the possibility of group specific strategies in handling stereotype threats. Earlier research has indicated that Hispanic/Latino(a)s tend to devalue the domain in response to the perception of ethnic-based injustice, whereas African Americans tend to discount the validity of performance feedback. The process of trying to identify remedial measures must take into account that groups experience and react differently to stereotype threat.

Stereotype threat has not been extensively researched in Europe, but Gaines et al. (2012) investigated the effect of different forms of racism on susceptibility to stereotype threat for individuals of African descent in Great Britain. In earlier studies racism had been categorized into individual, institutional, and cultural racism. Contrary to expectation, none of these forms had substantial impact on susceptibility to stereotype threat. However, a newly formed category, collective racism, defined as the stereotyping, prejudice, and discrimination on the part of specific groups of persons toward racially stigmatized individuals, did show a significant relationship with susceptibility to stereotype threat. The authors discuss their results in relation to the research and results found in the United States, and theorize concerning the strategies of coping that were used. They also conclude that much more research is needed in the area.

6.5 Environmental sources of overachievement

In the United States the outstanding achievements in terms of school grades, in scores on content-oriented achievement tests like the SAT, and in academic and professional achievements from the Asian American group have been noted (e.g. Neisser et al., 1995; Rushton & Jensen, 2005). The relation to scores on intelligence tests, however, is debated. Lynn (1987) found evidence of significantly higher scores on a G*f* factor, as well as on visuospatial reasoning, and saw hereditary factors as the major influence. However, he based his finds on testing in Asia, not the United States, and a number of questions can be asked concerning cross-cultural comparisons regarding for instance norming, cultural practices, and schooling. Sackett, Schmitt, Ellingson, and Kabin (2001) found that Asians scored higher than Whites on measures of mathematical-quantitative ability but lower than Whites on measures of verbal ability and comprehension.

Sue and Okazi (1990) reviewed the research on differences concerning Asian and American groups, and found no conclusive evidence of innate characteristics being the sole or even major cause for the exceptional achievement patterns they found in Asian Americans. They also found the suggested influence from cultural values lacking in strong empirical support. They suggested that the concept of relative functionalism might be fruitful. Relative functionalism implies that a group excels in areas where it finds life opportunities. The relatively higher performance of Asian Americans at a level beyond their measured IQ could be caused by highly motivated and persistent performances in the academic area since this offers opportunities for upward social mobility, while other areas may be perceived as closed.

6.6 Attempts to reduce impact of group differences in selection

The practical consequence of group mean differences on cognitive tests is unequal group representations in activities where acceptance is based on cognitive test results, such as entering higher education, or access to some jobs. This has inspired an area of research with focus on how other measures can bring about more desired results. Sackett, Schmitt, Ellingson and Kabin (2001), working from an American point of view, discuss different approaches aimed at finding a desirable balance between performance and diversity. They concluded that technical adjustments, such as eliminating items that are culturally laden, will bring only a small reduction in different selection ratios, as will individual test preparation or test orientation. However, these adjustments will bring favorable reactions from the individuals who are tested. The only really effective measure in reducing selection ratio differences is to include predictors other than cognitive tests, such as personality measures, that are unrelated to the cognitive tests.

Several other authors (e.g. Halpern, 2000) have suggested non-cognitive measures (often measures of personality) to be included in a selection procedure, in order to enhance validity and reduce the effect of group differences on cognitive tests. However, Halpern also warns that many of the discussed alternative measures may be more biased than the standardized tests they are replacing. Bias in alternative measures may also be more difficult to detect. Potosky, Bobko, and Roth (2005) reviewed the research in the area of predictor composites, and found less than satisfactory methodologies, often measures suffering from range restrictions, data based on incumbents, and lack in coverage of measures. In a meta-analytic approach they extended earlier analyses

and applied corrected validity estimates. The outcomes indicated that adding non-cognitive measures produced relatively modest effects on adverse impact and validity. Adding a measure of conscientiousness to a *g*-measure reduced adverse impact from d = .72 to d = .68, and increased validity from .51 to .55; while adding an interview reduced adverse impact from d = .72 to d = .65 and increased validity from .51 to .61. Including bio data increased adverse impact. In order for a composite measure to reduce adverse impact or substantially increase validity the added measure must contribute these aspects, as well as be uncorrelated to the cognitive measure. The cost and feasibility of developing and applying such measures must be considered. Thus, they conclude that expectations on adding non-cognitive measures must be realistic.

Gamliel and Cahan (2007) support the use of predictive criteria measures with smaller manifest group differences from a statistical point of view. They contrast the much higher group differences in cognitive ability measures used for prediction with the smaller group differences in the actual job- and educational related criteria, and propose that lower scoring groups are biased in their selection rates as an inherent statistical result of any imperfect prediction. When predictive validity is imperfect, although unbiased, proportionately more false-negative decisions are expected in lower scoring groups. The contradiction that unbiased prediction necessarily leads to biased selection could be a justification for the use of measures with smaller mean group differences than those typical of cognitive tests.

Roth, Buster and Bobko (2011) investigated the possibility of reducing bias by using trainability tests as an alternative to cognitive tests. Trainability tests are distinct from traditional cognitive ability tests "because one learns, within the testing procedure itself, job-relevant information" (p. 35). Earlier studies have indicated that trainability tests could be an alternative to tests of cognitive ability since they had similar validity and were less likely to have any adverse impact. Roth, Buster and Bobko investigated these assumptions using two large databases that allowed them to examine the consistency of effects across jobs and locations. They found considerably larger group mean differences than in earlier work, which measured as an effect size amounted to a Black-White group difference of around 1. Criterion-related validity was also greater than in earlier studies. Mean prediction for training performance was around 0.7, and mean prediction for job performance around 0.4. Thus, trainability tests seem to be associated with substantial group mean differences on level with tests of cognitive ability. Given that cognitive ability is hypothesized to predict training results this is not surprising, although disappointing, considering how earlier

research had indicated a hope that trainability tests could be an effective alternative with none or reduced negative group impact in job selection to cognitive tests. In evaluating different selection procedures it should also be noted that the trainability tests were time consuming to develop, costly to administer, and time consuming to score. However, they were often positively perceived by the applicants.

6.7 Summary

Most research on group differences in cognitive test results has its origin in the United States, and concern minority groups that are native to that country, such as African Americans and Hispanics. For these groups measures of general mental ability show mean results of around one standard deviation below the results of the majority for the African American group, and a somewhat smaller difference for the Hispanic group. Much debate has focused on the sources of these differences, and questions of estimates of heritability influence have been especially passionate. However, environmental and hereditary factors interact, sometimes in complex ways, and influences on group level are difficult to disentangle.

Attempts to reduce adverse impact in selection ratios for minority groups have included analyses of measurement properties of cognitive tests, as well as introduction of alternative methods. These attempts have had limited success. Measurement properties have been shown to be only marginally responsible for differential test outcomes. Use of measures other than cognitive tests have been problematic, as these methods often have weaker prognostic properties, or are more complicated and expensive to implement, or both.

The European research in this area is limited, but presents results that are similar to the American, in that minority groups as a rule perform on a cognitive level below that of the majority. However, since Europe has a great diversity of groups and languages much more research is necessary.

7 Research questions and design of studies

The theme of this thesis is the special circumstances that come into play when cognitive tests, developed in a Western context and validated for a Swedish setting, are used in relation to groups with other languages or cultural backgrounds. The background chapters have discussed the lack of validation evidence that exists for use of cognitive test results as a base for decisions concerning vocational training and employment when immigrant groups are concerned. Immigrant groups form a substantial minority in the Swedish work force. Thus, aspects of test use in relation to immigrant groups need to be investigated.

Messick's integrated model of validity highlights the different aspects that use of test results pose, and is thus chosen as the theoretical framework. Aspects of construct validity form the base for test result interpretation. Cattell's Investment theory will be the frame of reference when investigating how differences in cultural background affect the relations between cognitive functions. The hierarchical model of intelligence will provide a theoretical frame when questions of test score evaluation form a base for decisions of acceptance or rejection. Questions of utility and social consequences of test use are discussed in relation to prognostic properties for admission into vocational training and for employment. Value implications concern all these steps.

In Study I the focus is on Cattell's Investment theory (Cattell, 1987) and its explanation of the relation between the broad cognitive factor of fluid intelligence G*f* and a general cognitive factor *g*. The Investment theory explains the general factor as a result of G*f* having been invested in different cognitive activities. *Gf* then is a factor of General Intelligence (*g)* because it is involved in all domains of learning. Since most cognitive activities are dependent upon the cultural context of the developing individual the theory can be tested through investigating the effects on the relation between G*f* and g of differential learning opportunities. In addition to investigating the characteristics of the general cognitive factor and its relation to fluid intelligence under different conditions of learning opportunity the study investigates the differential correlation patterns for a number of broad cognitive abilities, as well as level of functioning on these abilities. The outcomes will provide information on the nature of the *g* factor and thus contribute to construct validity. Study I also makes an attempt at

questioning some of the earlier results concerning the causes of cognitive group differences.

In Study II the challenges that the interpretative process poses when test takers come from diverse groups are investigated. Specifically this study investigates how the cognitive test information is used by the psychologists, who integrate test scores in order to assess the suitability of each applicant, and in the next step by the employment officers.

The study investigates how the psychologists evaluate test results in relation to the requirements, and how this may differ in relation to the cultural background of the test taker. This area relates to construct validity also, but has additional focus on the value implications that test interpretation generates. The aim is to evaluate how cognitive test information is used in Sweden by psychologists, who integrate test score information intuitively when assessing suitability for vocational training, and by employment officers, who make the decisions to grant vocational education based on the assessments. Characteristics of these assessments and decisions are analyzed through comparisons with statistical, model-based, integrations of test-score information. The psychologists did not have access to factor scores, but were obliged to make an intuitive or subjective integration of the available normed scores on tests. This approach was contrasted with the outcomes in factor scores, which can be seen as an actuarial method of integrating test information.

Actuarial methods have consistently shown higher predictive power (Grove & Meehl, 1996), and thus the factor scores were hypothesized to be objective and reliable estimates of actual function. The general difficulties of intuitively integrating large amounts of test information could be expected to increase when evaluating the performances of immigrant groups, since there were considerable differences in group means, and these differences were more pronounced on some test outcomes than others.

Thus, the focus is on the interpretation of the test scores, and possible differences in this process related to the cultural background of the test taker. Since the assessment of the psychologists can be expected to be influenced by their understanding of the cognitive functions that the tests aim to measure, their interpretations reflect on the construct validity of the tests, while value implications will emerge from the rates of suitability assessments and granted courses.

Study III has its focus on the prognostic properties of tests and possible differential impact of test outcomes and vocational training in relation to employment for different groups. In this study the influence of cultural

background of the test taker on the prognostic properties of the test outcomes is investigated. Here the focus is on test use, which, although based on the interpretative aspects of the former studies, has its focus on the possible differential utility of test scores and their social consequences. In the study employment rates as a function of cognitive test scores and vocational course participation for individuals with different migrational backgrounds are investigated. The research focus concerns the relative influence of vocational training, cognitive functions and immigrant status on employment. In addition the effect of cognitive complexity of courses is investigated. While aspects of construct validity can be illuminated by this study also, the main focus is on the value implications and social consequences, i.e. the consequential aspects of validity.

8 Method and data

In this section description of the subjects and the test data will be presented. The methods, with their strengths and limitations, will be discussed in relation to the research questions. Some reflections concerning possible interpretations, given the methods, conclude the chapter.

8.1 Subjects

For the research questions to be investigated a reasonably large group of individuals with diverse cultural backgrounds who had taken the same, or almost the same, combination of tests, was needed. Such a group was present at the Employment Service. The subjects all took part in a structured assessment procedure as a step in their application for a vocational training course of higher or lower theoretical complexity.

Cultural background was defined empirically throughout the studies, and based on the individual's country of basic schooling. The individuals were grouped according to geographical area of basic schooling into three categories; Swedish non-immigrants (SNI), European Immigrants (EI), and Non-European immigrants (NEI). Immigrants had spent a mean of 8.2 years in Sweden, with a range from 1 to 31 years (sd = 5.5). The EI group was dominated by persons from the Balkan area, but also included some persons educated in the United States, Australia, or New Zeeland. In the NEI group a majority of the applicants were from the Middle East and northern Africa. Most of the vocational training courses were aimed at work in the industrial and technical sector, and a majority of the applicants were male. The characteristics of the three groups are shown in Table 1.

Table 1. *SNI, EI, and NEI groups by number, proportion (percent), gender, and age in years*

	N	%	Male %	Age M	Age SD
SNI	2358	66.1	84.2	33.0	9.3
EI	620	17.4	87.7	34.9	8.0
NEI	591	16.6	92.7	34.7	7.2
∑	3570	100.0	86.1	33.6	8.8

All three groups had twelve years of schooling as the most common educational background, but the immigrant groups had a greater dispersion, with larger groups with very low or very high educational level. Both with respect to the pattern in educational background and the most frequent areas of emigration the immigrant groups in this study are similar to the immigrant groups as a whole in Sweden, as reported by Myrberg (2001).

8.2 Testing procedure

The general testing procedure was adapted to the groups involved. The applicants were often inexperienced test takers, and their school experiences not recent. For the immigrant groups it was an additional obstacle to be tested in a Swedish speaking environment. The testing procedure was consequently modeled to avoid bringing other factors than competence assessment into focus. The subjects were given advance notice as to what achievement areas would be tested, as well as suggestions on where additional information on the subjects could be found, and in some cases what books to read to prepare. Examples of content areas covered were 9th grade mathematics for electrician applicants, and driver's theoretical test on level B (private car) for applicants to commercial driver, level C. Information about the requirements was communicated to the applicants in group information meetings as well as in writing on the individual notice to attend a testing session on a certain day. If the applicant found the time period too short to do the necessary preparations it was possible to receive a later testing date.

The actual testing procedure was initiated with information about the purpose of the testing and the routines for test data interpretation and access. The outcome of the testing procedure was also communicated, i.e. the possible alternatives given the individual results. At this point it was possible for the applicants to refuse assessment; a decision which generally would close the possibility of accessing the specific training area. However, other areas of training or other measures, such as practice sessions with employers could still be open to the applicant.

Tests were selected to be as closely aligned with the area of assessment as possible. This was both to secure face validity and to establish content validity. Speeded tests and tests with overly complicated language were avoided. However, demands of speed and proficiency in Swedish were not avoided when adequate, since the target criterions made these demands also.

Regular follow-up sessions with the educators at the training centers secured that the requirement profile and selection procedure worked as planned. In some

cases the evaluations from practice periods at prospective employers were also included.

8.3 Data

The studies use cognitive test data from assessments made at the employment agencies in two counties of southern Sweden over the years 1993-2004. Over a ten year period the selection effects of market fluctuations on recruitment to vocational training should be diminished. Around 2500 assessments were performed in one county, and around 1000 assessments in the other. In the studies these data sets are pooled. Since the subjects came from different locations, patterns of selection based on geography should also be weak.

A broad array of tests was used. The psychologists at the local employment agencies had considerable freedom to choose tests of their liking for different purposes. The tests employed were selected for maximum efficiency for each requirement profile, but there were also some differences in selection of tests due to different traditions in the two counties. A few tests, such as Instructions, WIT Puzzle and WIT Antonyms, were taken by almost all participants. A core selection of 17 tests had been used with greater frequency. The data from these 17 tests have been used in the studies. Most of the tests are Swedish versions of well-known tests based on Thurstone's Primary Mental Abilities (Thurstone, 1938). Presentations of the tests are available in Study I, and in Appendices of Studies II and III.

The classification of vocational courses according to level of cognitive demands was also a part of the data base. This variable was established on the categorization of every course made by the project leaders, and consisted of a four level scale, where level 1 denoted orientation and guidance courses with no specific cognitive demands, level 2 denoted demands on compulsory school level (up to grade 9), level 3 on qualified vocational level corresponding to approximately twelve years of school training, and level 4 denoted demands on university level. Outcome data on granted courses and employment were provided by Statistics Sweden. The match between the applied course and the granted course was checked by the AMSYK[4]-number, which designated each course.

[4] AMSYK is the slightly adapted version of SSYK (Swedish Standars of Classification of Occupations) used by the National Labour Administration. SSYK is based on ISCO-88 (International Standard Classification of Occupations), which is published by the International Labour Office (ILO), Genève, 1988.

8.3.1 Missing data

Since different areas of training demanded different sets of skill, a diversity of tests was used. However, no single participant took every single test, which means that there are missing data. Missing data are troublesome, since they can lower the strength of the relations found, or distort the correlation patterns, if they are not random. Problems of differential patterns in data can be handled in different ways. A traditional method would be to exclude units were data are incomplete. Another traditional method is substituting missing data with means. Both these methods are regarded as inefficient and obsolete. Schafer and Graham (2002) recommend taking advantage of modern computerized technique and consider missingness as a probabilistic phenomenon. They find two procedures "highly recommended"; maximum likelihood (ML) and Bayesian multiple imputation (MI). However, these procedures can only be used when data are missing at random (MAR). Missing at random implies that the missingness is random, given the information in the data. This is a less restrictive assumption than "missing completely at random". The major cause for missing data was that the two test sites relied on partly different tests for the same area. There is no reason to assume that this should cause a threat to the validity of the MAR assumptions. Another major factor that determined the composition of the test battery was for which program the applicant was tested. While there is likely to be a certain amount of self-selection to different training programs, this should not cause any serious threat to the MAR assumption. Also, there were high interrelations among observed variables which are exchangeable indicators of a limited set of latent variables. This implies that there is much information in the data, which should allow for good possibilities to satisfy the MAR assumption. Consequently, the analyses were conducted with the maximum likelihood estimation techniques implemented in the Mplus program (Muthén & Muthén, 2004).

8.4 Choice of methods

The choice of method or methods must respond to the type of research questions that are asked. Different methods shed light on different types of questions, allow elaborations at different level of abstraction, and have procedures for validation that differ. Methods in social science are of diverse kinds. Some methods focus on elaborate, detailed and nuanced information that allow empathic understanding and insight into experienced relationships. These methods are often based on phenomenological theories, sometimes of a narrative nature, or based on hermeneutics, often investigating how a

phenomenon is situated, and what environmental circumstances shape and constrain the phenomenon of interest. Other methods focus on abstract, law-like relations investigated from an external perspective. In this perspective insight and knowledge is based on the standardized and formalized ways data are treated and aim to generalize to other situations of similar character or constitution. Jensen (2008) describes how these two approaches were formally distinguished early on in German philosophy as Verstehende Psychologie, on one hand, and Naturwissenschaftliche Psychologie, on the other. The first places psychology among the literary humanistic approaches to understanding experience. The other treats psychology as an empirical natural science, based on objective measurements, explicit testing of hypotheses by experimental methods, and statistical tests of significance. The perspectives have also been described as idiographic, which describes the individual case, versus nomothetic, indicating the search for lawlike relations.

A grouping of this kind will imply a dichotomy into qualitatively or quantitatively oriented methods, respectively. However, such a dichotomy is neither fruitful nor accurate, in describing the variations in existing methods. Gustafsson (2008) suggests that the dichotomy between qualitative and quantitative approaches should be replaced with distinctions between low and high-level inference approaches with respect to data, generalization and explanation. This stance more clearly indicates that different methods suit different types of problems. The first type of methods often captures important insight in scanning problems of new or complicated types and can form a starting point for more formalized hypotheses. The latter type can be used to investigate a variation of phenomena of a more well-known nature, where variations in one or more aspects can be identified, investigated and generalized to situations of similar kind. Thus the same knowledge area can be investigated with different types of methods at different points of development.

The different types of methods are thus more complementary than mutually exclusive in their perspectives and will have their strengths at different levels of abstraction. Several metaphors have been used to illustrate this point. Moss, Girard and Haniford (2006) citing Lampert, use the analogy of a camera lens, shifting focus and zooming in and out. The same motif will appear different depending on the focus. Gustafsson (2008) draws an analogy to meteorology, where climate studies can give substantial information about long term changes and differences in geographical locations, but say little about the weather at a certain spot tomorrow, whereas weather forecasts have focus on predicting the immediate, local conditions, but say little about long term changes.

The research questions of this thesis all concern culturally related differences on cognitive tests; their causes, implications and consequences. The substantial individual variations are not the major focus of interest. There is a primary interest to establish the magnitude of group differences (if any); their relationship to a theory of cognitive structure and cognitive development; the implications of the differences when test results are interpreted, evaluated, and used as a basis for predictions; and finally the effect of culturally caused differences on valued outcomes on group level. As a consequence the general approach is more of a "climate" than a "weather" study, mainly utilizing statistical methods. The aim is to investigte broad outcomes on group level.

8.5 Cognitive functions and confirmatory factor analysis

Structural equations and confirmatory factor analysis are methods that can be used to assess a wealth of theoretical problems. Here the method is presented in relation to cognitive functions, since this is a major concept of the dissertation.

Cognitive functions, defined as latent variables, are presumed to be the cause of the manifest differences in test results. Latent cognitive functions are in themselves not directly observable but are established on the basis of theory and hypotheses. They are studied through measures that they should relate to, given the validity of the theory. The hypotheses concerning relationships of latent cognitive functions are formulated as structural equations models, which state the relations between the latent variables. The measurement part of the models describes the relation between manifest and latent variables. Measurement models are also called factor models (Loehlin, 2004). The main source of information concerning the measurement models is correlational evidence. The correlations will show patterns of relations in test result data that are typical. Thus models that describe the relationship between the latent cognitive variables and their relationship to manifest outcomes for different groups have been hypothesized in the studies.

The purpose of the models is to capture general trends, while the single case is of subordinate interest. Although the feasibility of the models has to be discussed on grounds of theory and credibility, the probability of the models can be tested against available data, using statistical measures that give an estimation of the certainty of the estimations. This approach makes the reliability of the inferences quantifiable, and thus open for an objective critical evaluation.

The fit between models and the structure of manifest data is estimated using confirmatory factor analysis. In confirmatory factor analysis the researcher uses theory to state explicit hypotheses about causal relations. These are then tested

against the material available. In exploratory factor analysis, as a contrast, there are no qualifications or causal relations assumed. Exploratory factor analysis is sometimes used as a preliminary step with the purpose to find typical relations that can then be studied further. However, "It is by no means a necessary step" (Loehlin, 2004, p.17). The fit of the model in relation to the empirical material is estimated, and stated in quantitative terms. Although there is almost always a choice between several estimation methods and model fit measures, in this mode of procedure the choices become transparent, and open for criticism. Thus, both results and validity issues can be confronted on a shared information base that facilitates communication.

In all three studies the Mplus Version 3 program (Muthén & Muthén, 2004) under the STREAMS 3.0 modeling environment (Gustafsson & Stahl, 2005) was used. Mplus is a program designed to facilitate modeling in a computerized environment, bringing together many statistical methods, and using standardized model building language. However, the process of model building with many steps from importing data, setting up models, choosing a particular procedure for estimation and interpreting the outcome is still complex and a single mistake along the way can have far reaching consequences. This problem is to a large part solved by resorting to the STREAMS modeling environment. STREAMS, which stands for Structural Equation Modeling Made Simple, standardizes the preparation of data, executes the necessary steps in data import, and structures the output. All this adds to the reliability of the procedure.

In theory a model is set up, estimated and assessed. However, in practice there can be obstacles on the way. The chosen estimation method may not be possible to execute in relation to the empirical material. In Studies II and III the models set up in estimating relationships of cognitive models in relation to outcome variables were too complicated to converge. Here, instead, the work had to proceed in two steps: first, factor scores on latent variables were computed from the cognitive model, and second, the factor scores were used as independent variables assessing their predictive power on outcomes in education and work.

8.6 Measuring fit

When models are compared to data it implies comparing the estimated covariance matrix to the observed covariance matrix. A straightforward measure would be χ^2; this measure however, will produce significant results with only small differences between the models when the number of observations is large and vice versa. An alternative is RMSEA, Root Mean Square Error of

Approximation (Loehlin, 2004). RMSEA is a population based fit index which takes the number of observations, as well as the number of estimated parameters into account. Estimates below 0.05 indicate good fit between the estimated and observed covariance matrix, whereas estimates up to 0.08 are acceptable.

RMSEA should be computed with a confidence interval. The width of the confidence interval gives information about the precision of the estimate. The upper limit of a 0.90 confidence interval should not exceed 0.08.

Another measure used to determine fit is SRMR (Standardized Root Mean Square Residual). The SRMR measure indicates an average of the residuals defined as the standardized difference between the observed correlation and the predicted correlation. An SRMR measure of 0.08 or less is considered satisfactory (Loehlin, 2004).

8.7 The use of contrasting methods

In Studies I and II there are examples of comparing outcomes produced by two different methods. In both cases there is an argumentation of the relative validity of each method.

8.7.1 Contrasting principal component analysis and structural equation modeling

In Study I the structural equation models describing the relations of hypothesized latent variables to specific tests for different migrational groups were set up and estimated. The estimations of the models were conducted through structural equation modeling. Group differences have been studied with other methods. Jensen (1998) has used a procedure of principal component analysis. Both principal component analysis and structural equation modeling are based on the correlation matrices produced from the manifest variables (i.e. test results) and should produce similar results. However, there are also differences. The principal component method allows the first principal component to be defined in a way that captures a maximum of correlations, regardless of what functions that give rise to the correlation. It is thus an explorative method. The structural equation method is confirmatory in that it specifies the expected patterns of relations and thus which sources of influence are allowed. The principal component method has been criticized for bias since the result often will overestimate loadings for G*c*-test and underestimate loadings of G*f*-tests. When testing a similar one-factor model in confirmatory factor analysis the one-factor model showed poor fit, since it presupposed a structure that did not fit the data.

8.7.2 Contrasting statistical and intuitive methods of integrating data

Computing factor scores for latent variables is a statistical method of combining a large number of test result data into a single measure, using set rules deciding how each component should be weighted. This can be contrasted to a more intuitive process in integrating data, such as when the psychologists integrate a number of test scores into an assessment, relying on their theoretical understanding of the concepts that the scores measure. The intuitive approach is the method actually employed by the psychologists in the directed aptitude assessment procedure. A large number of studies on the subject of intuitive versus statistical integration show that the statistical integration in most cases produces more reliable results (e.g. Grove & Meehl, 1996). In Study II there is a comparison between the outcomes in terms of assessments of suitability and of granted courses in relation to factor scores, computed from the latent variables, and the psychologists' intuitive integration of test results.

8.8 Dummy variables

Some of the variables used in this study are of a non-quantitative nature, such as migrational grouping. An individual is assigned to any of the three groups labeled SNI, EI or NEI. In order to be able to study the influence of these variables in statistical models they have to be transformed into a numerical value. A choice of numbers indicating "greater than" or "smaller than", such as 1, 2, 3, would be misleading. Instead, each group will form a category, where a single individual is only a member of one of the three. Group membership will be noted as present (=1) or absent (=0) in each category. Only two variables are needed to represent the three groups, since one group is a reference category which will automatically be the alternative when the outcomes of the other two are zero. In the studies the SNI is usually the reference group.

Since migrational group has scale values minimum 0 and maximum 1 the unstandardized regression coefficients can be interpreted as percent increase or decrease contributed by the migrational group variable on the dependent variable. An example would be a regression coefficient of -0.17 for the EI dummy variable on the outcome employment. This can be interpreted as a 17 percent lower chance for a person in the EI group to be in employment, compared to the SNI group, when all else is equal.

8.9 Regression analysis

A variety of regression analyses are used in the studies. In a regression analysis the relationship between one or several independent variables to a dependent variable is studied. The terms "independent" and "dependent" imply causal relationships, but the regression analysis cannot truly verify this, even if a relationship is significant. The regression only verifies the correlation. Questions of causality must be determined on other grounds. However, the outcome of the regression analysis will indicate the degree of change in the dependent variable, when an independent variable varies, given that the other independent variables are fixed.

In Study II a number of regression analyses are conducted. The independent variables are group status, i.e. belonging to one of the immigrant groups, and results on the cognitive tests expressed as latent variables. The dependent variables are assessment of suitability, being granted a vocational course and cognitive level of granted course. The regression equations are set up for the different dependent variables and the independent variables are added one by one. First the effect of group status only on the dependent variable is investigated. Then the other independent variables are introduced one by one. The effects on the group status variable of introducing other explanatory variables are noted. By introducing the cognitive variables one by one differential effects of the different cognitive functions can be noted. Since the cognitive variables correlate with one another it is not possible to study a separate effect if they are introduced together. However, in a final regression equation they are all combined, to study their total effect. A combined measure of the effects of the independent variables on the dependent variable is the amount of explained variance, R^2.

8.9.1 Multiple and logistic regression

Since there were two types of dependent data, two types of regression equations were used. When the outcome variable was continuous a multiple linear regression model was used. When the outcome variable was of a non-quantitative categorical nature, such as being granted a training or not, or becoming employed or not, the relations were investigated using a logistic regression procedure. In this procedure the relations between the factor scores and the outcome variables are expressed as a probability of being in either of the outcome categories as a result of the influence of the predictors. The probability is reported as a log value, and the model is called logistic regression. When the coefficients for the independent variables have t-values equal to or larger than

1.96 this indicates that the probability of reaching this coefficient by chance is less than 5%.

The logistic models are less sensitive than linear models, since there is smaller variation in the outcome data. A single individual is either in the outcome category or not. For instance, it is not possible to be "slightly" or "almost" granted a vocational course. Thus, some information is lost in this process. When possible, a linear model which takes better advantage of the data should be used. Depending on the assumptions about the nature of data different methods of calculating the regression are possible. Violations of these premises may produce slanted results or estimates of reliability that are not trustworthy. With more advanced uses of computer technology methods have been developed that can handle slight deviations in the distribution of data.

An example and discussion of logistic versus linear models can be found in Study III. In this study the outcome studied was if the individual was in employment or not, i.e. a binary variable. A logistic model was first attempted, but became too complicated to converge when the independent variables were added. Instead a Maximum Likelihood procedure with Robust standard errors (MLR) was used. The MLR procedure presupposes linear relationships. Despite the fact that this was not the case it could be shown that the estimates produced were reasonably reliable. For instance, the correlation between the intercept and the slope in the growth model that was created in Study III was -0.47 when estimated with the MLR procedure, while it was -0.38 when estimated with WLSMV (Weighted Least Squares Estimator with Robust Standard Errors and a Mean- and Variance adjusted Chi-Square; a logistic estimator), and thus highly similar. Also, the use of an estimation model which presupposes linearity was supported by the fact that there were no extreme proportions in the outcome variables (the employment rates ranged from 37.9 percent in the lowest group to 78.6 percent in the highest). A discussion of the use of the two procedures can be found in Beauducel and Herzberg (2006).

8.10 Growth models

Sometimes measures on the same variable repeated over time are used. This will introduce special complications. A great number of outcome measures can be difficult to sum up into a coherent picture. Possible interventions of disturbing processes can cause random variations that are not related to the independent variables but may affect the variable of interest. These problems can be reduced by creating a model which captures the essential elements of change in the variable over time. A growth curve model can be set up as a structural equation

model, where a model is created which estimates individual differences in growth with latent factors. This procedure was used in Study III, where the outcome in employment status was measured at three points in time, the year of the assessment and the two consecutive years. As could be observed directly from descriptive data all groups increased their employment rates over the years, but at different rates and from different starting points. The effects of migrational group, achievement on the latent cognitive factors, of being granted AMU, and on the cognitive level of granted course were assumed to have systematic influence on the outcome employment. In order to simplify an overview of development in employment rates, a model was created where employment status was modeled by two variables; employment at the year of testing was modeled by a latent variable for the intercept of the growth curve, while the development over the two consecutive years was modeled by a latent variable for the slope of the curve. The effect of the independent variables on the intercept and slope of this model were then investigated in a number of regression equations.

8.11 Standardizing scales; z-scores and effect size

Descriptive data, such as means and standard deviations on a number of measures, are presented in the three studies. The scales often differ in means and distributions, which make comparisons difficult. By standardizing different measures on a common scale comparisons are facilitated. Much used for this purpose is z-standardization. In z-score standardization the mean of a reference group is set to zero, and the standard deviation is set to one. The formula for z-transformation is $z = (x-m)/sd$, and thus a z-score is a measure of the deviation from the mean of the reference group (m), presented in a scale that is based on the variation in the measure. The standard deviation (sd) used should be that of the control group, if such a group is identified.

In this work the SNI group is often the reference group, which means that the mean of this group is subtracted from a measurement, and the remainder is divided by the standard deviation for the SNI group. Differences between groups in z-score means, for instance, can then be directly interpreted as effect size. Effect size gives information on the magnitude of a difference, not just if it is significant or not. When differences of means are concerned, Cohen (1969) has described effect sizes of around 0.2 as "small", an effect size of 0.5 as "medium", and an effect size of 0.8 as "large".

8.12 Summary

Testing with groups from the Employment Service provided the data for analysis of cognitive functions. On the basis of the country of schooling the individuals were grouped into Swedish non-immigrants, European Immigrants, and Non-European Immigrants. Statistics Sweden provided outcome data, such as acceptance to vocational training, level of course, and employment.

The methods used to analyze the data can all be categorized as supporting questions concerning variations on group level. At times contrasting methods are used, since this can support a conclusion with different arguments. The methods are described in a way that should support an understanding when reading the studies.

9 Results

The empirical part of this thesis consists of three studies, where the research questions have been investigated. The studies all concern test use with immigrant groups and progress from questions concerning construct validity, via interpretation and evaluation of test results, to the use made of test results in relation to vocational training. The studies thus touch on both evidential and consequential aspects of test interpretation and use. The combined outcome of the studies will form the basis for a validity discussion.

9.1 Study I

In Study I the main focus was on Cattell's Investment theory (Cattell, 1987). The hypothesis that G*f* equals *g* was tested both in the pooled group with mixed cultural background, and in the separate cultural groups. It could be shown that a model where G*f* equals *g* had an acceptable fit for all three separate cultural groups, while the relation was far from perfect when the groups were pooled. Thus, the learning environment must be reasonably similar for the individuals in a group for the relation to hold true. Given the condition of similar learning environments this study gives strong support for the inference that the *g* factor in a hierarchical model built on a sufficiently broad array of tests is identical with G*f*.

In Study I it was established that there were considerable group mean differences with respect to Stratum II cognitive factors. These differences were most pronounced for the G*c* factor, as could be expected, since the immigrant groups had had restricted opportunity to learn and practice Swedish vocabulary, grammar and general knowledge. The differences were also quite considerable on the visuo-spatial factor G*v*, which indicates that cultural factors influence these areas also. The group mean differences on the general factor G*f*, however, were of smaller magnitude. The group means on all cognitive factors are presented in Table 2.

Table 2. *Group means on cognitive latent factors. The means were standardized on a scale where the means for the Swedish non-immigrant group was set to 0, with a standard deviation of 1*

G*f*	G*c*	G*v*	G*s*
0.00	0.00	0.00	0.00
-0.14	-1.84	-0.63	-0.31
-0.40	-3.38	-1.28	-0.72

The distribution of the differences in means over factors differs from those often found in studies investigating group differences. Here the largest differences are found on the G*c* factor, but there are also considerable differences on the G*v* factor.

Jensen (1998) has expounded the much-cited "Spearman hypothesis" (Spearman, 1927), which states that the magnitude of the group differences in performance on cognitive tests are a function of their loading on the *g*-factor. The differences are thus expected or interpreted to be large on tests with high loadings on *g*. In this study several methods were employed to test the Spearman hypothesis. When a Structural Equation Modeling (SEM) technique was used the results did not support the hypothesis. When the frequently used method of correlated vectors was performed with a well-fitting model including all four latent variables/dimensions, the results were similar to those obtained with the method of SEM. There was no significant correlation between G*f* and group differences, but substantial and significant correlations between group differences and G*c*. However, when a simplified one-dimensional model was used the results were in agreement with earlier research, in that the differences in means over groups correlated highly with this dimension. It could be shown, however, that this model had poor fit, and that the dimension investigated had strong loadings on verbal tests. As discussed in the section "The *g* factor in the work of Jensen" the often used principal component method has been shown to be sensitive to the composition of the test battery upon which it is based, and specifically it has been shown to be biased in favor of G*c*-tests (Ashton & Lee, 2005). It can be concluded that the method of estimating a principal component and relating it to group differences can be misleading and cause unsubstantiated conclusions concerning group differences to be made.

Study I thus brings information on several validity aspects concerning the general intelligence factor. It shows that the relation between the general factor and G*f* is unity when the culture in which the individual is educated is reasonably constant, and thus it brings support to Cattell's investment hypothesis. The

pattern of mean differences on latent variables shows that the cultural background of the individual has a strong influence on G*c*, as could be expected, but also on G*v*. This finding is important since it has been argued that tests with visuo-spatial content are culture free, and thus can be used as a neutral measure of intelligence for groups with different cultural backgrounds. The conclusion that the visuo-spatial problem solving process is culturally influenced is also supported by the work of Cliffordson and Gustafsson (2008), who found differential schooling effect on the G*v* factor, and by Nisbett, Choi, Peng, & Norenzayan (2001), who investigated holistic versus analytic cognition in different societies, and found differences both on the perceptual input and visuo-spatial processing.

In the measurement part of the model it could be shown, that on the whole the pattern of influence from the latent variables on test outcome was similar over groups. However, there were also some minor differences. A tentative conclusion of the patterns of loadings is that most of the G*f*-tests capture aspects of this latent variable as well for the EI and NEI groups as for the SNI group. An exception is Raven's SPM, which captures less of the variation in the SNI group. Possibly there could be a restriction of range in the Swedish group.

The influence of the latent variable G*c* on tests aimed to measure this aspect is considerable and highly equivalent for all groups. On the instrument level it can be concluded that the Instructions is mainly a test of G*c* for all groups.

The tests aimed to capture influence from G*v* have their major loading from the latent G*v*-factor to a similar degree in all groups. Tests that are designed to be measures of G*f* or G*v*, but were verbal information and texted instructions bring in a component of G*c* tax somewhat more on the G*c* resources in the immigrant groups.

The G*s*-tests seem on the whole to be equally effective in capturing aspects of speed in all three groups. It is interesting to note that the tasks based on the Latin alphabet are no exception.

In sum some tasks are more verbally challenging for the immigrant groups. For some tasks this is also the case concerning G*v* for the NEI group. However, on the whole, the patterns of influence of latent variables on tasks are similar over groups. This indicates that generally the same functions are measured in the three groups.

9.2 Study II

The focus of Study II was on the interpretation and use of test information by the psychologists when assessing suitability, and on the employment officers when granting training. Considering the substantial differences in mean scores it is not surprising that the number assessed as suitable by the psychologists differed over the migrational groups. While 82.3 percent of the SNI group was assessed suitable for the chosen vocational course the corresponding number for the EI group was 68.0 percent, and for the NEI group 49.7 percent. These differences were significant ($p < .01$).

The differences between the groups in number of courses granted by the employment officers were slightly less pronounced. In the SNI group 65.0 percent were granted a vocational course, in the EI group 62.6 percent, and in the NEI group 54.8 percent. Although these group differences were smaller than those for psychologists' assessment they were still significant ($p<.01$). The difference in employment officers' decisions compared to the psychologists' recommendations could be due to the fact that the officers had access to information on actual jobs and assessed the need for vocational training with this in mind. They could also grant a lower level course, or an orientation course, for an individual who was expected to have difficulties in the course originally applied for.

Also significant was the difference in cognitive level of granted course over the three groups. For both the SNI and the EI groups most of the granted courses were on the two intermediate levels, i.e. 2 and 3. For the NEI group most of the granted courses were on level 2. The NEI group also received more orientation courses on level 1. However, the three groups had different patterns of application for courses, in addition to the differential patterns of achievement and of being assessed suitable for the wanted course. When all these variables were taken into consideration the differences in level of granted course over the three groups were only minor.

The independent influence of factor scores and migrational group status, respectively, were investigated in logistic regression equations where psychologists' global evaluation of suitability and vocational officers' decisions to grant courses were dependent variables.

The influence of factor scores was investigated in detail by analyzing each cognitive factor separately. Controlling for level of G*f*, the psychologists assessed members of immigrant groups as significantly less suitable for vocational training and the vocational officers granted significantly lower cognitive level courses for these groups. For the NEI group there was also a significant negative influence

on the vocational officers' decision to grant a course. The same pattern appeared when level of performance on the *Gs* factor was taken into account. Controlling for the level of performance on the *Gv* factor the psychologists assessed the EI group suitable on par with the SNI group. The NEI group were less often assessed suitable, a negative influence which was significant with a very small margin. Controlling for level of performance on the *Gc* factor, however, the outcomes for the immigrant groups were reversed. All outcome factors, psychologists' assessment of suitability, officers' decision to grant a vocational course, and level of granted course were all significantly more positive for both the immigrant groups.

Thus, both psychologists and vocational officers actively compensated the immigrant groups for the disadvantage to be tested in Swedish in making their assessments and decisions. Considering the level of performance on the visuo-spatial factor there were only minor differences in the assessments and decisions for the three groups. But given the same level of functioning on G*f*, and also on G*s*, the immigrant groups were not assessed suitable and granted courses to the same extent as the SNI group. The adverse impact of belonging to an immigrant group was only slightly reduced by the impact of the *Gf* and *Gs* factors. This is especially remarkable when the *Gf* factor is considered, since this factor can be expected to have great impact on the possibility to profit from instruction and perform in the work situation.

When all four cognitive factors were used in the regression equations the effects of migrational status on psychologists' evaluation of suitability was much reduced. It was non-significant for the EI group, and just barely so for the NEI group. This model explained 44 percent of the variation in the psychologists' evaluation of suitability, indicating that the outcomes on the cognitive factors were an important source of influence on psychologists' assessments. Using the combined influence of all factor scores neutralized the effect of belonging to an immigrant group on the decisions of the vocational officers to grant a course, and produced a significant positive effect for the immigrant groups on level of granted course. The influence of the four cognitive factors explained 5 percent of the variance in officers' decisions to grant a course and 14 percent of the variance in level of granted course, indicating that the results on the cognitive tests had some influence on the decisions of level of granted course, but only very minor influence on officers' decision whether to grant a vocational course or not. The combined effect of all the factor scores is not easily interpreted in causal terms, however, since the factors are correlated.

The lack of positive impact of the G*f* factor could possibly be explained by a difficulty for the psychologists to discern this factor. The psychologists did not have access to the computed factor scores, but based their assessments on the normed scores of actual tests. Since the immigrant groups had considerably lower mean performance level on verbal tests, and also somewhat reduced level of performance on the visuo-spatial tests, the psychologists' intuitive impression of problem solving capacity might be negatively influenced, especially since some of the tests used in measuring general problem solving ability also had considerable influence from the G*c* factor. The psychologists might then have difficulties in fully recognizing the problem solving potential in the immigrant groups. With an assessment of the general problem solving factor that could be expected to be not fully reliable, and possibly also evading some of the problems concerning the verbal capacity of the immigrant groups, the psychologists might make less than optimal recommendations for training area and level. Implementing factor scores as a routine in testing procedures that concern immigrant groups could increase the possibility for valid conclusions to be made.

9.3 Study III

In Study III the focus was on employment rates, and the influence of AMU and cognitive functions on this outcome. The outcomes in terms of employment rates were considered for three years; the year of assessment, and an additional two years, since the full effects of the training programs were only seen after this extended time. The considerable delay in effects could be caused by the lengthy process of receiving and finishing a vocational education. Some individuals needed to rehearse school subjects before starting a course. Especially in the middle years of the 1990s there was considerable waiting time, sometimes up to a year, before a place on a course was available. There could also be waiting time between a base module and later training. Some of the more qualified courses were quite lengthy, in some cases more than a year. Thus, only after three years was the full impact of the training completely visible. However, some applicants did not receive any training at all, and were thus immediately ready for employment and others after a short time.

During all three years in the follow up period the immigrant groups had lower employment rates than the SNI group. The immigrant groups showed a higher increase rate over time than that for the SNI group, which somewhat reduced the differences between the groups in employment rates over the years. However, in the three years of follow up the gap was never completely closed.

The rates of employment for the three groups with and without AMU are shown in Table 3.

Table 3. *Rates of employment with and without AMU for the SNI, EI, and NEI groups, in percent over the years 0-3*

Year	SNI		EI		NEI	
		No		No		No
	AMU	AMU	AMU	AMU	AMU	AMU
0	58.7	61.4	40.2	40.0	38.6	37.9
1	70.7	66.8	57.6	49.1	51.8	47.2
2	78.6	66.3	72.0	54.3	64.0	53.7

The effects of having been granted AMU, of belonging to an immigrant group, and of cognitive factors were investigated in a series of models, where the independent variables were added successively. The outcome in terms of employment rates over the three years was expressed as a growth curve. A latent variable for the intercept of the growth curve indicated employment rates in year 0, while a latent variable for the slope indicated the development in employment rates over the two following years. In the first model the influence of AMU on the growth curve model was estimated. There was no significant effect of AMU on year 0 (the intercept of the model), but significant positive influence on the slope, indicating a positive development over time. The effect amounted to 7 percentage units increase each year. Entering a variable for immigrant group status showed a significant negative effect on the employment rates year 0, with 20 percentage units for the EI group and 22 percentage units for the NEI group. However, both groups had steeper growth rates over years 1 and 2, indicating that the gap between the SNI group and the EI and NEI groups diminished somewhat over time.

Considering that the effect of granted AMU might be different for the cultural groups, interaction variables between granted AMU and the EI and NEI groups were introduced. None of the coefficients for these variables were significant, which supports an interpretation that granted AMU had similar effect for all cultural groups.

The cognitive factors all had positive influence on employment rates, but the influence was rather weak. Adding cognitive factors to the models only slightly reduced the negative impact of belonging to an immigrant group, indicating that the main part of the differences in employment rates for the three groups had

other causes. Again interaction variables for the cognitive factors were introduced for the EI and NEI groups. All were non-significant, except for the interaction variable for EI and G*c*. Introducing this factor made the initial (year 0) negative influence of belonging to the EI group smaller. However, the positive influence of the EI variable on the slope was also diminished and the increase in amount of explained variance small. For these reasons the more complex model with interaction effects was not pursued further. On the whole, the cognitive factors were interpreted to influence employment rates to a similar extent for all three groups.

A final result in Study III was the differential influence pattern on employment depending on the cognitive demand of the offered course. Courses at level 2 had a direct positive effect year 0 (the year of assessment). However, in years 1 and 2 this effect waned. In contrast, courses at level 3 and 4 had a negative effect on employment in year 0. Courses of longer duration probably lowered the rate of active job-seeking, especially in the initial phase. Still, the long term (employment years 1 and 2) effects of having been granted level 3 courses compensated for this initial negative effect. For level 4 courses the initial negative effect was stronger and in the time of follow-up the effect was just positive for year 2. Part of this differential effect may be explained by the length of the courses, but there is no absolute relation between course length and complexity. When evaluating effects of AMU the time span of the follow up must thus be sufficiently extended in order to catch the differential effect of courses of different complexity, especially for courses granted in successive modules. Since the migrational groups had different patterns of application, with immigrant groups more often applying for level 2 courses, this is especially important in evaluating possible differential effect of AMU.

10 Discussion

The focus of this dissertation is on validity issues in relation to cognitive test use with immigrant groups in Sweden. Messick (1989) has provided a useful matrix that serves as a guide in covering relevant aspects of validation as well as integrating evaluative judgments into a coherent picture. The matrix will be a framework for a discussion that concerns aspects of test interpretation as well as test use, and consider both evidential and consequential aspects. The process will move back and forth over these aspects, since they are interrelated and influence each other.

According to Kane (2006) validation employs two kinds of arguments; the interpretative argument and the validity argument. The interpretative argument specifies the assumptions and inferences made from test scores and the resulting interpretations and decisions. The validity argument evaluates these chains of conclusions according to their degree of coherence, reason and plausibility. This chapter will present some interpretive arguments and evaluate their credibility. The discussion will concern not just aspects of test use when diverse groups are concerned, but also general validity questions.

10.1 Construct validity of higher order cognitive factors and test interpretation

At the center of Study I is the question of the nature of the third order factor *g* and its relation to the second order latent variable G*f* as defined by Cattell (1987). The relation between these latent variables is of central importance when the attributes of the general factor *g* are to be understood. It is also fundamental in validating Cattell's Investment theory, which gives a theoretical explanation concerning the development of intellectual powers into different domains. In Study I it was shown that the relation between G*f* and the general factor *g* is perfect, when groups with similar educational or cultural background are considered. This makes the *g* factor defined, since its characteristics can be understood to be those specified for the G*f* factor. The characteristics bring substance and understanding concerning the construct of *g* and thus indications as to how it can be interpreted. In this way the construct validity of tests used to assess *g* is supported, both for immigrant and non-immigrant groups.

Study I also shows that the method of estimating the *g* factor influences the content of this factor, and that the presumed *g* factor in many studies has

substantial influence from especially verbally loaded tests. The cultural influence on the presumed *g* factor is often unintended and sometimes unrecognized, but may influence the predictive properties for education and work in the specific culture. Since the G*c* factor has been shown to have considerable predictive power for adults in a learning situation (Ackerman & Beier, 2006) the presence of G*c*-aspects in a presumed G*f* measure may in fact increase the prognostic properties. However, it will also affect the interpretations that can be made from the measure, and may also have adverse impact for individuals, such as non-native Swedes, who as a rule perform lower on tests with Swedish verbal content. Thus, the nature of the *g* factor must be defined since this has implications both for interpretation and prognostic properties.

While the results from Study I support the theoretical understanding of G*f* as a driving force in cognitive development it also shows the influence of cultural and educational background in producing different patterns of structuring the intellectual process. Thus there are somewhat different patterns of relations between the stratum II latent variables and the test scores over the migrational groups. A specific test could challenge mainly G*f* reasoning potential or G*v* visuo-spatial skills in the Swedish group, while it could put some demands also on G*c* reasoning skills in the immigrant groups. While these differences are only minor in groups that are culturally and geographically close (SNI versus EI), they increase somewhat with the cultural distance involved (SN versus NEI). It is important to note that cultural background influences not only G*c*, as expected, but also G*v*. The latter find is important since visuo-spatial tests have sometimes been considered to be culture-free, and thus suitable as unbiased measures in groups with diverse cultural backgrounds.

10.2 Group means on Stratum II latent variables; construct and value implications

Study I demonstrates that all four cognitive factors yield different group means on Stratum II latent variables for different migrational groups, but to varying degree. It is likely that the way a cognitive concept is viewed in a certain culture will influence the way an individual from that culture processes problems that tax on that cognitive concept, and that this in turn is at least in part responsible for the differences in performance level on Stratum II variables such as G*v* and G*s*. The differences in group means must be tentatively interpreted when it concerns the content and construct validity of these functions. The difference in factor loadings for single tests could generate culture–related hypotheses, such as which type of processes are used when the individual with immigrant

background solves cognitive tasks. However, it should also be noted that – with the exception of some tests of G*v* and G*f* with verbal components – the overall pattern of influence of cognitive functions on specific test outcomes is quite similar for all groups.

10.2.1 Crystallized ability, G*c*

Crystallized ability, G*c*, which is formed by the individual's interaction with educational and cultural experiences, is by its very definition culture dependent. In traditional interpretations this cultural influence has been restricted to the variations in schooling and home environment in the culture where the test was developed. When the perspective is widened to other cultures the interpretation must also be affected. Certainly the interpretation of G*c* must consider that language structure and vocabulary is explicitly culture specific. Other aspects, such as experiences gathered in the educational curriculum in social sciences, may be less so, and yet others, such as knowledge of how to apply mathematical rules, may be possible to transfer from one cultural context to another. When the level of *Gc* is measured in Swedish context for a person with non-Swedish background the tests can be seen as tapping into a content that may be partly familiar, but also partly quite alien. This, again, could indicate that a low measure for a person with immigrant background could be understood to be qualitatively different from that obtained by an individual with Swedish background. In the latter case the measure could be expected be more even over different sub-aspects of G*c*, while a variation in level related to the specific cultural content of different tasks could be expected from a person with immigrant background. Of course, an individual with immigrant background could also perform on a consistently low level as a genuine reflection of culture specific ability.

Study II showed that both psychologists, in evaluating suitability, and employment officers, in granting vocational courses, made significant positive adjustments in relation to the cut-score level for the immigrant groups, thus taking into account differences caused by the different educational background of the groups when interpreting the test results that related to crystallized ability, G*c*. The differential interpretation of a G*c* result depending on the cultural background of the applicant implies an understanding that this factor functions differently in Swedish speaking versus non-Swedish speaking groups. If the achieved level on the G*c* factor was under the cut-off value for a Swedish applicant this should indicate difficulty in the training programs, and the applicant thus advised against a specific vocational training program. For the immigrant groups an alternative interpretation is suggested by the standpoints of

the psychologists and vocational officers, who often chose to grant training at below cut-score level. The implication could be interpreted to be understood as a greater capacity for verbal tasks than the test outcomes indicate, and possibly also a developing potential. The way outcomes on the G*c* factor were evaluated indicated that the psychologists saw the whole G*c* complex, content and process, as being possible to stimulate into substantial growth in adult life, in circumstances such as that of vocational training. The individual setup of the courses could augment the stimulating effect. The interpretation supports a view where G*c* is seen as a factor with substantial potential to develop, even well into in adulthood. This differential interpretation could tentatively be supported by the fact that the positive effect of vocational training on employment is similar for the different groups, despite the fact that immigrants enter with much lower mean level of proficiency in the verbal area.

10.2.2 Visuo-spatial ability, G*v*

The visuo-spatial abilities have long been assumed to be culture independent, since they are not explicitly taught, but acquired in the individual's interaction with the physical environment. Based on this assumption visuo-spatial tests have often been presented as "culture free". However, as presented in the section "Intelligence in non-Western cultures" the cultural impact on the perception and interpretation of visuo-spatial stimuli can be considerable. In Study I outcomes show group differences in means on tests that support the interpretation that there is substantial cultural influence on the G*v* factor. However, Study II showed that the psychologists made no differential use of G*v* scores. In contrast to the differential interpretation of results on the G*c* factor level of functioning on the latent G*v* factor did not have differential influence on the decision making process. Possibly the psychologists have had a traditional view of the G*v* factor both as fixed early in life, and also unaffected by the specific culture of schooling. This could indicate that they saw this cognitive area as more firmly established in adult age than the G*c* area. It is also possible that the demands on visuo-spatial functions in the learning environment and later in work were interpreted to be so pronounced as to leave no margin for achievements below the cut-off score.

10.2.3 The fluid intelligence factor, G*f*

Mean level of functioning on the G*f* latent factor showed the smallest group differences of all latent factors. However, when level of functioning on the G*f* factor was considered, the psychologists, and in the next step the employment

officers, did not give the immigrant groups full credit, but withheld training to a higher degree. This is an unexpected outcome, since the psychologists should be aware of the central function of G*f* in new learning situations. The outcome could give rise to several hypotheses in interpreting this result. A possible hypothesis is that the immigrant groups were discriminated against. This hypothesis, however, seems less likely in the light of positive discrimination when evaluating results on the G*c* factor. Another hypothesis is that the immigrant groups are seen as needing higher level of G*f* to be able to be successful in the training programs. However, no such conclusions were articulated in the requirement profiles. The third hypothesis is a technical one and is based on the demonstrated difficulties in subjective integration of complex information versus actuarial methods. Since the psychologists did not have access to factor scores, but only test scores, they had no direct way of estimating the G*f* ability. The problem solving content of a test could then be hard to distinguish from the verbal and visuo-spatial contents. Since the immigrant groups had significantly lower means on the verbal tests, and somewhat lower means on the visuo-spatial tests, this could cause the psychologists to underestimate the level of functioning on the G*f* factor for the immigrant groups. If this line of reasoning is correct it should point to the need to establish testing routines which make it possible to estimate latent factors in a more standardized way at the Employment Service.

10.2.4 The speed factor, G*s*

The *Gs* factor showed moderate group mean differences for the SNI, EI, and NEI groups. Some cultural influence could also be expected for the speed factor G*s*, which can be valued differently in Western and non-Western cultures. However, the pattern for interpreting and evaluating outcomes on the G*s* factor was similar to that of G*f*, i.e. given the same result on the G*s* factor the immigrant groups were less often assessed suitable and granted training to a lower degree. There is no obvious theoretical or practical explanation for this practise. However, the G*s* factor had the weakest influence on outcome factors such as being granted vocational training and being in employment, so the effect of the differential treatment was weak.

10.2.5 Evaluating all factor scores combined

While the analysis of the interpretation and use of the single cognitive ability gives input to hypotheses about construct meaning the consequential effects of test use must be evaluated from the total influence of all factors. Here the

negative outcomes were cancelled by the positive, and the total effect on being granted a vocational course was similar for all groups. In addition, the differential pattern of evaluating test outcomes resulted in immigrant groups being granted higher cognitive level of vocational courses, given the combined influence of all factors. Thus, in this respect test use had a small positive differential consequence for the immigrant groups.

Although considering all factor scores simultaneously generated an outcome where the three groups had similar chances of being granted vocational training and the immigrant groups had significantly higher chances of being granted higher level courses, this need not be an optimal outcome. The information on the G*f* factor was not fully used, and the immigrant groups received considerable compensation on the G*c* factor. Since the cut-score levels of the different factors were established considering the demands of the different courses, granting access for individuals with less than required proficiency in Swedish could produce difficulties for these individuals to profit by the instruction. Te Nijenhuis, de Jong, Evers, and van der Flier (2004) have shown that when immigrant and Dutch children are matched on school achievements or intelligence tests, the immigrant children are more often recommended higher education, but also face a higher risk of repeating classes or dropping out. However, in the case of vocational training the courses should make allowances for individual needs, which could reduce the problem. There could also be an expectation from the psychologists that the learning environment in AMU should stimulate a rapid progress in the G*c* area for the immigrant groups. This, in turn, would imply an interpretation of the G*c* area as open for considerable development at an adult age.

10.3 Social consequences of test use

Study III investigated the relation of factor scores and vocational training to employment. While it could be concluded that the immigrant groups generally had lower levels of employment than the non-immigrant Swedish group, the overall influence of factor scores on employment rates, although low, was positive. In Study III the question of differential prognostic properties of factor scores on employment for the immigrant groups was explored through the use of interaction variables. All interaction factors but one were non-significant. The interaction between EI and G*c* showed a significant negative influence on the variable for the slope, making it more similar to that of SNI. However, it did not bring a substantial increase on the amount of explained variance, so the simpler model without interaction effects was preferred. This outcome brings support to

an interpretation that the prognostic properties for the cognitive factors in relation to employment rates are similar for the three groups, although it should be noted that the influence of cognitive factors on employment rates was restricted.

Study III also evaluated the relation between having been granted a vocational education and employment rates, i.e. the influence of AMU on employment. Since the three groups had been granted training on slightly different interpretations of their test results, and some advantage given to the immigrant groups concerning the G*c* factor, it could be expected that employment training might have a weaker outcome, in terms of employment. However, the positive prognostic properties of employment training were basically the same for all groups. This could be due to the way vocational training was offered. Training should be offered in a way adjusted and suited to the individual.

10.4 Summary evaluation

The theme of this thesis is validity issues concerning the use of tests developed for Swedish groups but applied with groups of other cultural origin. With the data available it has also been possible to address questions about the general nature of the *g* factor, and its function in a hierarchical model of intelligence. The general construct validity of measurements involving the *g* factor has thus been elucidated.

The power of the finds from the studies stems partly from the fact that the presence of data from different cultural groups allowed a quasi-experimental approach. It was possible to explore hypotheses about direction and causation, not just correlation, as requested by Bornstein (2011) and Borsboom, Mellenberg, and van Heerden (2004). Aspects of cognitive functioning were investigated in a way that could resemble a systematic variation in the prerequisites, where the cultural background is the cause of variations in measures.

Messick (1989) calls for validation of both evidential and consequential aspects of test use. Often many different, and competing, value implications must be considered. Messick (p. 81) mentions the following issues:

> "…to be fair to individuals in terms of equity, to groups in terms of parity or the avoidance of adverse impact, to institutions in terms of efficiency, and to society in terms of benefits and risks – all at the same time".

It is hardly possible to achieve all these objectives in an optimal manner. The desirable state to strive for is likely to be a compromise between different values

cherished by the society in which the tests are used. Also, this compromise is likely to be tentative, with a need to be scrutinized and re-evaluated, as circumstances in the surrounding society evolve. However, the goal of the validation process, according to Messick, is primarily to make these value conflicts explicit and articulated, while the validation process per se cannot determine which choices should be made. Halpern (2000) is another researcher who acknowledges the value laden compromises that must be made, and a purely scientific stance as inadequate.

Thus, in practice, questions of utility, such as the degree of prognostic efficiency of tests, must be weighed against the degree of equity and social consequences of the outcomes. The double aim of the Employment Service – to support growth in the economy and to support groups with a weak standing on the labor market – mirrors these aspects. The arrangement of the assessment procedure could be said to attempt to meet requirements of both evidential and consequential nature. The assessment ensures that individuals accepted to vocational training have (almost) the necessary qualifications. At the same time the assessment procedure does not create adverse impact for the immigrant groups.

Cognitive tests have been used in assessing and advising job seekers for over a century and are generally thoroughly validated. The meritocratic effect of decisions based on tests scores is long established, and according to Kane (2006, p. 29) "many inferences and assumptions are sufficiently plausible a priori to be accepted without evidence." Selecting individuals on basis of their resource profiles as measured by tests to an educational and work situation with a specific requirement profile should be an example of this. The more subtle questions of possible differential consequential aspects for the immigrant groups in the assessment procedure are addressed in the three studies and discussed above. The outcomes show that some differences related to cultural groups can be identified concerning construct interpretation and evaluation. The pattern of influence from different latent variables on specific tests varies somewhat over tests, although on the whole the similarities are substantial. Thus, although some "internal bias" (te Nijenhuis & van der Flier, 1999) can be identified, the magnitudes are small.

In contrast to the small differences in structure of cognitive abilities the differences on the mean level of functioning of different cognitive factors were substantial. Not surprisingly, this was especially true for the broad cognitive function of G*c*, but also to some extent for the visual-spatial area of G*v*. Smaller differences were seen in the areas of G*f* and G*s*, indicating that these data do not

support the Spearman hypothesis. The similarity in acceptance rates to AMU and training outcome, measured as employment rates in relation AMU, supports the interpretation that the social consequences of test use in this context were similar for all groups, i.e. that the "external bias" (te Nijenhuis & van der Flier, 1999) was negligible. This outcome, however, was based on test interpretations and evaluations that compensated the immigrants for being tested in Swedish.

While the use of tests had no general negative effect for the immigrant groups in accessing vocational training it may also not have been optimal. The differential interpretation of the G*c* factor may have been completely adequate in relation to the immigrant groups. It may also have overrated their competence, and thus have induced a greater amount of individual adjustments necessary in the vocational training situation. The lower estimation of the G*f* factor may have caused individuals from the immigrant groups to have been directed to courses of lower complexity than necessary. These reflections could indicate a need to introduce a treatment of test scores where latent variables are computed and thus made explicit. If scores on latent variables were made available to the psychologists this could bring benefits such as a better match on content and level of courses to the individual.

There has been no systematic evaluation of alternatives to the present procedure. However, the situation before Directed Aptitude Testing was introduced was an example of such a situation. There were many instances where an individual in training failed to reach acceptable results. Often a break in the training would occur after considerable time. Since there were considerably more applicants than training opportunities there was first a sometimes long wait for an opening. After starting training, a considerable time could pass before it was established that the area or level of training was unsuited for the applicant. This situation caused substantial emotional cost for the individual, who had spent valuable time and effort, and an equally substantial economic cost for the organizations. This general picture applied to immigrants and non-immigrants alike. Thus, the general benefit of the selection procedure is difficult to question, and there is no reason to believe that abstaining from assessment should be more beneficial for the immigrant groups.

Returning to Cronbach's (1971) question about the payoff of using tests, the conclusion here is that members from the immigrant groups on the whole reap the same benefits as the non-immigrants. Benefits of training, in terms of employability, also seem to be similar to that of the Swedish group. A slight disadvantage can be found in level of cognitive complexity in relation to results

on the G*f* factor. This could possibly be remedied if test outcomes were modeled in relation to latent variables.

10.3 Limitations and suggestions for further research

This dissertation investigates aspects of test use in immigrant groups. It does so using latent variables rather than test scores, since latent variables make the results interpretable in a theoretical context and comparable to other research in the area. However, in practice the outcomes in terms of psychologists' assessments of suitability and employment officers' decisions to grant a vocational course or not, were not based on factor scores but on test scores. It would therefore be of interest to see if the use of factor scores would influence these outcomes differently, as has been suggested especially concerning the G*f* factor.

Another source of complexity is that the assessments are based on medical, social and personal information, in addition to the cognitive factors. Non-cognitive skills have also been shown to influence employment (Heckman, Stixrud & Urzua, 2006), and were in fact used in the Directed Aptitude Testing. Non-cognitive and cognitive skills are somewhat related and including non-cognitive skills in the analyses could make the specific influence of cognitive skills more visible.

The subjects in this study were not randomly chosen individuals from their populations, but self-selected or selected by their vocational officer as suitable candidates for the courses offered. To be a client at the Employment Service involves processes of selection and self-selection, and so does being considered for vocational training. These selection mechanisms have created an obvious slant in the male/female proportions and possibly also attracted individuals with specific interests or talents in the areas offered. It is thus necessary to be cautious in making generalized conclusions about cultural differences, since the groups investigated are not representative samples from any well-defined cultural groups. It would be of great interest to have studies replicated with other groups, which should be more clearly defined in this respect and preferably also should be more balanced with respect to gender composition.

Subjects have been grouped according to school background from Sweden, European countries outside Sweden, and countries outside Europe. This is a crude grouping, concealing many varieties and profiles of cultural patterns in organizing knowledge and proficiencies. It would be of interest to conduct similar studies with more finely meshed grouping. There is a continual change in the composition of the groups that migrate into Sweden. For this reason also

replication with newer immigrant groups can bring important aspects into the picture.

Yet another factor worthy of more interest in the immigrant groups is the length of stay in Sweden. Under what circumstances and to what extent do cognitive and/or non-cognitive skills develop over time in the new environment? Are there interactions with age, gender, and/or initial level of functioning? Research on these subjects could teach us more about the properties of different constructs concerning cognitive functions under varying conditions.

Studies with more stringent outcome variables could be very elucidating in all of the areas considered above. Training outcomes have not been available and are only indirectly assessed through employment rates. Concerning employment it has not been possible to establish if it matches the original assessment area, and later training area. The vocational training courses also have a great variation in content, length, and prospects for future employment. The grouping into four cognitive levels only crudely represents this diversity.

The ultimate purpose of test use at the Employment Service is to increase the chances of suitable employment. It has been shown that the employment rates of immigrant males are lower in Sweden than for comparable groups in other countries. This could be due to differences in the surrounding social systems, as suggested by Kesler (2006), to negative discrimination, as suggested by Behtoui (2004), and to differences in competence, as suggested by Carneiro, Heckman, and Masterov (2005) with data from the United States. Study III indicates that differences in competence as expressed in measured cognitive functions are only marginally responsible for the differences in employment rates for the investigated groups. However, it is possible that the relations between cognitive functions and employment rates are not linear, and that there are certain thresholds above which the relations are stronger. In order to determine the role of cognitive scores as predictors for training and work success for immigrant groups in Sweden much more information is needed.

Svensk sammanfattning

Inledning

Alltfler individer rör sig över kultur- och språkgränser och strävar efter att etablera sig på en annan arbetsmarknad än den i uppväxtmiljön. Samtidigt ökar kraven på kognitiva förmågor i arbetslivet. Kognitiva test har visat sig användbara för bedömning av en individs resurser i samband med utbildning och arbete och har funnit omfattande användning i västvärlden. Testanvändning har skapat stora vinster (t.ex. Cook, 1993; Schmidt & Hunter, 1998), men även orsakat intensiv debatt (t.ex. Gould, 1996; Neisser et al. 1996). Förespråkarna lyfter ofta fram effektivitetsaspekter. Motståndarna är ofta engagerade i frågor om rättvisa och jämlikhet. Det är alltså angeläget att analysera båda dessa aspekter. Utvärdering av såväl effektivitet som rättviseaspekter berör validitet. Vid låg validitet minskar effektiviteten. Om det finns aspekter av testningen som fungerar olika för olika grupper äventyras rättvisan. Dessa båda validitetsaspekter kan förefalla olika, men ses mer fruktbart som två sidor av samma problem. Messick (1989), har föreslagit en validitetsmodell, som har funnit bred acceptans. Den innefattar inte bara testtolkning, utan även testanvändning och dess effekter i ett socialt sammanhang. Medan utvärdering av reliabilitet och generaliserbarhet kan ske via rent vetenskapliga kriterier, innebär en sådan validitetsmodell att även sociala och etiska överväganden måste göras, när validiteten bedöms.

Effektivitetsaspekter i samband med testanvändning är grundligt utvärderade. I en omfattande metaanalys fastslog Schmidt och Hunter (1998, 2004) att mått på generell begåvning var en effektiv prediktor för prestation i yrkesutbildning och arbete. Rättviseaspekter har fått stor vetenskaplig och samhällelig uppmärksamhet, särskilt i USA. Olika etniska eller kulturella grupper har ofta visat sig ha olika medelvärden på de kognitiva testen, och även om skillnaderna är små kan detta ge stora effekter på urvalskvoter. I Europa har forskning kring gruppskillnader i kognitiva förmågor och samband med olika gruppers möjligheter på arbetsmarknaden inte varit lika omfattande, men kunskapen ökar även här (Helms-Lorenz, van de Vijver, & Poortinga, 2002; Salgado, Anderson, Moscosco, Bertua & Fruyt, 2003; te Nijenhuis, de Jong, Evers & van derFlier 2004; te Nijenhuis & van der Flier, 1997, 2000, 2003, 2004). Resultaten går i stort i samma riktning som i USA, vilket innebär att minoritetsgrupper ofta har ett

lägre genomsnitt, men att den prediktiva validiteten mot arbete och utbildning i ett givet samhälle i stort är densamma.

Sverige har en aktiv arbetsmarknadspolitik, som bl.a. syftar till att på kortast möjliga tid integrera nyanlända grupper i arbetslivet. Ett redskap i arbetsmarknadspolitiken är arbetsmarknadsutbildning, vars effektivitet har blivit kontinuerligt utvärderad (Adda, Costa Dias, Meghir & Sianesi, 2007; de Luna, Forslund, & Liljeberg, 2008; Martinson & Lundin, 2003; Richardson & van den Berg, 2008). För att öka effektiviteten i programmet har Arbetsförmedlingen ett bedömningsförfarande (Riktad arbetspsykologisk utredning, "RA"), som syftar till att säkra att individen har de grundläggande, nödvändiga resurserna för en viss utbildning (Valentin Kvist, 1992). I RA genomgår den sökande ett standardiserat bedömningsförfarande, som inkluderar kognitiva och praktiska test, samt strukturerad intervju. Individens resursprofil jämförs med kravprofilen för det aktuella området, vilket resulterar i ett omdöme avseende lämplighet för den aktuella utbildningen.

Avhandlingens fokus

Avhandlingen fokuserar på frågor om validitet i testanvändning och testtolkning, när kognitiva tester används för invandrargrupper i Sverige. Specifikt undersöks testanvändning i samband med urval till arbetsmarknadsutbildning, samt i relation till anställning. Tre studier bildar grundvalen för ett resonemang om validiteten i testanvändning för invandrargrupper.

Metoder och data

Den empiriska delen av avhandlingen fokuserar på likheter och skillnader på gruppnivå. Med hjälp av statistiska metoder, framför allt strukturell ekvationsmodellering och konfirmatorisk faktoranalys, skapas utfall som bildar underlag för resonemang och slutsatser. Med detta arbetssätt blir slutsatserna möjliga att värdera i kvantifierbara termer och att diskutera kritiskt.

Studierna använder data som samlats in i samband med RA i södra Sverige under åren 1993-2004. Sammanlagt genomfördes ca 3500 testbaserade bedömningar. Ett stort antal kognitiva test fanns representerade. Studierna grundar sig på de 17 mest frekvent använda testen. På utfallssidan fanns data avseende psykologernas bedömning av lämplighet, arbetsförmedlarens beslut att bevilja en kurs, kognitiv kravnivå på beviljad kurs på en 4-gradig skala, samt uppgift om anställning.

De utbildningar som bedömningen avsåg var huvudsakligen inom industrisektorn och de arbetssökande som deltagit i bedömningsförfarandet var till största delen män (86.1 procent), med en genomsnittlig ålder på 33.6 år (sd 8.8). Uppdelningen i kulturella grupper har gjorts utifrån var den arbetssökande genomfört sin grundläggande skolgång och enligt denna definition var 66.1 procent icke invandrade svenskar (Swedish non-immigrants, SNI). Invandrargrupperna delades in i europeiska invandrare (EI), vars andel motsvarade 17.4 procent, och utomeuropeiska invandrare (NEI), vars andel var 16.6 procent. I alla tre grupperna var 12 års skolgång den vanligaste utbildningsbakgrunden, men invandrargrupperna hade större andel med mycket lågutbildade och mycket högutbildade individer.

Begåvningsteori

Begåvning beskrivs utifrån psykometriska teorier, med fokus på latenta faktorer. En utgångspunkt är Carrolls (1993) hierarkiska begåvningsmodell, där kognitiva faktorer beskrivs på tre nivåer – strata – utifrån bredden i deras inflytande. På stratum I finns i runda tal ett sextiotal faktorer med relativt smalt inflytande. På stratum II återfinns ett begränsat antal breda faktorer, som till antalet har varierat något, men där de som varit föremål för mest forskning är Fluid Intelligence (G*f*), Crystallized Intelligence (Gc), samt Broad Visual Perception (Gv). G*f* avser förmågan till problemlösning, där problemet har en okänd eller ostrukturerad form. G*f* avser förmågan att tänka nytt och fritt, att finna mönster i kaos, att skapa och använda sig av nya förhållningssätt. Även G*c* avser problemlösningsförmåga, men på ett sätt som funnit en struktur i den kultur man lever i. Det kan avse förmågan att använda språket för att bena upp och förstå ett sammanhang, eller applicera strategier och regler av matematisk eller formell natur för att lösa ett problem. G*v* avser förmågan att mentalt identifiera, visualisera och hantera två- och tredimensionella figurer av viss komplexitet. Ytterligare en faktor är G*s*, som avser förmågan att på ett effektivt sätt avväga snabbhet och noggrannhet i enklare uppgifter. De latenta kognitiva faktorerna G*f,* G*c*, G*v* och G*s* och deras egenskaper i relation till individens kulturella bakgrund belyses i de tre studierna.

På stratum III placerade Carroll (1993) en generell begåvningsfaktor, *g*.

Studie I: The relation between fluid intelligence and the general factor as a function of cultural background: A test of Cattell's Investment theory

I Studie I ligger fokus på förståelsen och tolkningen av den generella begåvningsfaktorn *g* på Stratum III i Carrolls modell. Den generella begåvningsfaktorn har varit föremål för omfattande vetenskaplig diskussion utifrån om den är en statistisk artefakt, och därmed varierar beroende på sammansättningen i ett testbatteri, eller om den har en entydig definition, och därmed kan förstås på ett invariant sätt. Cattells definition av den breda faktorn *Gf* (1963) har stora likheter med Spearmans (1904, 1927) beskrivning av den generella faktorn *g*, och det har därför föreslagits att detta skulle vara en och samma faktor.

Cattells Investmentteori (Cattell, 1987) postulerar att individen i sin utveckling föds med en förmåga att se mönster och samband, G*f*. Allteftersom individen mognar och socialiseras in i sitt samhälle investeras denna förmåga inom olika kognitiva områden, där individen utvecklar förmågor och färdigheter. Eftersom dessa breda förmågor har sitt upphov i individens G*f* kommer de breda förmågorna ofta fungera på likartad nivå, vilket ger upphov till korrelationer mellan mått på dessa funktioner. Ett antal studier har visat ett samband mellan *g* och *Gf* som pekar mot att de är identiska (Gustafsson 1984, 1988, 1994, 2002; Keith 2005; Reinolds & Keith 2007; Undheim 1981; Undheim & Gustafsson, 1987), medan andra studier inte visat detta samband. I Studie I utnyttjas det faktum att grupperna har olika kulturell bakgrund. Cattells (1987) Investmentteori prövas genom att den generella stratum III-faktorn *g* ställs i relation till begåvningsfaktorn G*f* på Stratum II, dels för varje grupp för sig, dels för alla individer sammanslagna i en enda grupp. Utfallet av detta upplägg visar, att relationen är lägre än 1.0 när individer med olika kulturell bakgrund blandas, men i praktiken blir identisk, när grupperna särskiljes utifrån kulturell bakgrund. Cattells Investmentteori finner alltså stöd, givet att den kulturella miljö där individen utvecklas har likartade drag. Detta bidrar till att den generella faktor kan definieras på ett entydigt sätt, vilket bidrar till dess begreppsvaliditet.

Den generella faktorn *g*, eller dess motsvarighet G*f*, är också i centrum för en genomgång av hur olika metoder att analysera gruppskillnader i resultat på kognitiva test ger olika utfall. Spearmanhypotesen (Jensen, 1998) indikerar att gruppskillnader är en funktion av testets laddning i den generella faktorn. Studie I visar på avsevärda gruppskillnader i prestation på de olika kognitiva faktorerna, men där dessa är av olika storlek, beroende på vilken kognitiv förmåga som avses. De största gruppskillnaderna återfinns inte inom G*f*, utan i G*c* och även i

viss mån i G*v*-faktorn. Genom att pröva två olika analysmodeller mot varandra visas att Spearmanhypotesen inte får stöd i detta material.

Analys av de latenta faktorernas (G*f*, G*c*, G*v* och G*s*) inflytande på konkreta testresultat pekar på att dessa mönster till stor del ser likartade ut för de olika kulturella grupperna. Detta ger stöd för att testen ställer krav på samma latenta kognitiva förmågor, oberoende av den testades kulturella bakgrund. Undantagen gäller test av förmågor inom G*f* respektive G*v*, som har skriftliga instruktioner eller skriftligt innehåll, vilket ställer krav på G*c* för grupperna med icke-svensk bakgrund. Samtidigt visar de avsevärda skillnaderna i gruppernas genomsnittliga prestation inom de olika kognitiva områdena att det kulturella inflytandet är stort, framför allt på faktorn G*c*, vilket kan förväntas, men även på faktorn G*v*.

Studie II: Interpretation of Cognitive Test Scores in Relation to Swedish and Immigrant Groups

I studie II ligger fokus på frågan hur psykologerna tolkar testutfall på olika kognitiva faktorer, relaterat till den testades kulturella bakgrund. I denna studie visas, att andelen individer som psykologerna bedömer lämpliga för sökt utbildning skiljer sig signifikant mellan de tre grupperna. Av SNI blev 82,3 procent bedömda lämpliga, av EI 68,0 procent och av NEI 49,7 procent. Detta utfall är inte oväntat med tanke på de avsevärda skillnader som finns mellan gruppernas resultat. Även arbetsförmedlarna beviljade utbildning i signifikant olika grad för de olika grupperna, men här var skillnaderna mindre. Av SNI blev 65,0 procent bedömda lämpliga, av EI 62,6 procent och av NEI 54,8 procent. Anledningarna till att arbetsförmedlarna gjorde annorlunda bedömningar än psykologerna kan vara flera. I arbetsförmedlarnas uppdrag ingår att göra en arbetsmarknadspolitisk bedömning av individens möjligheter på arbetsmarknaden. Det är möjligt att arbetsförmedlarna beviljat en annan utbildning än den som individen bedömts för. Det är t.ex. inte otänkbart att en individ som inte bedömts lämplig för en mer krävande yrkeskurs beviljats en kurs på lägre nivå, eller en vägledningskurs.

I sitt arbete har psykologerna inte haft tillgång till de statistiskt utvärderade resultaten på kognitiva faktorer, utan har fått utgå från en intuitiv integration av konkreta testresultat. Tidigare forskning (bl.a. Grove & Meehl, 1966) har funnit att statistiska metoder att integrera data genomgående ger ökad prediktiv styrka. De statistiskt beräknade faktorpoängen bedömdes därför vara reliabla mått på individens faktiska funktion.

I studie II ställs psykologernas bedömning av individens lämplighet för sökt utbildning, och arbetsförmedlarens beslut att bevilja utbildning, i relation till de

statistiskt beräknade resultaten avseende individens prestation på de olika kognitiva faktorerna i ett antal logistiska regressionsekvationer.

Utfallen visar att psykologerna, och även arbetsförmedlarna, gör olika bedömningar beroende på vilken kognitiv funktion som avses. När det gäller individer med samma nivå på G*c* blir både EI och NEI signifikant oftare bedömda lämpliga, beviljas utbildning i högre grad och utbildning på högre nivå än SNI. När det gäller individer med samma nivå på G*v* finns det ingen signifikant skillnad i andel som beviljas utbildning, eller i nivå på beviljad utbildning mellan grupperna. För EI och SNI finns heller ingen skillnad i andelen som bedöms lämpliga. För NEI bedöms en lägre andel lämpliga. Denna skillnad är signifikant med minsta möjliga marginal (t=1.96). När det gäller G*f* och G*s* är emellertid mönstret det att både EI och NEI, givet samma nivå på dessa faktorer, bedöms lämpliga i signifikant lägre grad än SNI, att NEI beviljas utbildning i signifikant lägre grad, och att både EI och NEI beviljas signifikant lägre nivå på utbildningen. Det finns alltså avsevärda skillnader i hur individer i de olika grupperna bedöms, beroende på vilken begåvningsfaktor som avses.

När begåvningsfaktorerna slås samman och bedöms som en helhet jämnar emellertid dessa effekter i stort ut varandra, så att det inte finns några signifikanta skillnader i bedömd lämplighet mellan SNI och EI, medan NEI däremot bedöms lämpliga i signifikant lägre grad, givet samma nivå på samtliga kognitiva faktorer. När samtliga kognitiva faktorer beaktats finns inga signifikanta skillnader mellan grupperna när det gäller andel beviljad utbildning, medan EI och NEI signifikant oftare beviljas högre nivå på utbildning.

Eftersom den sammantagna effekten av de olika begåvningsfaktorerna resulterar i avsaknad av skillnader mellan grupper i andel som beviljas utbildning, samt något högre nivå på beviljad utbildning för invandrargrupperna, ger användningen av test inga uppenbart negativa praktiska eller sociala konsekvenser för grupperna med invandrarbakgrund. De mönster som framkommer, med positiv bedömning för EI och NEI avseende G*c* och negativ bedömning av G*f* och G*s*, skulle ändå kunna tolkas som att testanvändningen inte blir optimal. Det finns en risk för att kraven på språkfärdighet blir för höga för EI och NEI, medan deras resurser framför allt avseende G*f* inte fullt ut tas tillvara.

Studie III: Immigrant groups, vocational training, and employment

I studie III ligger fokus på utfall i form av anställning. Det finns inget omdöme eller betyg efter avslutad arbetsmarknadsutbildning, men eftersom utbildningen

syftar till ökad möjlighet för individen på arbetsmarknaden kan anställning anses vara ett positivt resultat. Andelen i anställning följs över tre år; året då testningen genomfördes (år noll) och de följande två åren (år ett och två).

Den deskriptiva statistiken visar att andelen i anställning ökar över tid för samtliga grupper, men EI och NEI inleder år noll med en betydligt lägre andel i anställning, som den kraftigare tillväxten för dessa grupper under de två följande åren inte fullt ut lyckas utjämna. Genomgående, med undantag för SNI år noll, har också de individer som beviljats arbetsmarknadsutbildning högre andel i anställning än de som inte beviljats utbildning i samtliga kulturella grupper.

Utvecklingen av andel i anställning över tid sammanfattas i en tillväxtkurva, där en latent variabel för interceptet indikerar anställningsstatus år noll, medan en latent variabel för utvecklingen över år ett och två beskrivs av kurvans lutning (slope). Dessa blir de beroendevariabler som de oberoende variablerna relateras mot. De oberoende variablerna är beviljad arbetsmarknadsutbildning, grupptillhörighet (SNI; EI; NEI), resultat på de kognitiva faktorerna (G*f*, G*c*, G*v*, G*s*, och deras samlade effekt G*all*), respektive nivå på beviljad arbetsmarknadsutbildning. De oberoende variablernas prövas i en rad regressionsekvationer, där variablerna läggs till en och en i taget.

Först prövas effekten av att ha blivit beviljad arbetsmarknadsutbildning. Detta har ingen signifikant effekt på andel i arbete år noll, men bidrar signifikant med en ökning i andel anställda om 7 procent per år under år 1 och 2. Det finns ingen interaktionseffekt mellan beviljad arbetsmarknadsutbildning och tillhörighet till EI eller NEI på anställningsgrad, vilket tolkas som att den positiva effekten av arbetsmarknadsutbildning är i stort densamma för alla grupper.

I likhet med vad som kunde utläsas av den deskriptiva statistiken visar även den modellbaserade utvärderingen att tillhörighet till EI eller NEI har signifikant negativ effekt på anställningsgrad år 0, men samtidigt också kraftfullare ökning i anställningsgrad år 1 och 2. När de kognitiva variablerna förs in ändras denna bild endast marginellt. Det faktum att invandrargrupperna beviljas arbetsmarknadsutbildning med lägre genomsnittliga resultat på de kognitiva funktionerna kan alltså bara till liten del förklara den lägre anställningsgraden för dessa grupper. Även här saknas i stort interaktionseffekter, vilket kan tas till intäkt för att de kognitiva funktionerna uppvisar samma samband med andel i arbete för de olika grupperna.

I Studie III uppmärksammas också den skillnad i effekt som arbetsmarknadsutbildningar på olika nivåer visar. Medan beviljad utbildning på nivå 2 (med förkunskapskrav på grundskolenivå) ger effekt redan år noll

avklingar denna effekt därefter. Utbildning på nivå 3 (med förkunskapskrav på gymnasienivå) ger en negativ effekt på andel i arbete år noll, men därefter en kraftfullt ökad effekt över år ett och två. Beviljad utbildning på nivå 4 (högskolenivå) uppvisar ett mönster som liknar det för nivå 3, men den initiala inlåsningseffekten är större och hinner inte utjämnas under de tre år som studeras.

Sammanfattningsvis pekar studie III mot att beviljad arbetsmarknadsutbildning har en positiv effekt på andel i anställning som inte skiljer sig mellan de tre grupperna. Likaså förefaller de kognitiva funktionerna ha samma prognosförmåga mot anställning för de tre grupperna. Utvecklingen över tid för andel anställda skiljer sig mellan utbildningar på olika nivåer. Eftersom de olika grupperna inte har samma sökmönster (EI och speciellt NEI söker och beviljas generellt utbildningar på lägre nivåer) bör utvärdering av utfall av arbetsmarknadsutbildning i anställningsgrad ta hänsyn till detta mönster.

Diskussion och slutsatser

De olika studierna belyser olika aspekter av validitetsfrågor i relation till testanvändning i invandrargrupper.

I studie I belyses den generella begåvningsfaktorn och hur den utvecklas i relation till olika kulturella grupper, vilket ger bidrag till begreppsvaliditeten för den generella begåvningsfaktorn. Studie I ger även en bild av hur de latenta faktorerna laddar i olika test. Denna bild uppvisar rätt små och förväntade variationer för de olika kulturella grupperna, vilket även detta bidrar till begreppsvaliditeten.

I studie II visas att testutfall tolkas olika, beroende på individens kulturella bakgrund och att detta är relaterat till vilken kognitiv faktor som avses. Även här finns aspekter som bidrar till begreppsvaliditeten, men också en belysning av testningens konsekvenser, i form av beslut om lämplighet respektive beviljad utbildning, som har betydelse för individen. Sammantaget jämnar de positiva och de negativa konsekvenserna ut sig, sett i termer av beviljad utbildning. Givet samma kognitiva nivå beviljas EI och NEI dock utbildningar på något högre nivå jämfört med SNI, vilket möjligen inte är optimalt.

I Studie III visas att beviljad arbetsmarknadsutbildning har en positiv effekt över tid, på andel i arbete, och att denna effekt inte skiljer sig för de olika grupperna. Likaså visas att effekten av de kognitiva faktorerna på andel i arbete är mycket begränsad, men inte heller här finns någon avgörande skillnad mellan grupperna. Dock har EI och NEI en avsevärt lägre nivå avseende andel i arbete.

Denna nivå kan bara till mycket begränsad del förklaras med skillnader i de kognitiva prestationerna.

Sammanfattningsvis pekar de tre studierna på att användandet av kognitiva test med vissa undantag mäter samma funktioner i de olika grupperna. Vidare visas att urvalsproceduren med kognitiva test medför att individer med invandrarbakgrund beviljas utbildning i samma utsträckning som de med svensk bakgrund, och på något högre nivå, givet deras prestationer på de kognitiva variablerna. Även om detta innebär ett positivt utfall är det dock möjligt att det inte är optimalt, givet de krav utbildningen ställer. Till sist visas att andel i anställning visserligen är betydligt lägre i grupperna med invandrarbakgrund, men att detta till mycket liten del kan förklaras av skillnader på kognitiva faktorer, eller differentiell effekt av arbetsmarknadsutbildning.

Det samlade utfallet ger anledning till en positiv syn på testanvändning även för grupper med annan kulturell bakgrund. Användning av test ger i stort samma effekter för personer med olika kulturell bakgrund. Testresultat kan därtill ha ett värde för personer som saknar betyg eller andra formella dokument att styrka sina meriter. Testens användbarhet skulle möjligen öka ytterligare om värden på latenta faktorer användes genomgående. Dock bör resultaten prövas med andra grupper.

Begränsningar i studierna ligger bl.a. i undersökningsgruppens sammansättning, som inte är styrd av teoretiska överväganden. Det visar sig t.ex. i den kraftigt manliga slagsidan. Vidare hade skarpare slutsatser om testens prognoskraft och utfallet av arbetsmarknadsutbildning kunnat göras, om det funnits någon form av systematisk utvärdering efter avslutad utbildning. Sambanden mellan kompetens, så som den mäts genom kognitiva test och anställning, har här förutsatts vara linjära. Det är möjligt att det finns vissa tröskelvärden som behöver överskridas för att tydligare samband ska kunna ses. Uppföljning med andra grupper och med tydligare utfallsvariabler behövs för att ytterligare undersöka sambanden mellan testanvändning och invandrade gruppers möjligheter i arbetsmarknadsutbildning och anställning.

References

Ackerman, P. L., & Beier, M. E. (2006). Determinants of domain knowledge and independent study learning in an adult sample. *Journal of Educational Psychology, 98*(2), 336-381.

Ackerman, P. L., & Lohman, D. F. (2006). Individual differences in cognitive functions. In P. A. Alexander, P. R. Pintrich, & P. H. Winne (Eds.), Handbook of Educational Psychology, 2nd Edition (pp. 139-161). Mahwah, NJ: Lawrence Erlbaum Associates.

Adda, J., Costa Dias, M., Meghir, C., & Sianesi, B. (2007). Labour market programmes and labour market outcomes: a study of the Swedish active labour market interventions [Electronic version]. *IFAU Working Paper Series,* 27*,* 71. Retrieved January 28, 2009.

American Education Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testsand Manuals.* Washington, DC: Author.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques [supplement]. *Psychological Bulletin, 51*(2, Pt 2), 201-238.

APU-handboken. (2009). Part 1, Ch 1, *APU I praktiken*. Ch 2, *Yrkesvalsteorier*. Del 2, *Begåvning, teori och mätning*. Retrived 120511 from http://vis.arbetsformedlingen.se/arbetsformedling/verktyg/metoder/arbetspsykologiskutredningapuq.4.710196211278fb18bae80007879.html

Arbetsförmedlingen. (2008). *Riktad arbetspsykologisk utredning RA (Q): Metodbeskrivning.* Publication retrieved May 10, 2009, from https://prod.vis.ams.se/Default.aspx?a=4120)

Arbetsförmedlingen. (2010). *When you have your residence permit - come to Arbetsförmedlingen.* Retrieved May 23, 2012 from http://www.arbetsformedlingen.se/download/18.7cab701e12c9dc4a47c800044/uppehallstillstand-eng.pdf

Arbetsförmedlingen. (2011). *Faktablad till arbetssökande november 2011. Arbetsmarknadsutbildning.* Retrieved May 23, 2012 from http://www.arbetsformedlingen.se/download/18.46ccfec5127ddccec778000356/amu.pdf

Ardila, A., Bertolucci, P. H., Braga, L. W., Castro-Caldas, A., Judd, T., Kosmidis, M. H., Matute, E., Nitrini, R., Ostrosky-Solis, F., Rosselli, M. (2010). Illiteracy: The neuropsychology of cognition without reading. *Archives of Clinical Neuropsychology, 25,* 689–712.

Aronsson, G., Hellgren, J., Isaksson, K., Johansson, G., Sverke, M., & Torbjörn, I. (2012). *Arbets- och organisationspsykologi. Individ och organisation i samspel.* Stockholm, Natur & Kultur.

Ashton, M. C., & Lee, K. (2005). Problems with the method of correlated vectors. *Intelligence, 33*, 431-444.

Beauducel, A. & Herzberg, P. H. (2006). On the performance of Maximum Likelihood versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA. *Structural Equation Modeling: A Multidisciplinary Journal,* 13(2), 186-203.

Behtoui, A. (2004). Unequal opportunities for young people with immigrant backgrounds in the Swedish labour market. *Labour, 18*(4), 633-660.

Blair, C. (2006). How similar are fluid cognition and general intelligence? A developmental neuroscience perspective on fluid cognition as an aspect of human cognitive ability. *Behavioral and Brain Sciences 29*, 109-160.

Borén, M. (1995). RA (Riktad arbetspsykologisk utredning för fordonsmekaniker: del II. In Arbetsmarknadsstyrelsen (Ed.), Serie V/Vägledningsenheten, (Vol. 7). Solna: AMS

Borén, M. (1999). Riktad arbetspsykologisk utredning. Nätverkstekniker. Malmö *Fmpsy 1999:2*. Arbetsmarknadsstyrelsen.

Bornstein, R. F. (2011). Toward a process-focused model of test score validity: Improving psychological assessment in science and practice. *Psychological Assessment, 23*(2), 532-544.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061-1071.

Brigham, C. C. (1923). *A Study of American Intelligence*. Princeton: Princeton University Press. Retrieved June 2012 from http://archive.org/details/studyofamericani00briguoft

Byrne, B. M., Leong, F. T. L., Hambleton, R. K., Oakland, T., van der Vijver, F. J. R., Cheung, F. M., & Bartram, D. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology,3*(2), 94-105.

Carneiro, P., Heckman, J. J., & Masterov, D. v. (2005). Labor market discrimination and racial differences in premarket factors. *Journal of Law and Economics, 48*(1), 1-39.

Carroll, J. B. (1993). *Human Cognitive Abilities. A survey of factor-analytic studies*. Cambridge: University Press.

Carroll, J. B. (1997).Psychometrics, intelligence, and public perception. *Intelligence, 24*(1), 25-52.

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

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

Civil Rights Act. (1964). Retrieved December 20, 2012 from http://www .eeoc.gov/laws/statutes/titlevii.cfm

Cliffordson, C. & Gustafsson, J.-E. (2008). Effects of age and schooling on intellectual performance: Estimates obtained from analysis of continuous variation in age and length of schooling. *Intelligence, 36*(2), 143-152.

Code of Federal Regulations. (1978). *Part 1607—Uniform Guidelines on Employee Selection Procedures (1978)*. Retrieved January 2013, from http://www.gpo.gov/fdsys/pkg/CFR-2012-title29-vol4/xml/CFR-2012-title29-vol4-part1607.xml

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. NY: Academic Press.

Cook, M. (1993). *Personnel Selection & Productivity* (2 ed.). Sussex, UK: John Wiley & Sons.

Cook, M. (2009). *Personnel Selection. Adding Value Through People*. (5 ed.). Sussex, UK: John Wiley & Sons.

Council of the International Test Comission. (2000). *International Guidelines for Test Use, (Version 2000)*. Retrieved May 5, 2012 from http://www.intestcom.org/Guidelines/Test+Use.php

Crocker, L. (1997). Editorial: The great validity debate. *Educational Measurement: Issues and Practice, 16*(2), 4.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological test. *Psychological Bulletin, 52*(4), 281-302.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp 443-507). Washington, DC: American Council of Education.

Cureton, E.E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational Measurement* (pp. 621-694). Washington, DC: American Council on Edication.

de Luna, X., Forslund, A., & Liljeberg, L. (2008). Effekter av yrkesinriktad arbetsmarknadsutbildning för deltagare under perioden 2002–04 [Electronic Version]. *IFAU Working Paper Series*, 1, 53. Retrieved June 18, 2008.

De Meijer, A. L., Born, M. P., Terlouw, G., & van der Molen, H. T. (2008). Criterion-related validity of Dutch police-selection. Measures and differences between ethnic groups. *International Journal of Selection and Assessment, 16*(4), 321-332.

Deary, I. L., Lawn, M., & Bartholomew, D. J. (2008). A conversation between Charles Spearman, Godfrey Thomson, and Edward L. Thorndike: The International Examinations Inquiry Meetings 1931–1938. *History of Psychology, 11*(2), 122–142.

Dickens, W. T. & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review 108* (2), 346-369.

Dolan, C. V. (2010). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research, 35*(1), 21-50.

Dolan, C. V., Roorda, W., & Wicherts, J. M. (2004). Two failures of Spearman*s hypothesis: The GATB in Holland and the JAT in South Africa. *Intelligence, 32*, 155-173.

Emanuelsson, I., Reuterberg, S.-E., & Svensson, a. (1993). Changing Differences in Intelligence? Comparisons of groups of 13-year-olds tested from 1960 to 1999. *Scandinavian Journal of Educational Research, 37*(4), 259-277.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179-197.

Eriksson, F. (2011). *Utrikes födda på den svenska arbetsmarknaden. Långtidsutredningen 2011, Bilaga 4.* Retrieved June 6, 2012 from http://www.regeringen.se/sb/d/12401.

Equality Act, 2010. Retrieved August 2012 from http://www.homeoffice.gov.uk/equalities/equality-act/

Fagan, J. F., & Holland, C. R. (2007). Racial equality in intelligence: Predictions from a theory of intelligence as processing. *Intelligence, 35,* 319-334.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin 95*(1), 29-51.

Flynn, J. R. (1987). Massive IQ gains in 14 Nations: What IQ tests really measure. *Psychological Bulletin, 101*(2), 171-191.

Freidlitz, H. & Willquist-Gustavsson, I. (1996). Riktad arbetspsykologisk utredning inför val av industriteknisk utbildning. In Arbetsmarknadsstyrelsen (Ed.), *Vra, 9921681680* (Vol. 2). Solna: AMS.

Freidlitz, H. & Willquist-Gustavsson, I. (1997). Riktad arbetspsykologisk utredning inför val av industriteknisk utbildning: automationsteknik, CNC-operatör, svetsning, verktygstillverkning. In Arbetsmarknadsstyrelsen (Ed.), *Vra, 9921681680* (Vol. 2). Solna: AMS.

Freidlitz, H. & Willquist-Gustavsson, I. (2000a). Riktad arbetspsykologisk utredning. Nätverkstekniker. *Rapporter och publikationer. Fmpsy 2000:3.* Arbetsförmedlingen.

Freidlitz, H. & Willquist-Gustavsson, I. (2000b). Riktad arbetspsykologisk utredning. Verktygsmakarbedömningar. *Rapporter och publikationer. Fmpsy 2000:5.* Arbetsförmedlingen.

Freidlitz, H. & Willquist-Gustavsson, I. (2001). Arbetsförmedlingen

Rapporer och publikationer. Fmpsy 2000:5. Riktad arbetspsykologisk utredning. Bussförare. *Rapporter och publikationer. Fmpsy 2001:4.* Arbetsförmedlingen.

Frisby, C. L. (1999a). Culture and test session behavior: Part I. *School Psychology Quarterly, 14*(3), 263-280.

Frisby, C. L. (1999b). Culture and test session behavior: Part II. *School Psychology Quarterly, 14*(3), 281-303.

Förordning (1994:895) om svenskundervisning för invandrare. (1994). Retrieved May 24, 2012, from http://www.riksdagen.se/sv/Dokument-Lagar/Lagar/Svenskforfattningssamling/Forordning-1994895-om-svens_sfs-1994-895/

Förordning (2000:634) om arbetsmarknadspolitiska program. Retrieved July 11, 2011 from http:// http://www.riksdagen.se/sv/Dokument-Lagar/Lagar/ Svenskforfattnings samling/Forordning-2000634-om-arbet_sfs-2000-634/?bet=2000:634

Gagnerud, S. & Haglund, B. (2005). Den hierarkiska begåvningsmodellen. En sammanfattande beskrivning av den hierarkiska begåvningsmodellen i samband med sådan intelligens- och begåvningstestning som har faktoranalytisk forskning som grund. *Fmpsy 2005:1*. Arbetsmarknadsstyrelsen.

Gaines, S. O., Bagha, S., Barrie, M., Bhattacharjee, T., Boateng, Y., Briggs, J., …Rodriques, L. (2012). Impact of experiences with racism on African-descent persons' Susceptibility to stereotype threat within the United Kingdom. *Journal of Black Psychology, 38*(2), 135-152.

Gamliel, E. & Cahan, S. (2007). Mind the Gap: Between-group differences and fair test use. *International Journal of Selection and Assessment, 15*(3), 273-282.

Gardner, H. 1983. *Frames of mind: The Theory of Multiple Intelligences.* New York: Basic Books.

Gardner, H. 1998. Are there additional intelligences? The case for naturalist, spiritual, and existential intelligences. In J. Kane (Ed.), *Education, Information, and Transformation.* Englewood Cliffs, NJ: Prentice-Hall

Goddard, H. H. (1917). Mental tests and the immigrant. *The Journal of Delinquency, 2*(5), 243-277. Retrieved January 8, 2012, from http://harpending.humanevo.utah.edu/Documents/goddard.html

Gottfredson, L. S. (1997). Mainstream Science on Intelligence. *Intelligence, 24*(1) 13-23.

Gottfredson, L. S. (2000). Skills gaps, not tests, make racial proportionality impossible. *Psychology, Public Policy, and Law, 6*(1), 129-143.

Gottfredson, L. S. (2005). What if the hereditarian hypothesis is true? *Psychology, Public Policy, and Law 11*(2), 311–319.

Gould, S. J. (1996). *The mismeasure of man*. New York: Norton.

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: the clinical-statistical controversy. *Psychology, Public Policy, and Law, 2*(2), 293-323.

Guenole, N., Englert, P. & Tyler, P. J. (2003). Ethnic group differences in cognitive ability test scores within a New Zealand applicant sample. *New Zealand Journal of Psychology, Vol 32*(1), Jun 2003, 49-54.

Guilford, J. P., & Hoepfner, R. (1971). *The analysis of intelligence.* New York: McMillan.

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. Sternberg (Ed.), *Advances in the psychology of human intelligence,* (Vol 4, pp 35-71): Erlbaum

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

Gustafsson, J.-E. (2002). Measurement from a hierarchical point of view. In H. I. Braun, D. N. Jackson, & D. E. Wiley (Eds.), *The role of constructs in psychological and educational measurement* (pp. 73-95). London: Lawrence Erlbaum Associates, Publishers.

Gustafsson, J.-E. (2008). Effects of international comparative studies on educational quality on the quality of educational research. *European Educational Research Journal, 7*(1), 1-17.

Gustafsson, J.-E., & Stahl, P.-A. (2005). *STREAMS 3.0 User's Guide.* Mölndal, MultivariateWare.

Halsted, M. J. (2004). An islamic concept of education. *Comparative Education. 40*(4), 517-527.

Haglund, B. (1987). *Normkontroll av testen i Arbetspsykologisk utredning.* Stockholm: Vägledningsenheten AMS.

Halpern, D. F. (2000). Validity, fairness, and group differences. Tough questions for selection testing. *Psychology, Public Policy, and Law, 6*(1), 56-62.

Heckman, J. J., Stixrud, J. & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. NBER *Working Paper Series.* Retrieved March 12, 2008, from http://www.nber.org/papers/w12006

Helms-Lorenz, M., Van de Vijver, F. J. R., & Poortinga, Y. H. (2003). Cross-cultural differences in cognitive performance and Spearman's hypothesis: *g* or *c*? *Intelligence, 31*, 9-29.

Herrnstein, R. J., & Murray, C. (1994). *The Bell Curve: intelligence and Class Structure in American Life.* New York: the Free Press.

Higuera, L. A.-Z. (2001). Adverse impact in personnel selection: The legal framework and test bias. *European Psychologist, 6*(2), 103-113.

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology, 57*(5), 253-270.

Hülsheger, U. R., Maier, G. W., & Stump, T. (2007). Validity of general mental ability for the prediction of job performance and training success in Germany: A meta-analysis. *International Journal of Selection and Assessment, 15*(1), 3-18.

Hunter, J. E. (1986). Cognitive ability, cognitive aptitude, job knowledge, and job performance. *Journal of Vocational Behavior, 29*(3), 340-362.

Integration report. (2005). *Summary and Agenda for Integration and Diversity.*Integrationsverket, Norrköping.

Integrationsverket. (2003), *Rapport Integration 2002.* Retrieved from http://www.mkc.botkyrka.se/biblioteket/Publikationer/Integration_2002.pdf

Integrationsverket. (2004). *Rapport Integration 2003.* Retrieved january 2013 from http://www.temaasyl.se/Documents/%C3%96vrigt/helarapportintegration.pdf

Integrationsverket. (2007). *Integrationspolitikens resultat – På väg mot ett samlat system för uppföljning och analys vid 16 statliga myndigheter.* (2007). Retrieved from http://mkc.botkyrka.se/bibliotek/integrationsverkets-publikationer/2007-2

Johnson, W., & Bouchard, T. Jr (2005). The structure of intelligence: it is verbal, perceptual and image rotation (VPR), not fluid and crystallized. *Intelligence*, *33*(4), 393-436.

Jensen, A. R. (1986). *g*: Artifact or Reality? *Journal of Vocational Behavior 29*, 301-331.

Jensen, A. R. (1998). *The g factor. The science of mental ability.* Westport, Connecticut: Praeger.

Jensen, A. R. (2000). Testing. The dilemmma of group differences. *Psychology, Public Policy, and Law 6*(1), 121-127.

Jensen, A. R. (2008). Book review of Howard Gardner under fire: The rebel psychologist faces his critics. *Intelligence, 36,* 96-97.

Järnefors, E. B. (2012). Mer arbetspsykologi på Arbetsförmedlingen. *Psykologtidningen*(4), 14-15.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (pp. 17-64). Westport, CT: American Council on Education/Praeger.

Kane, M. T. (2008). Terminology, emphasis, and utility in validation. *Educational Researcher, 37*(2), 76-82.

Kapuściński, R. (2003). Ebenholtz (A. Bondegård, Trans. 2 ed.) Viborg: Nörhaven Paperback A /S.

Keith, T. Z. (2005). Using confirmatory factor analysis to aid in understanding the constructs measured by intelligence tests. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment:Theories, tests and issues* (pp. 581-614). (2nd edition),. New York, NY: Guilford Press.

Kesler, C. (2006). Social policy and immigrant joblessness in Britain, Germany and Sweden. *Social Forces, 85*(2).

Lemaître, G. (2007). The integration of immigrants into the labour market: the case of Sweden (Publication. Retrieved March 16, 2009, from Directorate for Employment, Labour and Social Affairs: http://www.oecd.org/els/workingpapers.)

Loehlin, J. C. (2004). *Latent Variable Models. An Introduction to Factor, Path, and Structural Equation Analysis* (4 ed.). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Lynn, R. (1987). The intelligence of the Mongoloids: A psychometric, evolutionary and neurological theory. *Personality and Individual Differences, 8*(6), 813-844.

Lynn, R. (1990). The role of nutrition in secular increases in intelligence. *Personality and Individual Differences 11*(3), 273-285.

Lynn, R., & Owen, K. (1994). Spearman´s hypothesis and test score differences between Whites, Indians, and Blacks in South Africa. *The Journal of General Psychology, 121*(1), 27-36.

Lynn, R., & Vanhanen, T. (2002). IQ and the wealth of nations. Westport, CT: Praeger.

Martinson, S., & Lundin, M. (2003). Vikten av arbetsgivarkontakter: en studie av den yrkesinriktade arbetsmarknadsutbildningen i ljuset av 70-procentsmålet [electronic version]. *IFAU Working Paper Series*, 10, 65. Retrieved January 28, 2009.

Maruyama, M. (1999). Self-heterogenization and cultural milieu selection: Two new directions in counseling. In P. Pedersen (ed.). *Multiculturalism as a Fourth Force* (pp. 37-72). Philadlephia, PA. Brunner/Mazel, Taylor & Francis Group.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities. Past, present and future. In D. P. Flanagan, & P. L. Harrison (Eds.) *Contemporary intellectual assessment: Theories, tests and issues (2nd ed.)*, (pp. 136-181). New York, NY: Guilford Press.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35*(11). 1012-1027.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd* ed., pp. 13-103) Washington, DC: American Council on Education, MacMillan.

Messick, S. (1993). *Keynote address.* Paper presented at the second European Conference on Psychology of the European Association of Psychological Assessment, Groningen, August 25-27, also at Symposium "Psykologisk mätning och bedömning – Psychological Assessment", Stockholm, June 7-8, 1994.

Moss, P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, Vol. 62*(3), 229-258.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education 30*(January), 109-162.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide* (Third Edition), Los Angeles, CA: Muthén & Muthén.

Myrberg, M. (2001). *Att Öppna Språkgränser: Klyftor och broar i vuxna invandrares läs- och skrivutveckling.* Stockholm: Skolverket.

Mårtensson, U. (2009). *Sfi i förändring. En studie ur ett lärarperspektiv. Examensarbete, Lärarutbildningen.* Malmö Högskola, Malmö.

Nell, V. (2000). *Cross-Cultural Neuropsychological Assessment: Theory and Practise.* Mahwah, New Jersey: Lawrence Erlbaum.

Neisser, U., Boodoo, G., Bouchard, J. T., Boykin, A. W., Brody, N., Ceci, S. J., …Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*(2), 77-101.

Nisbett, R. E., Choi, I., Peng, K., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review, 108*(2), 291-310.

Nisbett, R. E. (2005). Heredity, environment, and race differences in IQ. A commentary on Rushton and Jensen (2005). *Psychology, Public Policy, and Law, 11*(2), 302-310.

Nisbett; R. E.; Aronson, J.; and Blair, C.; Dickens, W.; Flynn, J.; Halpern, D. F.; Turkheimer, E. (2012). Intelligence. New Findings and Theoretical Developments. *American Psychologist, 67*(2), 130-159.

OECD. (2004). *The Integration of Immigrants Into the Labour Market: The Case of Sweden.* Organization for Economic Co-operation and Development

Petersson, K. M., Reis, A., Ingvar, M. (2001). Cognitive processing in literate and illiterate subjects: A review of some recent behavioral and functional neuroimaging data. *Scandinavian Journal of Psychology, 42*(3), 251-267.

Plomin, R., Pedersen, N. L., Lichtenstein, P., & McClearn, G. E. (1994). Variability and stability in cognitive abilities are largely genetic later in life. *Behavior Genetics, 24*(3), 207-215.

Pocket facts – Statistics on Integration. (2006). Retrieved from http://www.mkc.botkyrka.se/biblioteket/Publikationer/Pocketfacts.pdf

Popham, J. W. (1997). Consequential validity: Right concern-Wrong concept. *Educational Measurement: Issues and Practice, 16*(2), 9-13, 24.

Potosky, D., Bobko, P., & Roth, P. L. (2005). Forming composites of cognitive ability and alternative measures to predict job performance and reduce adverse impact: Corrected estimates and realistic expectations. *International Journal of selection and Assessment, 13*(4), 304-315.

Prop. 1975:26. *Regeringens proposition om riktlinjer för invandrar- och minoritetspolitiken m. m.* Retrived 130107 from http://www.riksdagen.se/sv/Dokument-Lagar/Forslag/Propositioner-och-skrivelser/Regeringens-proposition-om-rik_FY0326/?text=true

Prop. 1983/84:199. *Om svenskundervisning för vuxna invandrare.* Retrieved 120524 from http://www.riksdagen.se/sv/Dokument-Lagar/Forslag/Propositioner-och-skrivelser/om-svenskundervisning-for-vuxn_G703199/

Prop. 1997/98:16. *Sverige, framtiden och mångfalden - från invandrarpolitik till integrationspolitik.* Retrieved June 6, 2012 from http://www.riksdagen.se/sv/Dokument-Lagar/Utskottens-dokument/Betankanden/Sverige-framtiden-och-mangfal_GL01SfU6/

Raven, J. C.; Raven, J., & Court, J. H. (2000). *Raven manual: Section 3. Standard Progressive Matrices*. Oxford, Oxford Psychologists Press.

Reinolds, M. R., & Keith, T. Z. (2007). Spearman's law of diminishing returns in hierarchical models of intelligence for children and adolescents. *Intelligence, 35*(1), 267-281.

Richardson, K., & van den Berg, G. J. (2008). Duration dependence versus unobserved heterogeneity in treatment effects: Swedish labor market training and the transition rate to employment [Electronic Version]. *IFAU Working Paper Series*, 7, 54. Retrieved January 28, 2009.

Roth, P. L., Bevier, C. A., Bobko, P., Switzer_III, F. S., & Tyler, P. (2001). Ethnic group differences in cognitive ability in employment and educational settings: A meta-analysis. *Personnel Psychology, 54*(2), 297-330.

Roth, P. L., Buster, M. A., & Bobko, P. (2011). Updating the trainability tests literature on Black–White subgroup

differences and reconsidering criterion-related validity. *Journal of Applied Psychology, 96*(1), 34-45.

Roth, P. L., Huffcutt, A. I., & Bobko, P. (2003). Ethnic group differences in measures of job performance: A new meta-analysis. *Journal of Applied Psychology, 88*(4), 694-706.

Rotundo, M.,& Sackett, P. R. (1999). Effect of rater race on conclusions regarding differential prediction in cognitive ability tests. *Journal of Applied Psychology, 84*(5), 815-822.

Rushton, J. P. (1998). The "Jensen Effect" and the "Spearman–Jensen hypothesis" of Black–White IQ differences. *Intelligence, 26*(3), 217–225.

Rushton, P. J. (2004). Placing intelligence into an evolutionary framework or how g fits into the r–K matrix of life-history traits including longevity. *Intelligence, 32,*321-328.

Rushton, P. J., & Jensen, A. (2003). African-White IQ differences from Zimbabwe on the Wechsler Intelligence Scale for Children-Revised are mainly on the *g* factor. *Personality and Individual Differences, 34,* 177-183.

Rushton, P. J., & Jensen, A. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law, 11*(2), 235-294.

Rushton, P. J., & Skuy, M. (2000). Performance on Raven's Matrices by African and White university students in South Africa. *Intelligence, 28*(4), 251-265.

Rushton, P. J., Skuy, M., Bons, T. A. (2004). Construct validity of Raven´s Advanced Progressive Matrices for African and non-African engineering students in South Africa. *International Journal of Selection and Assessment, 12*, 220-229.

Sackett, P. R., Borneman, M. J., & Connely, B. S. (2008). High-Stakes testing in higher education and employment: Appraising the evidence for validity and fairness. *American Psychologist, 63*(4), 215-277.

Sackett, P. R., Laczo, R. M., & Lippe, Z. P. (2003). Differential prediction and the use of multiple predictors; The omitted variables problem. *Journal of Applied Psychology, 88*(6), 1046-1056.

Sackett, P. R., Schmitt, N., Ellingson, J. E., & Kabin, M. B. (2001). High-stakes testing in employment, credentialing, and higher education. Prospects in a post-affirmative-action world. *American Psychologist, 56*(4), 302-318.

Salgado, J. F., & Anderson, N. (2002). Cognitive and GMA testing in the European Community: Issues and evidence. *Human Performance, 15*(1/2), 75-96,

Salgado, J. F., & Anderson, N. (2003), Validity generalization of GMA tests across countries in the European Community, *European Journal of Work and Organizational Psychology, 12*(1), 1-17.

Salgado, J. F., Anderson, N., Moscosco, S., Bertua, C., & de Fruyt, F. (2003). International Validity Generalization of GMA and Cognitive Abilities: A European Meta-Analysis. *Personnel Psychology, 56,* 573-605.

Salgado, J. F., Anderson, N., Moscosco, S., Bertua, C., Fruyt, F. d., & Rolland, J. P. (2003). A meta-analytic study of general mental ability. Validity for different occupations in the European Community. *Journal of Applied Psychology, 88*(6), 1068-1081.

Schafer, J. L., & Graham, J. W. (2002). Missing Data: Our view of the state of the art. *Psychological Methods, 7*(2), 147-177.

Schmidt, F. L., (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance, 15*(1/2), 187-210.

Schmidt, F. L., & Hunter, J. (1998). The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin, 124*, 262-274.

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86*(1), 162-173.

Schmidt, F. L., Pearlman, K., & Hunter, J. E. (1980). The validity and fairness of employment and educational tests for Hispanic Americans: A review and analysis. *Personnel Psychology, 33*, 705-724.

Schmidt, F. L., Shaffer, J. A., & Oh, I.-S. (2008). Increased accuracy for range restriction corrections: Implications for the role of personality and general mental ability in job and training performance. *Personnel Psychology, 61,* 827-868.

Segendorf, Å. O., & Teljusto, T. (2011). *Sysselsättning för invandrare – en ESO-rapport om arbetsmarknadsintegration.* Retrieved from http://www.eso.expertgrupp.se/Uploads/Documents/2011-5-till-webben.pdf

Seibt, B., & Förster, J. (2004). Stereotype threat and performance: How self-stereotypes influence processing by inducing regulatory foci. *Journal of Personality and Social Psychology, 87*(1), 38-56.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5-8, 13, 24.

Sjögren, A., & Zenou, Y. (2007). *Vad förklarar invandrarens integration på arbetsmarknaden – en teoriöversikt.* Retrived May 21, 2012. from http://www.mkc.botkyrka.se/biblioteket/Publikationer/Stencilserie/2007_09Teorioversikt.pdf

Skolinspektionen. (2010). *Kvalitetsgranskning. Svenskundervisning för invandrare (sfi) - en granskning av hur utbildningen formas efter deltagarnas förutsättningar och mål.* Retrieved May 24, 2012 from http://www.skolinspektionen.se/Documents/Kvalitetsgranskning/sfi/webb-slutrapport-sfi.pdf

Skolverket. (2009). *Förordning om kursplan för svenskundervisning för invandrare. SKOLFS 2009:2.* Retrieved May 19, 2012.from http://www.skolverket.se/skolfs?id=1492

Skolverket. (2011). *Fler än 100 000 elever i sfi*. Retrieved November 19, 2012, from http://www.skolverket.se/statistik-och-analys/statistik/2.4402/2.4514/fler-an-100-000-elever-i-sfi-1.178149

Skuy, M., Gewer, A., Osrin, Y., Khunou, D., Fridjohn, P., & Rushton, J. P. (2002). *Intelligence, 30*, 221-232.

Sohlman, Å. (2006). *Arbetsmarknadspolitiska åtgärder för integration, expertbilaga till Rapport Integration 2005*. Retrieved from http://www.mkc.botkyrka.se/biblioteket/Publikationer/ri2005/bilagor/2006-507.pdf.

Spearman, C. (1904). "General intelligence", objectively determined and measured. *American Journal of Psychology, 15*, 201-292.

Spearman, C. (1923). *The Nature of 'Intelligence' and the Principles of Cognition*. Oxford, England: MacMillan.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York: Macmillan

Spearman, C. (1932). *The Abilities of Man: Their Nature and Measurement* (2 ed.). London: Macmillan.

Spearman, C. (1939). Thurstone's work reworked. *Journal of Educational Psychology, 30*, 1-16.

Statskontoret. 2012:28. *Kostnader för arbetsmarknadsutbildning och yrkesvux – en jämförelse*. Retrieved December 12. 2012 from http://www.statskontoret.se/upload/Publikationer/2012/201228.pdf

Stauffer, J. M., & Buckley, M. R. (2005). The existence and nature of racial bias in supervisory ratings. *Journal of Applied Psychology, 90*(3), 586-591.

Steele, C. M. (1997). A threat in the air. How stereotypes shape intellectual identity and performance. *American Psychologist, 52*(6), 613-629.

Sternberg, R . (2004). Culture and intelligence. APA Presidential Address. *American Psychologist*, 59(5), 325–338.

Sternberg, R. (2005). There are no public-policy implications. A reply to Rushton and Jensen. *Psychology, Public Policy, and Law, 11*(2), 295–301.

Sternberg, R., & Kaufman, J. C. (1998). Human abilities. *Annual Review of Psychology, 49*, 479-502.

Sue, S. & Okazaki, S. (1990). Asian-American educational achievements. A phenomenon in search of an explanation. *American Psychologist, 45*(8), 913- 920.

Suzuki, L., & Aronson, J. (2005). The cultural malleability of intelligence and its impact on the racial/ethnic hierarchy. *Psychology, Public Policy, and Law, 11*(2), 320-327.

Svensk författningssamling 2007:1030. *Förordning (2007:1030) med instruktion för Arbetsförmedlingen*. Retrived March 9, 2009 from http://www.riksdagen.se/sv/Dokument-Lagar/Lagar/Svenskforfattningssamling/Forordning-20071030-med-ins_sfs-2007-1030/

Svensk författningssamling 2010:197. *Lag (2010:197) om etableringsinsatser för vissa nyanlända invandrare*. Retrieved November 12, 2012, from http://www.riksdagen.se/sv/Dokument-Lagar/Lagar/Svenskforfattningssamling/Lag-2010197-om-etableringsi_sfs-2010-197/

Sveriges Psykologförbund. (2000). *Guidelines on Test Use: Swedish Version*. (2000). (Sveriges Psykologförbund, Trans. 2 ed): International Test Comission. Retrieved January 2013 from http://www.psykologforbundet.se/yrket/Tester/STP%20Riktlinjer%20för%20testpolicy%20i%20organisationer.pdf

Te Nijenhuis, J., de Jong, M.-J., Evers, A., & van der Flier, H. (2004). Are cognitive differences between immigrant and majority groups diminishing? *European Journal of Personality, 18,* 405-434.

Te Nijenhuis, J., & van der Flier, H. (1997). Comparability of GATB scores for immigrant and majority group members: Some Dutch findings. *Journal of Applied Psychology, 82*(5), 675-687.

Te Nijenhuis, J., & van der Flier, H. (1999). Bias research in the Netherlands: Review and implications. *European Journal of Psychological Assessment, 15*(2), 165-175.

Te Nijenhuis, J., & van der Flier, H. (2000). Differential prediction of immigrant versus majority group training performance using cognitive ability and personality measures. *International Journal of Selection and Assessment, 8*(2), 54-60.

Te Nijenhuis, J., & van der Flier, H. (2003). Immigrant–majority group differences in cognitive performance: Jensen effects, cultural effects, or both? *Intelligence, 31*, 443-459.

Tenopyr, M. L. (1996). The Complex Interaction Between Measurement and National Employment Policy. *Psychology, Public Policy, and Law, 2*(2), 348-362.

Thorndike, R. L. (1985). The central role of general ability in prediction. *Multivariate Behavioral Research, 20*, 241-254.

Thurstone, L. L. (1938). Primary Mental Abilities. *Psychometric monograph, 1*

Thurstone, L. L. (1947). *Multiple factor analysis: A development and expansion of The Vectors of Mind.* Chicago: University of Chicago Press.

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

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

U. S Equal Employment Opportunity Commission. (2007). *Employment Tests and Selection Procedures.* Retrieved January 2013, from http://www.eeoc.gov/policy/docs/factemployment_procedures.html

Valentin Kvist, A. (1992). RA (Riktad arbetspsykologisk utredning – RA: en metod att kartlägga individens möjligheter inför ett utbildningsbeslut. In Arbetsmarknadsstyrelsen (Ed.), *Serie V/Vägledningsenheten,* (Vol. 3). Solna: AMS

Valentin Kvist, A. (1995a). RA (Riktad arbetspsykologisk utredning för el-tele-tekniker: del II. In Arbetsmarknadsstyrelsen (Ed.), *Serie V/Vägledningsenheten,* (Vol. 5). Solna: AMS.

Valentin Kvist, A. (1995b). RA (Riktad arbetspsykologisk utredning för fordonsmekaniker: del I. In Arbetsmarknadsstyrelsen (Ed.), *Serie V/Vägledningsenheten,* (Vol. 6). Solna: AMS.

Valentin Kvist, A., Book, C., Borén, M., Danielsson, K.-G., Grenner, M., Meijer, L. (1995). RA (Riktad arbetspsykologisk utredning för el-tele-tekniker: del II. In Arbetsmarknadsstyrelsen (Ed.), *Serie V/Vägledningsenheten,* (Vol. 5). Solna: AMS.

Vernon, P. E. (1950). *The Structure of Human Abilities.* London: Methuen.

Wechsler, D. (1958). The measurement and appraisal of adult intelligence (4th ed), Baltimore, Willliams & Wilkins.

Wittmann, W. W., & Süß, H.-M. (1999). Brunswik symmetry. Investigating the paths between working memory, intelligence, knowledge, and complex problem-solving performances via Brunswik Symmetry. In P. L. Ackerman, P. C. Kyllonen, & R. D. Roberts (Eds.).

Learning and individual differences: Process, trait, and content determinants (pp 77-108). Washington, DC: American Psychological Association.

Woodcock, A., Hernandez, P. R., Estrada, M., & Schultz, P. W. (2012, July 2). The consequences of chronic stereotype threat: Domain Disidentification and Abandonment. *Journal of Personality and Social Psychology*. Advance online publication. doi:10.1037/a0029120