Group differences in intelligence test performance

Wicherts, J.M.

Publication date
2007
Document Version
Final published version

Group differences in intelligence test performance

Jelte M. Wicherts

ISBN/EAN 978-90-9021622-5

Printed by Printpartners Ipskamp, Enschede
This project was supported by a grant from the Netherlands Organization for Scientific Research (NWO) awarded to Conor Dolan.

Group differences in intelligence test performance

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. J.W. Zwemmer
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit
op dinsdag 6 maart 2007, te 14.00 uur

door
Jelte Michiel Wicherts
geboren te Amersfoort

Promotiecommissie

Promotor:	Prof. Dr. H.L.J. van der Maas
Co-promotor:	Dr. C.V. Dolan
Overige leden:	Prof. Dr. D.I. Boomsma Dr. A.V.A.M. Evers Prof. Dr. H. Kelderman Prof. Dr. G.J. Mellenbergh Prof. Dr. D.H.J. Wigboldus

Faculteit der Maatschappij- en Gedragswetenschappen
Universiteit van Amsterdam
Afdeling Psychologie

Contents

1

Introduction

1.1 The Different Types of Psychometrics

During his PhD project the author of this thesis learned that there exist at least three different types of psychometrics. To avoid any confusion, I will introduce these three types of psychometrics, highlight their differences and similarities, and discuss the role the current thesis may play in bridging the gap between two of these types.

The first and oldest type of psychometrics, also known as psychometry (or *psychometrie* in Dutch), is defined as "the psychic ability in which the user is able to relate details about the past condition of an object, usually by being in close contact with it" (Wikipedia, 2006). I shall denote this first type of psychometrics by *ψ-psychometrics*. The primary method of ψ-psychometrics is placing objects to one's forehead, with the eyes closed. This method provides vague statements concerning the history of an object or of its (former) owner. This information is often, albeit not always successfully, used to find missing persons or to solve crimes (Roll, 2003). An American physiologist by the name of J.R. Buchanan has studied and taught ψ-psychometrics for a while at one American university (Buchanan, 1889), but academic interest in the topic has mostly waned. Nonetheless, the merits and nature of ψ-psychometrics remain to be heatedly debated within and beyond the field of ψ-psychometrics (e.g., Randi, 1982; Roll, 2003).

The second type of psychometrics, which I shall denote it by *a-psychometrics*, is the most well-known of the three types of psychometrics. It is taught at most universities and is the topic of hundreds of books (e.g., Jensen, 1998). The primary methods of α-psychometrics are the administration of IQ tests, the computation of correlation coefficients, and the occasional use of Principal Components Analysis. The goal of α-psychometrics is to gain understanding in the hypothetical construct of general intelligence or *g*,[1] and to study the degree to which *g* can explain a host of psychological and societal phenomena. α-Psychometrics produces global verbal assessments on the nature and potency of *g*, which are often published in a journal called *Intelligence*. α-Psychometrics is heatedly debated within the field of α-psychometrics and beyond. In fact, virtually anyone has some opinion on α-psychometrics, particularly when group differences are involved.

The third type of psychometrics is only taught at good universities (one of which happens to be the University of Amsterdam). It is perhaps the least known of the three types of psychometrics, although it has its own journal called *Psychometrika*, and is well-organized in the Psychometric Society. I shall denote this third type of psychometrics by *β-*

[1] Note that *g* does not stand for "god", but for the adjective "general". Nonetheless, in some circles of α-psychometrics, *g* appears to have an almost religious status. For instance, the concept of *g* is the driving force behind all sorts of phenomena all over the world (e.g., Lynn & Vanhanen, 2002), but the nature of *g* itself need not be explained further.

psychometrics. This type of psychometrics is concerned with understanding the relation between test or item scores and the latent variable(s) supposed to underlie those scores. The field of β-psychometrics employs statistical models to understand more fully this complex relation. β-Psychometrics is mainly debated *within* the field of β-psychometrics (cf. Borsboom, 2006a). Outside this field, however, β-psychometrics is often considered too difficult and is mostly ignored.

The gap between ψ-psychometrics and the two other types of psychometrics is irreconcilably large, and we do not consider ψ-psychometrics further. However, β-psychometrics and α-psychometrics are more strongly related, for the simple reason that they have a common ancestry (e.g., Spearman, 1904). The aim of the current thesis is to bridge the gap between β-psychometrics and α-psychometrics. In the studies of this thesis, tools developed in β-psychometrics are applied to address problems in the field of α-psychometrics. In other words, this thesis is aimed to further our understanding of the relation between intelligence test scores and the underlying dimensions of cognitive ability, in order to gain insight in several phenomena in intelligence testing. Specifically, the studies in this thesis are concerned with group differences in intelligence test scores that have made α-psychometrics both famous and controversial. Next, I will shortly discuss the gap that has emerged between β-psychometrics and α-psychometrics. After that, I will provide an overview of this thesis.

1.2 The Gap Between α-Psychometrics and β-Psychometrics

There was a time when β-psychometrics and α-psychometrics were one and the same. Psychologists like Thorndike, Thurstone, and Spearman were all founding members of the Psychometric Society, and lay the foundations for both α- and β-psychometrics. They were interested in the substantive aspects of intelligence, as well as in the statistical characteristics of intelligence test scores. As these psychometricians were succeeded by later generations of researchers, and as the field of intelligence research expanded considerably, the field slowly evolved into α- and β-psychometrics. This development is indicated by changes in the editorial board of the journal Intelligence. In the early 1980s, three past- or later presidents of the Psychometric Society were members of the 18-headed editorial board. Anno 2006, of the 24 members of the editorial board of Intelligence, only one (4%) has once made an appearance in Psychometrika. Likewise, of the 26 current members of the editorial board of Psychometrika, only one (4%) has published in Intelligence. This is a striking development given the interrelated history and the large overlap between those two fields. After all, both these types of psychometrics are concerned with understanding latent traits by measuring them. β-Psychometrics and α-psychometrics appear to be two old friends who somehow have lost contact over the years.

The ensuing gap between α- and β-psychometrics was already evident in Jensen's (1980) impressive book (i.e., 799 pages) *Bias in Mental Testing*. This book is an α-psychometric work with many β-psychometric components. Nonetheless, Jensen chose not to focus on modern test theory, but on classical test theory instead (cf. Lord & Novick, 1968). For instance, Jensen's conclusion that measurement bias at the item level was not present in the comparison of cognitive ability test scores of Black and White Americans

was primarily based on classical test theory methods that were criticized in β-psychometrics as early as the 1970s (Ironson & Subkoviak, 1979; Lord, 1977). Since then, the field of β-psychometrics has developed more advanced tools to detect item bias or Differential Item Functioning (Holland & Wainer, 1993; Millsap & Everson, 1993). However, these contemporary methods are applied rarely in the field of α-psychometrics. Jensen's overview of bias research also drew heavily on the comparison of predictive regression lines across groups as a method to detect measurement bias of tests (Cleary, 1968). In the field of β-psychometrics, it is well established that differential prediction is not informative for the issue of measurement bias (Millsap, 1995, 1997a, 1998; cf. Reilly, 1973). However, as of 2005, the field of α-psychometrics still uses this method to claim that measurement bias with respect to ethnic groups does not exist (e.g., Rushton & Jensen, 2005a).

In his 1980 book Jensen devoted *one* sentence to factorial invariance within the common factor model, despite several studies in β-psychometrics that showed the suitability of this approach in studying group differences in factorial structure (Jöreskog, 1971; Meredith, 1964; Sörbom, 1974). Jensen devised his own method to study the nature of group differences in multivariate test scores, viz. the method of correlated vectors (Jensen, 1998). This method remains to be used in the field of α-psychometrics (e.g., Hartmann, Kruuse, & Nyborg, 2007; Te Nijenhuis, Tolboom, Resing, & Bleichrodt, 2004), despite extensive work by β-psychometricians that has shown that this method is all but flawless (Dolan, 2000; Dolan & Lubke, 2001; Lubke, Dolan, & Kelderman, 2001; Lubke, Dolan, Kelderman, & Mellenbergh, 2003a).

Old friends have a lot in common and generally enjoy being reunited, although reunions may be a bit awkward in the beginning.[2] The best way to reunite α- and β-psychometrics is to focus on the strengths of both approaches. This thesis shows that the application of methods from β-psychometrics can contribute to understanding several phenomena in α-psychometrics. It also illustrates that the use of β-psychometric models can be greatly improved when substantive theories are translated to measurement models.

1.3 Overview of This Thesis

Group differences in intelligence test scores are among the most controversial topics of psychology. Essential to the understanding of the nature of these group differences is whether or not groups can be reasonably compared in terms of the latent traits that the tests at hand are supposed to measure. Such a comparison of latent traits requires that the relation between test scores and latent cognitive variables is identical across groups. Whenever groups differ in this relation, we speak of measurement bias. Clearly, measurement bias complicates the comparison across groups of test scores. On the other hand, when test scores are characterized by the same measurement properties over

[2] To illustrate this, consider the following incident that took place at an α-psychometrics conference in December 2004. The presentation of a study in which rigorous β-psychometric methods showed that *g* was not the source of sex differences in intelligence test scores (Dolan et al., 2006) was reacted upon by Richard Lynn, an α-psychometrician and a strong proponent of the view that there are sex differences in *g*. Lynn asserted that the use of advanced psychometrics may lead to confusingly inconsistent results, and that we should just focus on IQ scores to study sex differences in intelligence. Unfortunately, reunions cannot be successful when attendees decide to ignore each other (see, e.g., Hartmann et al., 2007).

groups, we speak of measurement invariance. Measurement invariance is an interesting ideal, but it does not always arise in real data. Measurement invariance is not only interesting from a β-psychometrics perspective, but also highly relevant for many substantive issues in the field of α-psychometrics.

This thesis mainly draws on the work of Mellenbergh (1989) on measurement invariance and the work of Meredith (1993) on how to study measurement invariance by means of Multi-Group Confirmatory Factor Analysis or MGCFA (Lubke et al., 2003a; Lubke, Dolan, Kelderman, & Mellenbergh, 2003b). MGCFA is a model-based approach with which group differences in multivariate test scores can be studied (Dolan, 2000). As such, this approach is well-suited to study group differences in intelligence test scores.

The focus of Chapter 2 (Wicherts, Dolan, & Hessen, submitted) is on so-called intercept differences. This chapter provides an introduction of measurement invariance as defined within the common or confirmatory factor model. It shows that the suboptimal use of MGCFA is very common. Moreover, the study in this chapter illustrates how the use of suboptimal methods to study group differences in multivariate test scores can result in incorrect assessment of the appropriateness of tests for particular groups. The application of measurement invariance testing in this chapter is of the traditional type, viz. a comparison test scores of different ethnic groups in order to study the "fairness" of tests for ethnic minorities. The results of the re-analysis in Chapter 2 show that a commonly used Dutch IQ test underestimates IQ of ethnic minority children by about 7 IQ points. Such results signal a strong need for more research on measurement bias in the common factor model, particularly for tests that are used in applied settings.

A little known frustration of β-psychometricians is that they often encounter measurement bias, but are not able to understand the reasons for measurement bias. That is, they do not know "the biasing variables" (Mellenbergh & Kok, 1991). Chapter 3 (published as Wicherts, 2005b; Wicherts, Dolan, & Hessen, 2005) is focused on one of these biasing variables, namely the effects of stereotype threat on test performance. Stereotype threat (e.g., Steele & Aronson, 1995) is the pressure on a test taker arising from stereotypes related to the academic proficiency of one's social group. Numerous studies in experimental social psychology have shown that this effect may lower test performance of members of stigmatized groups (Steele, Spencer, & Aronson, 2002). With the notable exception of Jensen (1998), few α-psychometricians have discussed the relevance of stereotype threat to the issue of group differences in intelligence test scores (cf. Stricker & Bejar, 2004). An interesting aspect of Chapter 3 is that it combines the individual differences approach with the experimental approach (Cronbach, 1957). Like most experimental psychologists, social psychologists who studied stereotype threat were mainly interested in mean differences between groups, and employed Analyses of Variance (ANOVA) to analyze these. ANOVA has its drawbacks when used to study phenomena that are related to individual differences, but the use of MGCFA circumvents such problems. The results of the studies in Chapter 3 show that stereotype threat indeed results in measurement bias. This suggests that the use of MGCFA or other bias detection methods can shed light on the generalizability of stereotype threat effects to real-life test settings.

Chapter 4 (published as Wicherts et al., 2004) focuses on the fascinating phenomenon of secular increases in average IQ test scores of populations over time. For instance, in The Netherlands a version of a well respected IQ test (i.e., Raven's Progressive Matrices; J. C. Raven, 1960) was administered to basically all male 18-year-old military draftees from 1952 to 1982. The 1982 cohort scored approximately 20 IQ points higher than the 1952 cohort (Flynn, 1987). Political philosopher James Flynn (1984; 1987; 1998c; 2006) established the gain in IQ test scores as a robust phenomenon all over the developed world, and the effect is now commonly known as the Flynn Effect.[3] The Flynn Effect baffled many in the field of α-psychometrics, particularly those who subscribed to the view that intelligence was strongly heritable. Moreover, the Flynn Effect led several authors to doubt the validity of IQ tests (Flynn, 1987). In Chapter 4, measurement invariance across cohorts is tested in order to better understand the nature of the Flynn Effect. The results show that the Flynn Effect is not accompanied by measurement invariance, which has important implications for our understanding of this effect. That is, these results imply that the gains in IQ test scores cannot be solely due to increases in latent cognitive ability.

Chapter 5 is the only empirical chapter without the results of factor analysis, although Principal Components Analysis is employed in this study. This chapter is concerned with the controversial topic of IQ in Africa. Richard Lynn (2006) maintained that average IQ in this part of the world lies below 70. Unlike others (e.g., Herrnstein & Murray, 1994), the author of this thesis was rather skeptical of this low estimate and set out to critically evaluate the research on which this claim of low average IQ was based. This resulted in a meta-analysis, the results of which indicate that Lynn's estimate of average IQ in Africa is too low.

In addition, the published studies of IQ in Africa illustrate how strongly β-psychometrics and α-psychometrics have lost contact over the years. A comparison of IQ test scores between western samples and African samples is probably the greatest challenge to the merits of an intelligence test. IQ scores in Africa have been claimed to be both valid (Rushton & Jensen, 2005a) and invalid (Greenfield, 1997; Nell, 2000). Such a dispute can be resolved by studying measurement invariance across cultural groups. As we will see in Chapter 5, rigorous β-psychometric techniques have rarely been applied to address the meaning of IQ test scores in Africa. The methods used by several α-psychometricians do not meet the standards of β-psychometrics. Therefore, it is entirely unclear what IQ test scores in Africa mean, and whether these can be compared to IQ scores in western samples in terms of differences in latent cognitive ability. Regardless of the unclear β-psychometric status of African IQ, the results of the meta-analysis do not sit well with theories that

[3] Some authors (e.g., Rushton, 1999; Te Nijenhuis, Voskuijl, & Schijve, 2001) suggested that the effect be renamed the Lynn-Flynn Effect, because Lynn (1982) also contributed to establishing the phenomenon. There are several reasons not to rename the effect as such. First, Flynn (1984; 1987) did far more than Lynn to put the effect on the map. Second, Ms. Lynn Flynn is a real estate agent from Truckee, California, who has no involvement whatsoever in IQ research. Third, if one were to name the effect after those who noticed it before Flynn did, Tuddenham (1948) and Cattell (1950) should also be honoured. However, then all articles concerning the secular increase should use the term Tuddenham-Cattell-Lynn-Flynn Effect, which would be a waste of precious journal space. Fourth, there is no need to add to the term Lynn's name, for the simple reason that his name is already included in the term "FLynn Effect".

assign a substantial role to genes in race differences in intelligence (e.g., Lynn, 2006; Rushton, 2000b).

In Chapter 6 it is argued that the use of β-psychometric modeling can contribute greatly to the understanding of cognitive abilities. In addition, this chapter discusses the results of the studies in Chapter 2-5, and concludes with the scientific cliché that more research is needed. This research should more fully integrate the merits of β-psychometrics and α-psychometrics, because these two old friends can contribute greatly to each others' work. Finally, Chapter 7 (published as Wicherts & Dolan, 2004) is an appendix concerned with the use of fit measures in applications of MGCFA with mean structure.

The author sincerely hopes that β-psychometricians, α-psychometricians, and others will read this thesis with much interest. The author doubts whether any ψ-psychometricians will actually read this thesis. But then again, they will probably already know its contents after holding the book shortly against their foreheads with their eyes shut.

2

Measurement invariance and group differences in intercepts in confirmatory factor analysis

Measurement invariance with respect to groups is an essential aspect of the fair use of scores of intelligence tests and other psychological measurements. In this chapter, it is shown why establishing measurement invariance with confirmatory factor analysis requires a statistical test of the equality over groups of measurement intercepts. Without this essential test, latent mean differences are ambiguous and measurement bias may be overlooked. The implications and meaning of group differences in measurement intercepts are discussed. A re-analysis of a study by J. Te Nijenhuis, E. Tolboom, W. Resing, and N. Bleichrodt (2004) illustrates that ignoring intercept differences may lead to the conclusion that bias of IQ tests with respect to minorities is small, while in reality bias is quite severe.

2.1 Introduction

The valid and fair use of psychological tests in clinical psychology, education, and other settings requires that tests measure what they are supposed to measure, and that test scores are not affected by irrelevant characteristics associated with membership of demographic groups (e.g., ethnicity, gender). In the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999, henceforth the Standards), test fairness is defined as a situation in which "examinees of equal standing with respect to the construct the test is intended to measure should on average earn the same test score, irrespective of group membership" (p.74). For instance, suppose members of an ethnic minority group underperform on an IQ test, because of their unfamiliarity with certain words in the instruction texts. If, as a consequence, this test underestimates IQ of a group by, say, one third of a standard deviation (i.e., 5 IQ points), this test would generally be considered unsuitable for use in high-stakes decisions in education. Moreover, individual test scores based on such a test should be interpreted very cautiously, if at all. Fortunately, various statistical methods have been developed that can be used to detect measurement bias at both the scale and the item level (e.g., Millsap & Everson, 1993; Raju, Laffitte, & Byrne, 2002). However, the suboptimal or incomplete use of these methods may still result in the conclusion that measurement bias is absent when in fact measurement bias is present. The aim of this chapter is to show that establishing measurement invariance (i.e., *un*biasedness) by means of multi-group confirmatory factor analysis (MGCFA) requires a model that incorporates between-group mean structure (Meredith, 1993). Despite the ubiquitous use of the MGCFA framework in testing measurement invariance or equivalence across groups

of psychological tests in many settings, the mean structure is often not modeled statistically (Vandenberg & Lance, 2000). Our aim is to show that ignoring (or not testing for) between-group differences in measurement parameters related to the mean structure (i.e., measurement intercepts) may lead to incorrect conclusions regarding the appropriateness of tests for certain groups. The reason for this is that a group difference in a measurement intercept is indicative of a trait-irrelevant depression (or elevation) of test scores within a particular group. Such an effect on test scores violates measurement invariance. Therefore, the failure to identify group differences in measurement intercepts may have serious social and individual consequences, particularly in test settings in which test scores are used for psychological assessment and/or selection purposes.

Although the development of tests for measurement invariance was motivated by the ideal of fairness in intelligence and achievement testing for various demographic groups, tests of measurement invariance are applied widely in studies of comparability of many kinds of psychological measurements in areas such as education, cross-cultural psychology, applied psychology, intelligence research, and clinical psychology. Throughout this chapter, we use latent cognitive ability as an example. However, our argument applies to any kind of latent variable (e.g., depression, mood, personality, etc.). Moreover, we use terms like bias and fairness, whereas in many applications in which intercept differences may play a role (e.g., cross-cultural research), one would normally not denote these differences as unfair because fairness is simply not an issue. The technical term bias refers to any group difference on test scores, which cannot be accounted for by group differences on the construct that the test purports to measure. These additional group differences often show up as group differences in measurement intercepts. As we aim to show, these intercept differences may provide valuable information on the causes of group differences in test scores in many kinds of group comparisons.

In what follows, we first show that it is quite common that groups are compared on multivariate test scores without a rigorous modeling of the mean structure. After that, we provide an explicit definition of measurement invariance, which underlies the definition of fairness cited above, and explain how group differences in measurement intercepts violate measurement invariance under this definition. Next, we discuss conceptually how intercept differences may be detected by means of MGCFA, and how such differences may arise. Finally, we illustrate the importance of studying intercept differences in a re-analysis of data from a study into the appropriateness of an intelligence test for ethnic minority children in the Netherlands.

2.2 Disregard of Intercept Differences

Various tutorials have been written on how to investigate measurement invariance (or equivalence) using confirmatory factor analysis (Little, 1997; Lubke et al., 2003a; Ployhart & Oswald, 2004; Widaman & Reise, 1997). Most tutorials, although not all (e.g., Van de Vijver & Leung, 1997) stress the importance of modeling the mean structure when assessing measurement invariance across groups. However, in their exhaustive overview of the literature on empirical tests for measurement invariance with MGCFA from 1981 up to 1999, Vandenberg and Lance (2000) found that only in a small proportion (i.e., 12%) of

measurement invariance studies, intercept differences were actually studied. To gain insight in the current practice in the study of measurement invariance within MGCFA, we tried to locate all measurement invariance studies in psychology and related fields published in 2005.[4] We thus obtained a total of 110 studies in which MGCFA was employed to study group differences. However, only in 27 of these studies (24.5%) intercept differences across groups were tested. In a total of 69 studies it was concluded (in the abstract) that measurement invariance across groups was established. However, of these studies, only in 25% (i.e., 16 studies) intercept differences could be ruled out as a potential source of measurement bias. In the remaining 75% of these MGCFA studies, measurement invariance was claimed without a test of intercept differences.

Unfortunately, the investigation of mean differences in MGCFA is not part of many structural equation modeling courses (Stapleton & Leite, 2005). This may have contributed to the fact that the literature contains many examples of studies purporting to show that mean test score differences between groups are attributable to mean differences on latent factors, without the essential test of equality of measurement intercepts (e.g., Crockett, Randall, Shen, Russell, & Driscoll, 2005; Liu, Borg, & Spector, 2004; Te Nijenhuis, Tolboom et al., 2004). Moreover, it appears to be a commonly held view that the equality of factor loadings is sufficient for establishing measurement invariance (see, e.g., Coatsworth et al., 2005; de Frias & Dixon, 2005; Du & Tang, 2005; Ghorpade, Hattrup, & Lackritz, 1999; Woehr, Sheehan, & Bennett, 2005; Yao & Wu, 2005). For instance, de Frias and Dixon (2005) recently studied measurement invariance of the Memory Compensation Questionnaire (MCQ) across gender and age groups. Based on their finding that factor loadings were invariant across these groups, they claimed to have established measurement invariance, which according to them "provides assurance that the observation of group differences [...] is attributable to the process of memory compensation" (p.175). Although the equality over groups of the factor loading estimates is a necessary condition for measurement invariance, it is *insufficient* for attributing test score differences over groups to latent differences in constructs. Ipso facto, equality of factor loadings over groups does not allow the conclusion that a test is free of bias. For mean comparisons across groups to be valid, and for a test to be fair towards members of particular groups, group differences in (factor loadings and) intercepts need to be studied first.[5]

---

[4] To this end, we used the following search strings in PsychInfo: "invariance", "equivalence", "invariant", "equivalent and factor", "multiple and factor", "multi group and factor", "multi sample and factor", "factor analysis and differences", "factor analysis and comparison", "simultaneous and factor analysis", "MACS", and "mean and covariance structure". In addition, in Web of Science, we searched for all published papers referring to several seminal papers on measurement invariance. We restricted our interest to studies in which invariance was tested across existing groups (e.g., ethnic groups, gender). An overview of all studies is available upon request from the first author.

[5]Note that whenever the mean structure is modeled, the interpretation of latent (factor) mean differences across groups also requires a test of the equality of intercepts. Nonetheless, we came across several papers that included a latent mean comparison across groups without providing the results of the statistical test that the intercepts are indeed group invariant (Chirkov, Ryan, & Willness, 2005; Corwyn & Bradley, 2005; Hagger, Chatzisarantis, Barkoukis, Wang, & Baranowski, 2005; McInerney, Dowson, & Yeung, 2005). However, in the absence of these test results, it remains uncertain whether (or to what extent) the observed group differences are actually due to mean differences at the latent level.

Dutch Minority IQ Test Performance

Many studies of measurement invariance have direct practical consequences, especially when the tests involved are used for standardized assessment (with norms for the general population). In the Netherlands, several studies of invariance have been concerned with the suitability of Dutch intelligence tests for ethnic minorities who on average score lower than Dutch majorities (e.g., Helms Lorenz, Van de Vijver, & Poortinga, 2003; Te Nijenhuis, Evers, & Mur, 2000; Te Nijenhuis, Tolboom et al., 2004; Te Nijenhuis & van der Flier, 1997). These minority groups are mostly composed of first- or second-generation immigrants who are not necessarily as proficient in Dutch as native speakers (a situation comparable to that of many recent immigrants to the US). This may have a negative effect on their scores on cognitive ability tests. Unfortunately, however, in none of these invariance studies intercept differences have been tested statistically (but see Dolan, Roorda, & Wicherts, 2004). Nonetheless, conclusions are drawn concerning the appropriateness of tests for Dutch minority groups. For instance, Te Nijenhuis and colleagues (2004) studied measurement bias on a Dutch intelligence test (i.e., RAKIT) with respect to several groups of minority children. The results of their analyses, which ignored intercept differences, suggested "only little bias" (Te Nijenhuis, Tolboom et al., 2004, p. 24) with respect to minorities. However, our re-analysis by means of MGCFA with mean structure shows that, due to rather strong intercept differences, the underestimation of intelligence in a group of ethnic minorities amounts to at least 7 IQ points. This implies that the RAKIT should be used with caution in the assessment of intelligence in minority children in the Netherlands.

Intercept differences across groups are highly important to the issue of measurement invariance. Besides, such differences are rather common. Based on our review of MGCFA studies published in 2005, in two-thirds of the studies (18 of 27) that did model the mean structure, some intercept differences were detected. Nonetheless, our review of MGCFA studies indicates that these differences are often overlooked or simply ignored. This may be due to the fact that the importance of intercept differences is not fully appreciated. Moreover, some authors have expressed the need for more discussion on the meaning and nature of group differences in intercepts (Ployhart & Oswald, 2004; Raju et al., 2002; Vandenberg & Lance, 2000). Therefore, our aim is to elucidate why the equality of measurement intercepts over groups is important in understanding between-group differences, and in establishing that a certain test is fair or free from measurement bias for members of particular groups. To this end, we first discuss the definition of measurement invariance.

2.3 Measurement Invariance

The idea behind measurement invariance or unbiasedness is quite simple and intuitive. An important requirement of measurement invariance is that the expected (manifest) test scores of a person who has a certain level of latent ability (or abilities), are *independent* of group membership (e.g., Drasgow & Kanfer, 1985). Suppose, for instance, that a male and a female are equally proficient in mathematics. A systematic difference in their observed scores on a mathematics test would suggest the test is biased with respect to

gender. This is because measurement invariance (i.e., unbiasedness) requires that the expected test score given a certain latent ability, should not be influenced by, or depend on characteristics, other than the latent ability. To formalize this, let Y denote the manifest test scores, and let η denote a given fixed level on the latent trait that underlies the scores on Y. The expected test scores (denoted E(Y)), should depend on latent ability, but not on gender. So, when measurement invariance holds, and we condition on the level of the latent trait score, the expected scores should be equal for males and females:

$$\mathrm{E}(Y \mid \eta, \textit{male}) = \mathrm{E}(Y \mid \eta, \textit{female}) = \mathrm{E}(Y \mid \eta). \quad (1)$$

Note that this does not imply that females and males do not differ with respect to latent ability. Equation 1 concerns the conditional expectation given a fixed level of η and gender, it does not say anything about the conditional expectation given gender (i.e., E(Y | *male*) does not necessarily equal E(Y | *female*)).

This requirement of measurement invariance can be expressed more generally if we denote group membership by a grouping variable, which gives rise to group membership (e.g., gender, ethnicity, cultural group). Let v denote this grouping variable. Measurement invariance with respect to v requires that (Mellenbergh, 1989):

$$\mathrm{E}(Y \mid \eta, v) = \mathrm{E}(Y \mid \eta) \quad (2)$$

Equation 2 states that the expected values of Y given η and v should be equal to the expected values of Y given only η. Measurement invariance can be investigated empirically by formulating a measurement model, that relates the observed scores Y to the latent score(s) η (Millsap & Everson, 1993). As we demonstrate below, measurement invariance requires that the relationship between the test score(s) (i.e., measurement of ability) and the latent trait(s) (i.e., latent ability) of a person should not depend on group membership (Mellenbergh, 1989; Millsap & Everson, 1993).

In the case of a dichotomous (e.g., right/wrong) item measuring one latent trait (e.g., mathematical ability), the definition of invariance in Equation 2 requires that the probability of answering that item correctly (i.e., the expected value) given a particular latent trait score is identical for members of different groups. Within (parametric) item response theory, an item is considered to be unbiased if the parameter that links this probability to the latent trait is invariant over groups. For instance, the difficulty parameter of an item in a one-parameter logistic model should be identical across groups (e.g., Holland & Wainer, 1993). This aspect of item fairness is well known, it is explicitly mentioned in the Standards (i.e., Standard 7.3), and most studies of test fairness or test equivalence nowadays involve a test of Differential Item Functioning (DIF). However, measurement invariance also applies to the level of subtests in, for example, an intelligence test battery. That is, in most uses of such multivariate tests, the measurement aim exceeds the specific abilities tapped by particular subtests. Instead, the aim is to measure the ability that is common to several subtests. For instance, in general intelligence batteries such as the Wechsler scales (i.e., WAIS-III or WISC-IV; Wechsler, 1997, 2004), the measurement aim is either to get an indication of general intelligence, and/or of one of the four index scores (e.g., Verbal Comprehension, Perceptual Organization). Moreover, norm tables are usually not related to specific subtest scores, but to these broad factors. With such a measurement aim, measurement invariance requires that the expected subtest score conditional on latent

ability (e.g., Verbal Comprehension) be identical across groups. In that case, the intercept of the subtest needs to be group invariant, as we explain more fully below.

It is not generally recognized that a subtest from an intelligence test battery may display measurement bias. The Standards do not refer to this possibility, although they do stress the importance of studying group differences both in the internal structure of test responses (i.e., Standard 7.1), and in the effects of construct-irrelevant variance (i.e., Standard 7.2). These standards refer to the covariance structure, which is also an essential aspect of measurement invariance. For instance, the (error) variance around the expected test scores represents variance unaccounted for by the target trait(s). This variance may be due to some additional construct-irrelevant variable. Moreover, a test would normally be regarded unfair if its measurement precision in one group is considerably lower than the measurement precision in another group. Therefore, it is important to also consider group differences in the covariance structure.

In fact, the general definition of measurement invariance provided by Mellenbergh (1989) also relates to the covariance structure, because it is expressed in terms of the complete (conditional) distribution of Y, denoted by $f(Y \mid .)$. This definition states that measurement invariance with respect to v holds, if:

$$f(Y \mid \eta, v) = f(Y \mid \eta), \qquad \text{(for all } Y, \eta, v\text{)}. \tag{3}$$

Note that this definition does not depend on the exact nature of the distribution (i.e., continuous, discrete). If manifest data are (approximately) multivariate normally distributed, Equation 3 requires that, conditional on the latent trait scores, the expected values (i.e., Equation 2), the covariances between test scores (i.e., internal structure; cf. Standard 7.1), and the amount of variance unrelated to the latent trait(s) (cf. Standard 7.2) are equal across groups.[6] By adopting the linear confirmatory factor model as a measurement model (Mellenbergh, 1994), all these requirements of measurement invariance can be tested readily.

2.4 Multi Group Confirmatory Factor Analysis (MGCFA)

In this section we show how measurement invariance of continuously distributed test scores can be tested using MGCFA. Moreover, we show that group differences in measurement intercepts constitute a direct violation of the requirement in Equation 2. To ease presentation, we focus on the single common factor model in two samples. The elaboration to multiple-factor analysis in more than two samples is straightforward (cf. Bollen, 1989; Dolan, 2000; Lubke et al., 2003a).

The confirmatory factor model may be viewed as a measurement model in which the observed test or indicator scores (e.g., subtest scores) are regressed upon the scores on the latent, unobserved, construct η (Mellenbergh, 1994). As in ordinary linear regression, the model includes the following measurement parameters for each indicator: a regression weight or factor loading λ, a residual term ε, and an intercept τ. The test score y_1 of person j

[6]Multivariate normal distributions are characterized only by expected values and covariances. Therefore, full measurement invariance under normality requires that Equation 2 holds and that the (conditional) covariance structure, denoted by $\Sigma(Y \mid .)$, should follow: $\Sigma(Y \mid \eta, v) = \Sigma(Y \mid \eta)$.

in group i is predicted by the score on the latent variable or factor η (e.g., intelligence):

$$y_{1ij} = \tau_{1i} + \lambda_{1i}\eta_{ij} + \varepsilon_{1ij} . \quad (4)$$

Note that the expected value of the residual ε is assumed to equal zero, and that the residual is assumed to be uncorrelated with the factor score (as well as with the residuals of other indicators). The residual term of an indicator contains both random measurement error and specific factors tapped by that particular indicator (i.e., all uncommon sources of variance; DeShon, 2004; Meredith & Horn, 2001). The intercept is the value of y corresponding with the point where $\eta = 0$. In many applications (e.g., single-group studies) the mean structure is not of interest. However, in establishing measurement invariance over groups, the mean structure has to be incorporated in the analyses (Meredith, 1993).

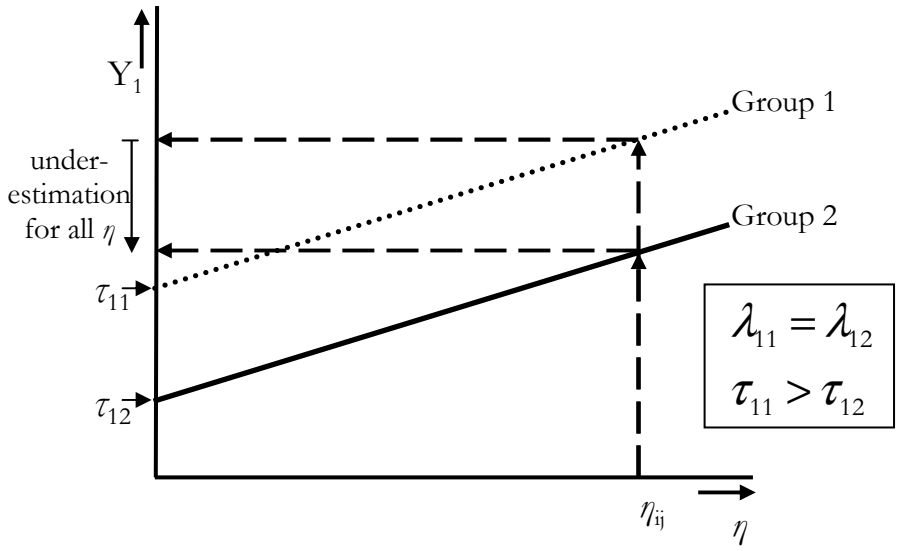

Figure 2.1 *Regression lines for the prediction of test scores Y_1 on a latent variable η in two groups when intercepts are unequal.*

With this measurement model in place, we can consider the implications of measurement invariance graphically. Figure 2.1 displays the regression lines relating the scores on a (sub)test to factor scores in two groups. In this figure factor loadings are identical in both groups, but intercepts are different. As can be seen, the intercept in Group 2 is lower than the intercept in Group 1. The consequences of this group difference in intercept are evident. Regardless of ability level, members of Group 2 with a certain ability, score lower than members of Group 1 with the same latent ability. Clearly, an intercept difference violates measurement invariance. Because the underestimation of ability in Group 2 is equal for all ability levels, this situation is denoted by *uniform bias* (Mellenbergh, 1982).

From the linear factor model of Equation 4, one can derive the expected test score given the factor score η_{ij} and group membership, as the sum of the intercept and the factor score weighed by the group-invariant factor loading (cf. Bollen, 1989). Suppose a person

from Group 1 and a person from Group 2 have the same latent ability, say: $\eta_{ij} = 1$. Then, the expected test score for the person from Group 1 and for the person from Group 2 equal:

$$\mathrm{E}\left(y_{1ij} \mid \eta_{ij} = 1, i = 1\right) = \tau_{11} + \lambda_1 \times 1, \tag{5a}$$

and

$$\mathrm{E}\left(y_{1ij} \mid \eta_{ij} = 1, i = 2\right) = \tau_{12} + \lambda_1 \times 1, \tag{5b}$$

respectively. In terms of Equation 2, it is clear that measurement invariance does not hold, because for *any* given value of η_{ij} , the expected test score for a person from Group 1 (i.e., Equation 5a) will be higher than the expected test score for a person from Group 2 (i.e., Equation 5b). The underestimation in Group 2 in this case is equal to the group difference in measurement intercepts (i.e., $\tau_{11} - \tau_{12}$). Depending on their direction, intercept differences may lead to an overestimation or an underestimation of group differences in latent ability.

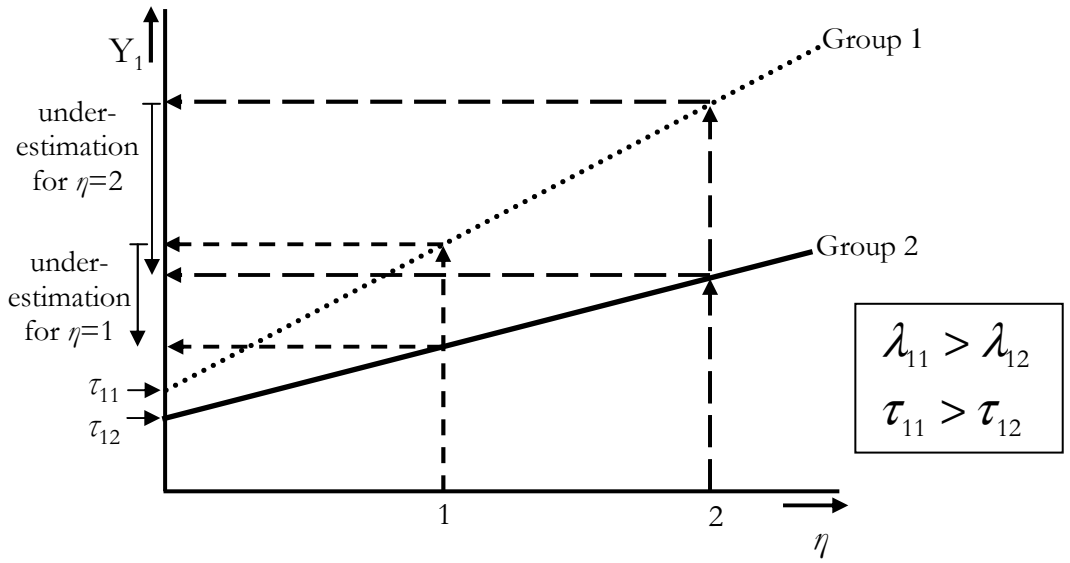

Figure 2.2 *Regression lines for the prediction of test scores Y_1 on a latent variable η in two groups when intercepts and factor loadings are unequal.*

Figure 2.2 displays a situation in which both the factor loading and the measurement intercept differ over groups. That is, for all members of Group 2, the ability is underestimated in *y* due to an intercept difference. Moreover, in this scenario the underestimation of ability depends on the particular ability level. Specifically, the underestimation of ability increases with increasing ability. In other words, besides a main effect for group, there appears to be an interaction effect such that higher ability levels suffer more from underestimation of ability. Note that this situation is denoted as *non-uniform bias* (Mellenbergh, 1982). The underestimation of latent ability in Group 2 now equals: $\left|\left(\tau_{11} - \tau_{12}\right) + \left(\lambda_{11} - \lambda_{12}\right) \times \eta_{ij}\right|$. Clearly, Equation 2 cannot hold in the presence of group differences in the factor loading λ and in the intercept τ. From Figures 2.1 and 2.2, it is apparent that both factor loadings and intercepts need to be invariant across groups for the fulfilment of Equation 2. Only when factor loadings and intercepts are group-invariant,

can we conclude that between-group mean differences on the indicators are a function of a latent group difference on the mean of the latent factor.

A further requirement of the general definition of measurement invariance (i.e., Equation 3) is that the variance around the expected values is group-invariant. Thus, the variance of *Y* conditional on the latent factor scores should be equal across groups:

$$\text{var}(y_{1ij} \mid \eta_{ij}, i = 1) = \text{var}(y_{1ij} \mid \eta_{ij}, i = 2) = \text{var}(y_{1ij} \mid \eta_{ij}). \quad (6)$$

Equation 6 implies that the variance of the residual term (i.e., residual variance) should also be equal across groups for measurement invariance to hold (DeShon, 2004; Lubke & Dolan, 2003; Meredith, 1993).

Detection of Intercept Differences

The detection of group differences in intercepts starts with the expansion of the model to several indicators (in fact, factor analysis is only feasible with several indicators of the common factor). Suppose we have four subtests measuring the same latent ability. Then, the linear models for each of the four subtests are equivalent to Equation 4. Although the intercepts and factor loadings may differ for each subtest, the latent ability score η_{ij} of person j is the same for all subtests, so we can conveniently arrange the four expressions of this factor model using vector notation:

$$\begin{bmatrix} y_{1ij} \\ y_{2ij} \\ y_{3ij} \\ y_{4ij} \end{bmatrix} = \begin{bmatrix} \tau_{1i} \\ \tau_{2i} \\ \tau_{3i} \\ \tau_{4i} \end{bmatrix} + \begin{bmatrix} \lambda_{1i} \\ \lambda_{2i} \\ \lambda_{3i} \\ \lambda_{4i} \end{bmatrix} \times [\eta_{ij}] + \begin{bmatrix} \varepsilon_{1ij} \\ \varepsilon_{2ij} \\ \varepsilon_{3ij} \\ \varepsilon_{4ij} \end{bmatrix}. \quad (7)$$

This, in turn, is more parsimoniously expressed by the following matrix notation:

$$y_{ij} = \tau_i + \Lambda_i \eta_{ij} + \varepsilon_{ij}. \quad (8)$$

Except for the difference in notation Equation 7 and Equation 8 are identical. For example, in Equation 8, Λ_i is a 4 x 1 matrix containing the factor loadings of group i. Equation 8 presents a model for the observations. To obtain estimates of the parameters in this model, we fit the observed covariance matrices and mean vectors to the covariance matrices and mean vectors that are implied by the model in Equation 8 (cf. Bollen, 1989). For instance, the covariance matrices that are observed within each group, can be used by a program such as LISREL (Jöreskog & Sörbom, 2003) or EQS (Bentler, 1995) to estimate model parameters and assess the fit of the model. The measurement parameters of interest are the factor loadings (Λ_i), the vector of intercepts (τ_i), and the variances of the residuals within each group, which are incorporated in a matrix denoted Θ_i. The distribution of factor scores (i.e., latent ability) within each group i is modeled by the factor means and factor variances, denoted by α_i and Ψ_i, respectively.[7]

Under measurement invariance, groups do not differ with respect to the relation between manifest test scores and the latent trait(s), and any group differences in manifest

[7] Given these assumptions, the observed variables are normally distributed $y_{ij} \sim N_p(\mu_i, \Sigma_i)$, where the implied mean vector equals $\mu_i = \tau_i + \Lambda_i \alpha_i$, and the implied covariance matrix equals $\Sigma_i = \Lambda_i \Psi_i \Lambda_i' + \Theta_i$ (superscript t denotes transpose). Note that one factor loading per factor is used for scaling purposes.

test scores are due to group differences at the latent level (i.e., a_i and Ψ_i). Therefore, under measurement invariance all measurement parameters should be invariant over groups (i.e., $\Lambda_i = \Lambda$, $\tau_i = \tau$, $\Theta_i = \Theta$),[8] which constitutes a situation denoted by *strict factorial invariance* (Meredith, 1993). The invariance of measurement parameters implies that the same constructs are being measured across groups. The tenability of measurement invariance can be studied by comparing the fit of models with and without the restriction that parameters are equal across groups. The preferred method is fitting a series of increasingly restrictive models, which are presented in Table 2.1 (cf. Lubke et al., 2003a; Vandenberg & Lance, 2000; Widaman & Reise, 1997). Because of the nesting of these increasingly restrictive models, equality over groups of each of the measurement parameters may be tested statistically by means of a likelihood ratio test or by using other indices of fit. The question arises how it is possible to disentangle group differences in measurement intercepts from group differences in latent ability.

Table 2.1
Equality constraints imposed across groups in steps towards strict factorial invariance

No.	Description	factor loadings	residual variances	intercepts	factor means
1	Configural invariance	Λ free	Θ free	τ free	α fixed at 0
2	Metric/weak invariance	Λ <u>invariant</u>	Θ free	τ free	α fixed at 0
3	Equal residual variances	Λ invariant	Θ <u>invariant</u>	τ free	α fixed at 0
4	Strict factorial invariance	Λ invariant	Θ invariant	τ <u>invariant</u>	α free[1]

Note: Each step is nested under the previous one; Underlined restrictions are tested in each step; free: freely estimated within each group; invariant: parameters estimated equally across groups; Factor (co)variances Ψ are freely estimated throughout. [1]Modeled as between-group differences in factor means by restricting factor means in one arbitrary group to equal zero.

One important aspect is that within confirmatory factor analysis with mean structure, mean structure and covariance structure are modeled simultaneously (Meredith, 1993). Factor loadings play an essential role in the connection between these two structures. The crux of the method to detect group differences in intercepts lies in the relation between factor loadings and between-group differences on the indicators. Namely, *if* between-group differences in the means of the indicators are due to between-group differences in the latent variable, one would expect that the relative size of between group differences on the indicators is collinear with the factor loadings. That is, the higher a subtest's factor loading, the better the scores on this subtest are predicted by the common factor, and the better this test is able to show (any) between-group difference at the latent level. Figure 2.3 displays the regression lines for two subtests loading on the same factor and the distribution of factor scores (η) in two groups. As can be seen, the two groups have a different mean on this factor (i.e., $a_2 > a_1$). In addition, the factor loading of subtest Y_1 (left-hand side) is smaller than the factor loading of subtest Y_2 (right-hand side). If factor loadings and intercepts are invariant over groups, the expected group difference is a

[8] That is, under measurement invariance the implied covariance equals $\Sigma_i = \Lambda\Psi_i\Lambda' + \Theta$, and the implied mean vector equals $\mu_i = \tau + \Lambda\alpha_i$. All group differences in Σ_i and μ_i are due to group differences in the covariances Ψ_i and means a_i of the factors.

function of the latent between-group mean difference (i.e., a_2 - a_1) weighed by the corresponding factor loading.[9] This means that on subtest Y_2, the expected mean group difference is larger than on subtest Y_1 due to the higher factor loading of the former subtest as opposed to the latter.

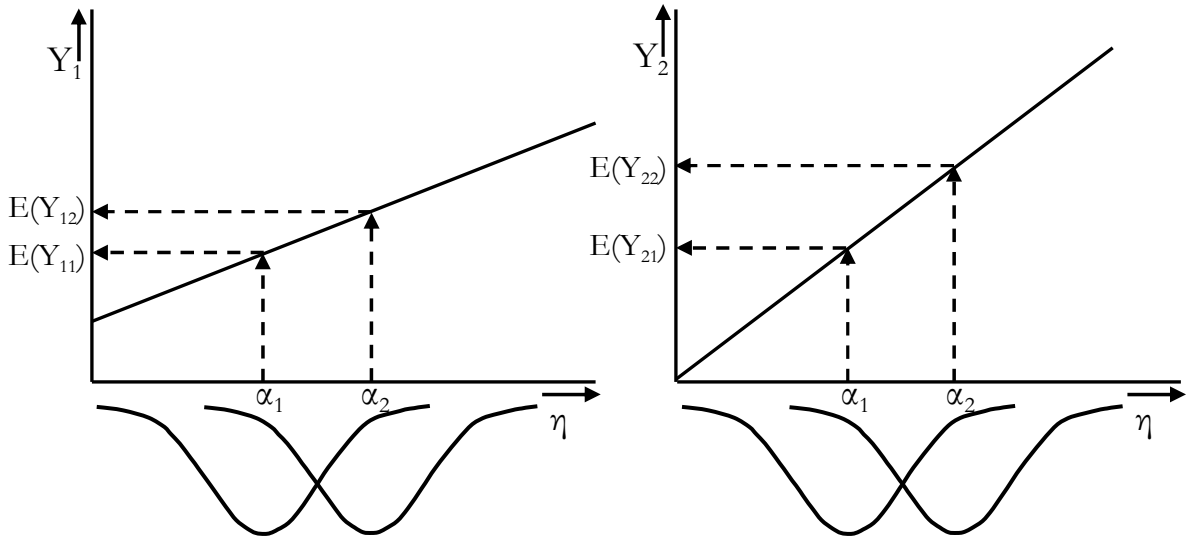

Figure 2.3 *Effect of larger (right) or smaller (left) factor loading on the expected between group difference on indicators.*

Thus, if mean group differences on the subtests are due to mean differences on the factor, this means that whatever mean score differences we might find, these should be expressed in a way that is compatible with the relative size of factor loadings. On the other hand, if the between-group difference on a subtest is not in line with the relative size of its factor loading, this implies that the between-group difference on this subtest could not be solely due to a between-group difference on the common factor. If such occurs, there is a group difference in the intercept on this subtest. Thus, different intercepts capture any between-group mean difference, which cannot be explained by between-group mean differences on the factor. If intercepts differ across groups, they should be estimated freely across groups. If there remain sufficient invariant indicators of a factor, this enables an unbiased estimation of factor mean difference, as well as an estimation of the degree of uniform bias on the biased indicator.

The statistical test of equality of intercepts is simply conducted by testing a model with group-invariant intercepts, while allowing for between-group differences in factor means (cf. Table 2.1, Step 4). It is crucial to assess the fit of equality of intercepts while allowing for differences in factor means (Meredith, 1993). The reason for this is simply that if there is *any* between-group difference in factor mean, and we would not allow for this

[9]Formally, the expected values in Groups 1 and 2 equal $\mathrm{E}(y_{11}) = \tau_1 + \lambda_1\alpha_1$ and $\mathrm{E}(y_{12}) = \tau_1 + \lambda_1\alpha_2$, respectively. If both λ_1 and τ_1 are group invariant, the expected mean group difference equals: $\mathrm{E}(y_{12} - y_{11}) = \lambda_1(\alpha_2 - \alpha_1)$.

possibility, this (latent) source of mean difference would be forced into differences in intercepts. This is equivalent to the requirement that a test of factor loadings must allow for between group differences in factor (co)variances (Meredith, 1993). Whether or not groups differ with respect to factor means (*α*) or factor (co)variances (Ψ), is not a matter of *measurement* invariance. Measurement invariance should be established before group differences at the latent level (e.g., 2nd order structure, latent means) are studied. In conclusion, between-group differences in intercepts are detectable with MGCFA because of the relation implicit in this model between covariance structure (within-group structure) and mean structure (between-group structure).

Meaning of Intercept Differences

According to the Standards "bias in tests [...] refers to construct-irrelevant components that result in systematically lower or higher scores for identifiable groups of examinees." (p.76). A difference in intercepts suggests that the mean difference between groups on that particular indicator cannot be accounted for by mean differences on the factor(s) that a test is supposed to measure. One may look at intercept differences as occurring because of a group-difference in the specific ability tapped by the corresponding indicator (Meredith & Horn, 2001). Another way to look at such a scenario is to imagine an additional factor "out there" that necessarily differs across groups, which results in a mean effect on that indicator (Lubke et al., 2003b).

An intercept difference may be due to bias in the traditional sense that certain words in the items of the corresponding subtest may be less familiar to members of one of the groups. If one expects such measurement artifacts, a further look at DIF may shed light on the source of the intercept difference. Mostly, however, one would not expect this to be the case because usually all indicators also tap specific abilities, which may simply differ over groups. For instance, in two studies of gender differences on the WAIS-III in Spain and in The Netherlands, it was found that the intercept of the Information subtest (which loads on Verbal Comprehension) was higher for males than for females (Dolan et al., 2006; Van der Sluis et al., 2006). This is in line with a reported gender difference in general knowledge (e.g., Lynn & Irwing, 2002), which suggests that males outperform females on Information. Such a subtest specific gender difference is not in line with the (non-significant) gender difference on the factor (i.e., Verbal Comprehension), resulting in an intercept difference. This effect results in an underestimation of female IQ as opposed to male IQ. On the Dutch WAIS-III this effect is small for Total IQ (about 1 point), but substantial for female Verbal Comprehension index scores (about 4 points or 0.25SD). In many applications of MGCFA there will be substantive reasons to expect intercept differences. For instance, older test takers may give slower responses than younger test takers on a timed test for abstract reasoning. This might be due lower processing speed in older test takers (e.g., Salthouse, 1996). In such a scenario, a measure of processing speed may be used to explain this intercept difference (Lubke et al., 2003a), thereby enabling a disentanglement of different aspects of aging on cognitive test performance.

It is important to stress that a common factor is defined within a particular factor model. It is quite possible that a subtest shows an intercept difference when it loads on one factor, but not when it loads on another factor in another model. Of course, the character

of a factor depends on its indicators. Reasons for intercept differences are not a characteristic of subtests per se, but of a characteristic of subtest *scores* as they relate to the common factor(s). Many test artifacts can give rise to intercept differences. For instance, the effects of stereotype threat on test performance may be seen as such an artifact. Stereotype threat (e.g., Steele & Aronson, 1995) is the pressure on a test taker arising from stereotypes related to the academic proficiency of one's social group. For example, it has been shown that (implicitly) reminding female test takers of the stereotype that women have lower math ability than men, may result in a lowering of female math performance, particularly when tests are difficult (O'Brien & Crandall, 2003; Spencer, Steele, & Quinn, 1999). Because such effects are often subtest specific, stereotype threat may also result in a lowering of measurement intercepts in stigmatized groups (Wicherts et al., 2005; Chapter 3). Therefore, a rigorous test of measurement invariance enables the detection of test artifacts that depress test scores of members of particular groups.

In conclusion, a between-group difference in intercept implies a uniform group-specific suppression (or elevation) of test scores, which may provide important information on the nature of group differences in test scores. We now turn to an illustration by means of a re-analysis of a study in which minority children and majority children are compared on intelligence test performance.

2.5 Illustration: IQ and Minority Children

Ignoring intercept differences between groups may have serious consequences, because such intercept differences may be indicative of an underestimation of ability in a particular group. We illustrate this by means of a re-analysis of a study by Te Nijenhuis and colleagues, who investigated whether a Dutch children's intelligence test (RAKIT) was suitable for children of immigrants from Turkey, Morocco, and the former Dutch colonies. In what appears to be a textbook example of a measurement invariance study, Te Nijenhuis et al. (2004) went to great length in studying invariance of the RAKIT across the different ethnic groups. They used DIF analyses, an analysis of differential prediction using school grades as a criterion, and MGCFA. Although they investigated the equality of factor loadings in the latter analyses by using a likelihood ratio test and a congruence measure, Te Nijenhuis et al. did not investigate whether measurement intercepts were equal across groups. Based on the findings of small DIF effects, only slight differential prediction, and group-invariant factor loadings, these authors concluded that the RAKIT "is highly, though not perfectly, valid for the assessment of immigrant children" (p.22). Our aim is to test for intercept differences in order to verify this claim of measurement invariance. Note that we restrict our attention to the test scores of a group of children of Moroccan and Turkish descent, aged 7, who were compared to a representative sample of Dutch majority children of the same age.

Method

Participants. The test scores of a representative sample of 196 majority children were used as comparison to the test scores of 131 children from Moroccan (N=60) and Turkish (N=71) descent. In view of power concerns we pooled these two minority groups for the

factor analyses (analyses per group gave similar results). Overall, the mean subtest performance did not differ significantly between the two immigrant groups: a MANOVA on the subtest scores resulted in a non-significant multivariate effect for group: F (12, 118) = 1.685, $p > 0.05$. In addition, a Box test showed that covariance matrices did not differ between Turkish and Moroccan children: F (78, 49603) = 1.08, $p > 0.05$. All minority children have followed education in Dutch. The minority sample is not explicitly sampled to be representative, but the children are from various schools in both rural and urban areas. The samples do not differ in age and in gender composition. Average age in both samples is 7 years and 8 months.

Intelligence Test. The RAKIT (Bleichrodt, Drenth, Zaal, & Resing, 1984) is an individually administered Dutch intelligence test for children (aged 4 to 11 years) composed of 12 subtests. RAKIT full scale IQ has been shown to correlate .86 with WISC-R full scale IQ (Bleichrodt et al., 1984). The subtests are Closure, Exclusion, Memory Span, Verbal Meaning, Mazes, Analogies, Quantity, Discs, Learning Names, Hidden Figures, Idea Production, and Storytelling. All instruction texts are in Dutch. Subtests with the largest language component are Verbal Meaning, Analogies, and Storytelling. Although subtest scores are standardized, and may be interpreted separately, the broad measurement aim of the RAKIT is to provide an indication general mental ability (i.e., IQ), and/or one of four factors, which are composed of the scores on 2 to 6 subtests.

Analyses. Based on Carroll's (1993) taxonomy, Te Nijenhuis et al. posited a factor structure with 4 factors: Hybrid (G_h), Visual (G_v), Memory (G_m), and Retrieval (G_r). This factor model is displayed in Figure 2.4.[10] Our focus is on the mean group differences on the subtest level, and we investigate whether these are attributable to group differences in the means of the four factors. As most of the RAKIT subtests have a rather strong language component, measurement bias with respect to minority children is a real possibility. In addition, item analyses by Te Nijenhuis et al. indicated that some subtests showed DIF. Despite this, Te Nijenhuis and colleagues concluded that only one of the subtests (i.e., Verbal Meaning) showed bias that was of any practical concern.

The tenability of strict factorial invariance with respect to groups is investigated by fitting a series of increasingly restrictive models, as presented in Table 2.1. In the first step, no between-group restrictions are imposed, although the configuration of factor loadings is invariant. The next steps involve restricting all factor loadings (Step 2) and all residual variances (Step 3) to be invariant over groups. In Step 4, the invariance of the mean structure is investigated by restricting the measurement intercepts to be equal across all groups. In the same step, factor mean differences with respect to an arbitrary baseline group are estimated.

[10]This factor model differs from the model which corresponds to the four factors in the manual (Bleichrodt et al., 1984). Using this alternative factor model to assess measurement invariance gave quite similar results.

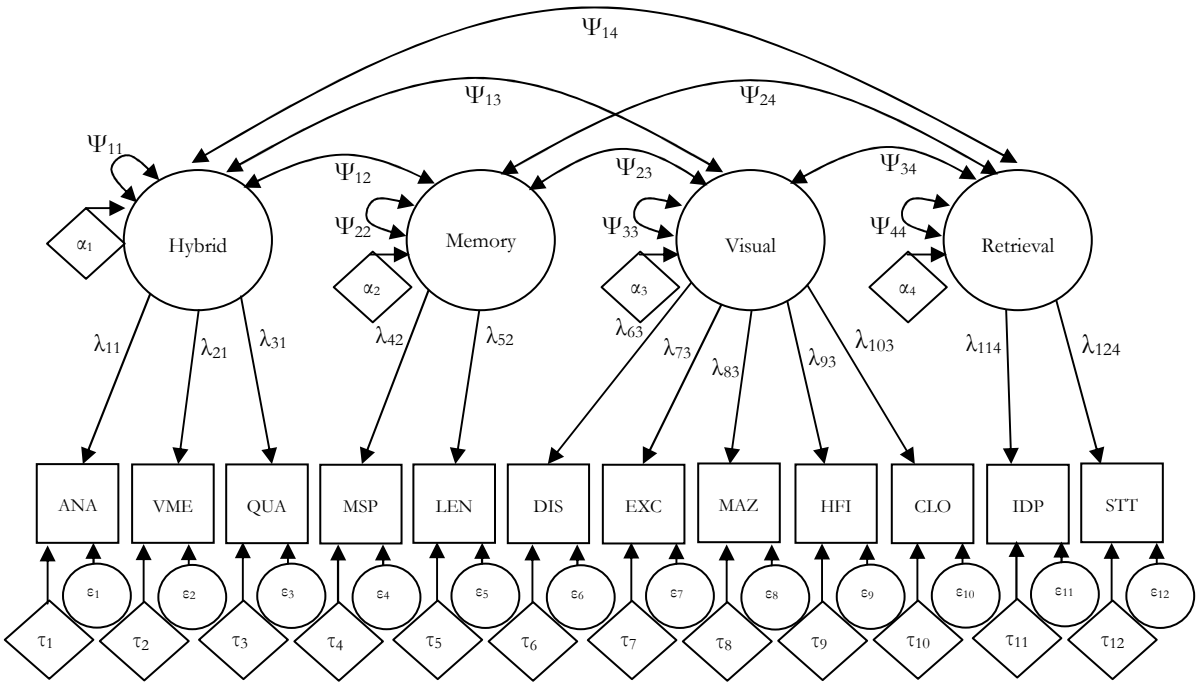

Figure 2.4 *Factor model for RAKIT subtests.*

The tenability of each restriction is judged by differences in fit between the restricted model and the less-restricted model. For instance, Step 2 vs. Step 1 involves the tenability of equality of factor loadings. As the successive models are nested (Bollen, 1989), a likelihood ratio test can be used to test each restriction. To assess model fit, and to assess the tenability of across-group restrictions on measurement parameters, we look at exact fit in terms of χ^2 and Degrees of Freedom (DF). We also consider the Comparative Fit Index[11] (CFI; Bentler, 1990) and Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1993). Based on their simulation study, Hu and Bentler (1999) suggested that CFI values above 0.95 and RMSEA values below 0.06 are indicative of good model fit. Besides these fit measures, we use the AIC for comparing the relative fit of models (cf. Wicherts & Dolan, 2004; Chapter 7). The AIC is a fit measure that takes into account the parsimony of models, with lower AIC values indicating better fit. In case a step is accompanied by a clear deterioration in model fit, the particular restriction is rejected. In such cases, modification indices can highlight the particular parameter(s) causing the misfit. A modification index (MI) is a measure of how much chi-square is expected to decrease if a constraint on a given parameter is relaxed, and the model is re-fitted (Jöreskog & Sörbom, 1993). A closer look at the magnitude of MIs of intercepts in Step 4 provides important information about intercept differences between groups. MI values larger than 3.84 indicate that model fit can be improved significantly ($p < .05$).

[11] Widaman and Thompson (2003) have argued that because of the nesting of models it is inappropriate to employ the standard null model within the MGCFA context with mean structure. Therefore, we use a model without any factor structure, in which intercepts and residual variances are restricted to be group invariant (i.e., model 0A in Widaman & Thompson, 2003) as the null model in computing the CFI values.

Results

The means and standard deviations of both groups of all subtests are displayed in Table 2.2. As can be seen, the mean differences between both groups are large. Figure 2.5 displays the effect sizes of the difference between the majority and minority group per subtest. For each subtest, Figure 2.5 also contains the estimates of factor loadings, as estimated in the majority group without across group-restrictions (i.e., Step 1). To ensure comparability of factor loadings and effect sizes, we restricted the highest factor loading for each factor to be identical to the effect size of the corresponding subtest.[12] This enables a comparison for each factor of the effect sizes per subtest in relation to the relative estimates of factor loadings. Recall that measurement invariance requires that mean group differences on the subtests should be collinear with the corresponding factor loading. That is, the higher a factor loading, the larger the mean difference should be. If effect sizes and factor loadings per factor are not collinear, this suggests intercept differences (a statistical test of which follows below).

Table 2.2

Means, standard deviations of RAKIT subtests for majority and minority group

Factor	Majority		Minority	
Subtest	M	SD	M	SD
Hybrid				
Analogies	15.03	4.94	10.66	4.73
Verbal Meaning	15.03	5.14	3.86	4.64
Quantity	15.21	5.10	9.51	5.53
Visual				
Discs	15.01	5.05	10.82	4.81
Exclusion	14.96	5.07	11.29	4.66
Mazes	15.02	5.03	11.60	4.88
Hidden Figures	14.94	4.93	10.95	4.81
Closure	14.85	5.06	10.37	5.89
Memory				
Memory Span	15.05	4.94	14.40	6.01
Learning Names	15.05	5.05	9.18	5.05
Retrieval				
Idea Production	15.06	5.18	11.05	5.42
Storytelling	14.99	5.05	10.19	5.22

[12] Usually, scaling of the common factor is achieved by restricting one factor loading per factor to equal 1. Because this value need not be necessarily 1, we used the effect size values here for illustrative purposes. Note that, because of this choice, the comparability of factor loading estimates across different factors is lost. Note also that the subtest scores reported are standardized norm scores. Hence, standard deviations are equal across subtests.

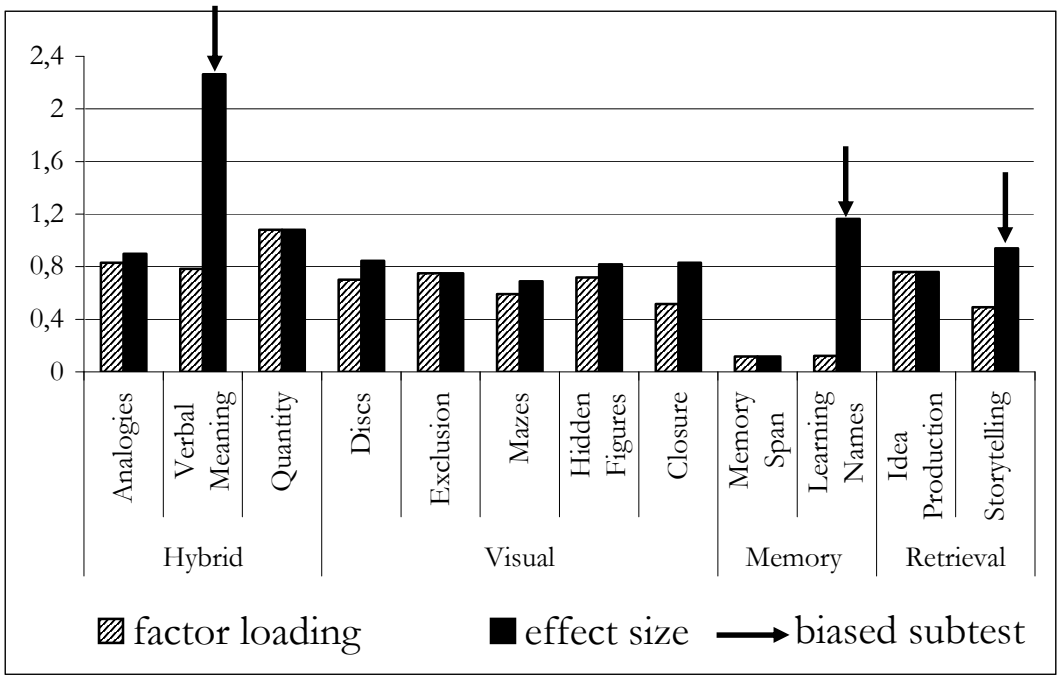

Figure 2.5 *Factor loadings and effect sizes per subtest.*

Consider the three subtests loading on the Hybrid factor. Of these subtests Quantity has the highest factor loading and Verbal Meaning the lowest, with the factor loading of Analogies assuming an intermediate value. As is strikingly apparent, the effect size of Verbal Meaning is far too large (d = 2.26) to be the result of a between group differences in the mean of the Hybrid factor. If this large mean difference were due to a latent mean difference, the (standardized) factor loading of the Verbal Meaning subtest would have been twice as large as the factor loading of the two other indicators of Hybrid ability. This is clearly not the case because we already know from the analysis of covariance structure in the majority group that this subtest has a factor loading smaller than the other two subtests. There may be several reasons for this result. It is conceivable, yet unlikely, that both the Analogies subtest and the Quantity subtest underestimate the ethnic difference on this factor. This would mean that both subtests are positively biased towards minority children. This explanation appears rather farfetched, because Verbal Meaning is a test measuring vocabulary knowledge and the minority group contains mainly non-native speakers of Dutch. Therefore, by inspecting the mean difference and the factor loadings, we would expect that the mean of the minority group on Verbal Meaning is too low. This suggests that the intercept of this subtest is considerably lower for minority children, and that this test is biased towards minorities.

Now consider very large difference between the effects sizes for the two indicators of the Memory factor, despite the fact that the factor loadings of both these subtests are very similar. Again, the subtest with the largest cultural component (Learning Names) shows the largest between-group difference. That is, the Learning Names subtest contains several Dutch names from various fairy tales, which may be unfamiliar to children from Moroccan and Turkish descent. The difference between the effect sizes of Learning Names

and Memory Span are so large, that is simply impossible for a single Memory factor to account for this effect. Finally, for the two indicators of the Retrieval factor, the subtest with the lowest factor loading is the Storytelling subtest. However, this subtest shows a larger between-group difference than the less-culturally loaded Idea Production subtest. Again we have to assume that the subtest with the smallest language or cultural component is biased in favor of minorities to circumvent the most obvious explanation, which is that the subtest with the largest between group difference is biased against minorities.

Thus, if we view the pattern of mean differences in light of the pattern of factor loadings based on the covariance structure in the majority group, we clearly see that these patterns are incompatible. This incompatibility is due to between-group differences in intercepts that are indicative of measurement bias. Of course, measurement invariance is investigated statistically by testing the fit of various models that differ with respect to between group constraints on factor loadings, residual variances *and* intercepts. The fit indices of the different models are reported in Table 2.3. First, we investigate the comparability of covariance structure (i.e., Steps 1-3). In Step 1, no between-group restrictions are imposed, although the configuration of factor loadings is equal across groups. As can be seen, the baseline model fits well in terms of RMSEA and CFI (cf. Hu & Bentler, 1999). In addition, the values of the Standardized Root Mean Square Residual (SRMR) indicate that the baseline model fits well in both the majority (0.054) and the minority group (0.062).

In Step 2, the factor loadings are restricted to be equal across both groups. As can be seen, this restriction is accompanied by a non-significant increase in chi-square. Moreover, all fit indices improve given this restriction. Therefore, factor loadings appear to be invariant across groups.

Table 2.3
Fit measures of steps towards strict factorial invariance

Step	Restrictions	DF	χ^2	ΔDF	$\Delta\chi^2$	p	RMSEA	CFI	AIC
1	-	96	152.52**				.059	.962	318
2	$\underline{\Lambda}$	104	157.13**	8	4.61	.798	.055	.964	306
3	$\Lambda, \underline{\Theta}$	116	179.96**	12	22.83*	.029	.059	.957	310
3a	$\Lambda, \Theta^{\underline{1}}$	115	170.93**	-1	7.03**	.008	.054	.962	299
4	$\Lambda, \Theta^1, \underline{\tau}$	123	240.80**	8	69.87**	.000	.077	.920	356
4a	$\Lambda, \Theta^1, \tau^{\underline{2}}$	122	196.91**	-1	43.89**	.000	.061	.949	311
4b	$\Lambda, \Theta^1, \tau^{\underline{2,3}}$	121	178.51**	-1	18.40**	.000	.053	.961	294
4c	$\Lambda, \Theta^1, \tau^{\underline{2,3,4}}$	120	174.37**	-1	4.14*	.042	.052	.963	292

Note: Underlined restrictions are tested by likelihood ratio test $\Delta\chi^2$. *p < 0.05; **p < 0.01; (-1): Parameter freely estimated; 1: Memory Span; 2: Verbal Meaning; 3: Learning Names; 4: Storytelling

In Step 3, the residual variances are restricted to be group-invariant. This step is accompanied by a slight deterioration in fit in terms of RMSEA, CFI, and AIC. In addition, the likelihood ratio test shows that this restriction appears untenable. A closer look at the modification indices shows that this misfit is mainly due to the residual variance of Memory Span (MI = 9). Indeed, freeing this parameter (Step 3a), leads to an improvement in model fit as can be seen by the significant decrease in chi-square, and improvements in RMSEA,

CFI, and AIC. In the majority group this residual variance is smaller (18.41, SE = 2.32) than in the minority group (31.48, SE = 4.19).

In Step 4, the intercepts are restricted to be equal across groups, while at the same time allowing a difference in the four factor means. As can be seen in Table 2.3, this restriction is accompanied by a clear drop in model fit. The increase in chi-square is highly significant, the RMSEA increases well above the cut-off for good fit, the CFI drops below 0.95, and the AIC is relatively large. As we already expected by visual inspection of Figure 2.5, mean subtest differences between the minority and majority group cannot be explained solely in terms of group differences in the means of the factors. Clearly, there are intercept differences between the groups.

A further look at the modification indices indicates that the intercepts of the following subtests differ across groups: Verbal Meaning (MI = 40), Learning Names (MI = 15), and Storytelling (MI = 5). Indeed, if we allow between group differences in these parameters, the model fit (in Models 4a through 4c) improves considerably. In all cases, the intercepts in the minority group are lower, indicating measurement bias with respect to this group.

One might ask whether these intercept differences are serious. Under the assumption that the remaining subtests are not biased, we can estimate the factor mean difference across groups. The multiplication of the factor loading with this factor mean difference provides the expected mean difference of the subtest (cf. Table 2.4) (see also Scholderer, Grunert, & Brunso, 2005). By comparing this expected mean to the mean difference actually obtained, we get the following underestimations per subtest: Verbal Meaning: 6.89, Learning Names: 5.12, and Storytelling: 1.79. For the total score, this means an underestimation of 13.8 points, which according to the transformation table in the test manual (Bleichrodt et al., 1984, p.128) represents an underestimation of the total IQ of 7 IQ points, or a little less than half a standard deviation.

Table 2.4
Estimation of bias due to intercept differences per subtest

Subtest	factor mean difference	factor loading	expected difference	actual mean difference	under-estimation
Expressions:	*A*	*B*	*C=A*B*	*D*	*=D - C*
Verbal Meaning	5.827	0.735	4.283	11.176	6.893
Learning Names	0.626	1.2061	0.755	5.870	5.115
Storytelling	4.003	0.754	3.018	4.804	1.786

Conclusion

By not testing for intercept differences, Te Nijenhuis and colleagues overlooked the fact that at least three of the twelve subtests in the RAKIT are biased for 7-year olds from Moroccan and Turkish descent.[13] These rather large intercept differences suggest that the RAKIT is not suitable for the assessment of minority 7-year-olds. A further analysis

[13] Te Nijenhuis et al. did notice the problems with Verbal Meaning, but missed the bias on Learning Names and Storytelling. The combined bias on the latter two subtests constitutes an underestimation of 3.5 IQ points or about 0.25 SD units.

(available upon request) of the data of Surinamese and Antillean children of the same age (at least 4.5 IQ points underestimation), and of children aged 5 and 9, gave similar results. Although the biasing effects for the other minority groups were less serious, the underestimation of ability was still large enough to render the RAKIT unsuitable for the use in these minorities. Even an underestimation of a few IQ points may have serious consequences. For instance, the Dutch Ministry of Education uses explicit cut-off IQ values (e.g., 70 or 85) for the selection of children for special education. An underestimation of the size we found for the children of Moroccan and Turkish may result in incorrect selection decisions. Although we would advice against use of the RAKIT in these immigrant groups for such purposes, a practical solution would be to discount the biased subtests. Alternatively, the intercept differences we found across groups may be used to correct upwards the subtests scores for these immigrant groups.

2.6 General Discussion

It is unfortunate that in many applications measurement invariance is assumed to hold without testing for the equality over groups of measurement intercepts. Our present aims were to show why a test of the equality of measurement intercepts across groups is essential for measurement invariance, what group differences in intercepts may mean, and how these differences can be detected. If the intercept of a particular subtest is different across groups, this implies that between-group differences on this subtest cannot be solely due to between-group differences in the construct(s) that the subtest is supposed to measure. In other words, an intercept difference indicates measurement bias in the sense there are one or more construct-irrelevant variables causing group differences in test scores. The importance of studying intercept differences was illustrated by a re-analysis of a study into the appropriateness of a Dutch intelligence test for minority children. The results indicated the presence of rather strong measurement bias, which was not fully appreciated in the original study, despite the fact that the analyses in that study appeared quite thorough.

It may be argued that the requirement of identical measurement intercepts over groups is too stringent, and will prove to be too restrictive in most data analyses. However, intercept differences do not render test scores completely incomparable across groups. Quite to the contrary, intercept differences may be taken into account, their size may be estimated (provided that there remain sufficient invariant indicators), and they may provide valuable information on the precise nature of between-group differences in test scores in many applications.

The seriousness of intercept differences depends on the measurement aim. If we allow for intercept differences, we also allow for group differences in the mean of the specific ability tapped by an indicator. Note that such an effect may or may not be due to DIF at the item level, which should be studied separately. A further issue refers to the size of intercept differences one is willing to accept (Borsboom, 2006b). Again, it depends on the use of the test. Fortunately, as we showed in our empirical example, the effect size of such bias is easily computed provided that the remaining indicators of a factor are unbiased. In our example, the effects of bias could be directly related to its effect on IQ

scores, which enabled the expression of bias in terms of IQ points. In most applications, effect size estimates can be readily computed and related to the effects on norm scores. Millsap and Kwok (2004) provide an alternative approach to the question of whether or not bias is acceptable in the context of correct or incorrect selection decisions.

Uniform Bias and Non-Uniform Bias

Most of the MGCFA studies we reviewed (98 out of 110), have involved a test of the invariance of factor loadings (cf. Vandenberg & Lance, 2000). Of course, this is an essential test for any meaningful between group comparison of test scores. However, in many invariance studies thoroughly developed and well-validated tests are compared across (demographic) groups. The question arises how likely it is that in such comparisons factor loadings would differ across groups. Specifically, for a biasing variable to have an effect on the covariance structure, that variable needs to vary across persons within a group. One may ask how likely it is for a biasing variable to have a significant amount of variance (in relation to the variance of the target construct) to have such an effect. If a biasing variable has an effect that is specific to one indicator, it is more likely that such an effect shows up in group differences in residual variances (DeShon, 2004; Lubke & Dolan, 2003), than that the biasing variable has an effect on the factor loading of the affected indicator. There are roughly three scenarios in which factor loadings may differ across groups, resulting in non-uniform bias: (1) If the biasing variable affects more than one indicator of a factor. (2) If the biasing variable covaries strongly with the latent variable. (3) If the biasing variable interacts with the latent variable, such that, for instance, with increasing ability levels the effect of the biasing variable on the indicator increases.

It depends on the test at hand and the groups under study, whether one would expect non-uniform bias. In about one-third of the 27 measurement invariance studies we reviewed, some group differences in factor loadings were found. On the other hand, in two-thirds of these studies researchers encountered intercept differences. Therefore, uniform bias (i.e., intercept differences) appears to occur more often than non-uniform bias. Furthermore, the effects of uniform bias are by their very nature (i.e., depression of scores for an entire group) more serious in settings where test fairness is a concern. Moreover, in many settings where test fairness is not an issue, group differences in intercepts may provide valuable information on the constructs tapped by (sub)tests and the nature of group differences.

Implications for Practice

Psychological tests of various kinds are used in countless applied settings. Many of these tests are either developed with a particular factor structure in mind (e.g., WISC-IV, WAIS-III), or are amendable to investigation by CFA.[14] There is general agreement that test scores should not be affected by irrelevant characteristics attached to the membership of demographic groups. We have argued that the requirement of fairness also relates to the subtest level, which implies that in multivariate tests (e.g., intelligence battery), the

[14] Note that measurement invariance can also be studied using exploratory factor analysis (Hessen, Dolan, & Wicherts, 2006).

invariance of subtests' intercepts should also be studied. This chapter was motivated by the fact that tests of measurement invariance in CFA are often not conducted to their full potential. As we saw from our re-analysis of the data of a Dutch intelligence test for minority children, this may have serious consequences. Unfortunately, tests of the equality of intercepts within this area are quite rare. For instance, we know of only one rigorous test of intercept differences in studying minority test performance on Dutch intelligence tests (Dolan et al., 2004). The results of that study indicated strong intercept differences, which indicated a construct irrelevant lowering of test performance of minorities on the GAT-B. The result of the current re-analysis also suggests that intercepts in several intelligence subtests of the RAKIT are lower for Dutch minorities. It is disconcerting that in the Netherlands both the RAKIT and the GAT-B are used widely for minorities in education, personnel selection, and in clinical settings.

Detecting intercept differences between groups should be an essential part of the validation of tests. Yet, to our knowledge, in the development of test batteries such as the WAIS-III or the WISC-IV, intercept differences across demographic groups are generally not studied. This implies that we cannot be certain that such tests are actually measurement invariant across groups. Unfortunately, the same applies to the majority of measurement invariance studies published in 2005, because in most of these the possibility of intercept differences was ignored. There are many advantages attached to the use of MGCFA with mean structure in testing measurement invariance. First, the approach is very flexible. For instance, variables that may account for measurement bias are easily incorporated in a factor model (Lubke et al., 2003a; Oort, 1992), which enables the understanding of the sources of bias and the eventual reduction of unfairness. Establishing why measurement bias occurs may contribute to more efficient test development. Second, uniform bias is perhaps the most obvious form of bias and it is easy to detect. The power to detect uniform bias in the common factor model is relatively large (Lubke et al., 2001).

Concluding Remarks

The use of DIF analyses in test development and test validation has become standard practice. Unfortunately, this still could not be said about tests for intercept differences in MGCFA, despite the fact that the CFA is commonly used. Intercept differences can have strong effects on test scores. Fortunately, however, intercept differences are easily detectable by means of MGCFA. Our hope is that a better understanding of the meaning of intercept differences and of ways to detect them, may contribute to the understanding of group differences in test scores, thereby increasing the fair use of tests.

3

Stereotype threat and group differences in test performance: A question of measurement invariance

Studies into the effects of stereotype threat (ST) on test performance have shed new light on race and sex differences in achievement and intelligence test scores. This chapter relates ST theory to the psychometric concept of measurement invariance, and shows that ST effects may be viewed as a source of measurement bias. As such, ST effects are detectable by means of multi-group confirmatory factor analysis. This enables research into the generalizability of ST effects to real-life or high-stakes testing. The modeling approach is described in detail, and applied to three experiments in which the amount of ST for minorities and women was manipulated. Results indicated that ST results in measurement bias of intelligence and mathematics tests.

3.1 Introduction

"The greatest social benefit will come from applied psychology if we can find for each individual the treatment to which he can most easily adapt. This calls for the joint application of experimental and correlational methods." (Cronbach, 1957, p. 679)

Recent developments in experimental social psychology concerning the effects of stereotypes on test performance have contributed to the understanding of the nature of race and sex differences in achievement and intelligence test scores. Specifically, the theory of stereotype threat (Steele, 1997) states that stereotypes concerning the ability of groups (e.g., women are bad at mathematics) can have an adverse impact on test performance of members of such groups, particularly in those who identify strongly with the domain of interest (e.g., female math students). Considering the widespread use of achievement and intelligence tests in college admission and job selection, and the high stakes involved in their use, stereotype threat effects on test performance may have serious personal and social consequences. There is general agreement on the importance of fair, unbiased, assessment in the sense that individual latent abilities should be measured validly and accurately. This means that *measurements* of ability should not depend on group membership based on, for instance, ethnicity or sex. Therefore, the absence of measurement bias with respect to groups (i.e., measurement invariance) is an essential aspect of valid measurement (e.g., Millsap & Everson, 1993). Both research into stereotype threat and research into measurement invariance are aimed at disentangling measurement artifacts related to group

membership from individual differences in the construct that a particular test is supposed to measure (e.g., latent mathematics ability). The aim of the current chapter is to explicitly relate stereotype threat to the concept of measurement invariance, and to show that stereotype threat effects on test performance may be viewed as a source of measurement bias.

This conceptualization of stereotype threat effects has statistical as well as practical advantages. It gives rise to an analytical framework in which individual and group differences in latent abilities and (experimental) stereotype threat effects on test performance can be modeled simultaneously. Of more practical importance is the fact that tests for measurement invariance with respect to groups can shed light on the degree to which stereotype threat plays a role in real-life and high-stakes settings. This provides a means to study the effects of stereotype threat in settings in which it is ethically and pragmatically difficult to manipulate the debilitating effects of stereotype threat on test performance (Cullen, Hardison, & Sackett, 2004; Sackett, 2003; Steele & Davies, 2003; Steele et al., 2002).

Below, we first discuss some methodological and statistical issues concerning experimental tests of stereotype threat effects on test performance. Next, we relate the effects of stereotype threat to measurement invariance, and discuss how such effects can be detected by means of multi-group confirmatory factor analysis. Finally, we illustrate this approach by analyzing the results of three experiments in which the effects of stereotype threat on the test performance of stigmatized groups were investigated.

3.2 Investigating Stereotype Threat Effects

The experimental paradigm, which is used to study the effect of stereotype threat on test performance, usually involves the comparison of existing groups (e.g., Blacks and Whites) and the manipulation of stereotype threat. The latter is accomplished, for instance, by labeling a test as either diagnostic or non-diagnostic for the stereotyped ability (e.g., Steele & Aronson, 1995, Study 2), or by asking for biographical information either prior to, or after completion of the test (e.g., Steele & Aronson, 1995, Study 4). Stereotype threat is expected to negatively affect test performance of stigmatized groups, but to have no (or a small positive; see Walton & Cohen, 2003) effect on test performance of non-stigmatized groups. Stereotype threat theory thus predicts an interaction between group and threat manipulation.

Generalizability of Stereotype Threat

Within laboratory experiments stereotype threat has been found to depress scores on various achievement and intelligence tests, in diverse stigmatized groups (Steele et al., 2002). The extent to which stereotype threat generalizes to test settings outside the laboratory is an important issue. Only few experimental studies have looked into the debilitating effects of stereotype threat on test performance in test settings high in ecological validity, and/or settings with consequential test outcomes. Stricker and Ward (2004) conducted two field studies within an actual high-stakes test situation, but were unable to replicate the strong negative effects of asking for biographical information prior

to taking a test (i.e., group prime) on minority and female test performance (cf. Steele & Aronson, 1995). In addition, three recent laboratory experiments addressed the effects of stereotype threat on Blacks' test performance in a job selection context (McFarland, Lev Arey, & Ziegert, 2003; Nguyen, O'Neal, & Ryan, 2003; Ployhart, Ziegert, & McFarland, 2003). In these studies, test-taking motivation was enhanced by the promise of financial rewards for high test scores. Despite the use of manipulations with well-established effects (i.e., race prime and test-diagnosticity), the debilitating effects of stereotype threat on minority test performance were generally absent. Sackett (2003) suggests that these results imply that the generality of stereotype threat effects to (motivational) job selection contexts is limited. Along similar lines, Stricker and Ward (2004) suggest that their studies indicate that high test stakes appear to be capable of overriding the negative effects of stereotype threat on test performance.

From a theoretical point of view, however, the internal validity of these real-life or contextualized experiments appears questionable. Steele and colleagues argue that stereotype threat probably always occurs within such settings, because of features that have been shown to elicit stereotype threat in the laboratory (Steele & Davies, 2003; Steele et al., 2002). For instance, promising incentives or placing a test in a selection context makes a test diagnostic for the stereotyped ability, thereby triggering stereotype threat even within control conditions. Heightening stereotype threat by means of explicit test diagnosticity or group prime then fails to depress test performance of stigmatized groups much further, resulting in ineffective stereotype threat manipulations (Steele & Davies, 2003; Steele et al., 2002). In that respect, stereotype threat theory predicts that stereotype threat studies, which are high in ecological validity, are low in internal validity, and vice versa. More importantly, whereas inductive reasoning leads one to expect that most real-life test settings do evoke stereotype threat, empirically the question of generalizability appears hard to answer (Steele et al., 2002).

3.3 Analyzing Stereotype Threat Effects

Given the pragmatic and ethical problems of experimentation within real-life settings, correlational methodology (e.g., regression analysis) may be used to investigate the presence of stereotype threat on actual achievement tests. Osborne (2001) reasoned that stereotype threat effects may be mediated by anxiety (cf. Blascovich, Spencer, Quinn, & Steele, 2001). He found that the racial gap, and to a lesser extent, the gender gap on several achievement tests in the High School and Beyond Study were partly mediated by self-reported anxiety, which supports the notion that stereotype threat affected test performance. Cullen et al. (2004) proposed that the strong identification of high-ability persons with the domain of interest (cf. Steele, 1997), renders them more sensitive to stereotype threat (Aronson et al., 1999). They reasoned that if stereotype threat affects test performance of stigmatized groups on a predictor (e.g., SAT), this differential sensitivity to stereotype threat would lead to group-specific and non-linear relations between the affected predictor and criteria that are supposedly *un*affected by stereotype threat, such as job performance or grade points of classes unrelated to stereotypes. However, Cullen et al. (2004) found neither prediction bias, nor any non-linear effects, and concluded that

stereotype threat effects on the predictors (SAT and Armed Services Vocational Aptitude Battery) were small or non-existent.

These seemingly inconsistent results may be due to the strong assumptions underlying the use of such regression approaches. For instance, Cullen et al. (2004) had to assume the absence of group differences on academic criteria (cf. the underperformance phenomenon; Steele, 1997), whereas Osborne (2001) rightly expressed some concern about the causal link involved. Moreover, these correlational studies address the effects of stereotype threat on test performance in an *indirect* manner. It is well established that group differences in prediction (i.e., prediction bias) do not necessarily imply that measurements are biased with respect to groups, and vice versa (Millsap, 1997a).

Measurement Models

The indirectness of these regression approaches can be avoided by adopting measurement models that explicitly relate test scores to the latent constructs that are supposed to underlie those test scores. Instead of the latent abilities, stereotype threat affects the test scores in a group-specific manner. As we shall see below, a comparison of stigmatized and non-stigmatized groups with respect to the test scores-construct relationship (i.e., test for measurement invariance) allows for a *direct* study of the presence of stereotype threat effects within a particular test situation.

An additional advantage of using measurement models is that they can be used to analyze experimental data (cf. Donaldson, 2003), thereby overcoming some difficulties associated with traditional use of analysis of variance within stereotype threat experiments. The groups under investigation in such studies are expected to differ considerably with respect to the latent ability that is supposed to underlie the dependent variable(s) (i.e., test scores). This may give rise to analytical problems because of pre-existing group differences in the average or variability of latent ability (e.g., gender differences in math variability; Hedges & Nowell, 1995). In numerous stereotype threat studies, prior test scores (e.g., SAT) and analysis of covariance or ANCOVA are used to equate groups for mean differences in ability. However, as we argue in Appendix C, several expectations derived from stereotype threat theory do not sit well with the assumptions underlying the traditional use of ANCOVA (see also Yzerbyt, Muller, & Judd, 2004). For instance, stereotype threat may lower the regression weight of the dependent variable on the covariate in the stereotype threat condition, which violates regression weight homogeneity over all experimental cells (cf. Appendix C). The use of statistical methods that differentiate between the construct (i.e., latent ability) and the measurement of that construct circumvents such problems. More importantly, measurement models equip us with ways to test for measurement invariance.

3.4 Measurement Invariance

Measurement invariance revolves around the issue of how groups differ in the way the measurement of a psychological construct (e.g., mathematics test score) is related to that construct (e.g., mathematical ability). Measurement invariance means that measurement bias with respect to groups is absent (Lubke et al., 2003a, 2003b; Meredith, 1993). Below,

we explain measurement invariance conceptually in relation to stereotype threat. Let us first look at the formal definition of measurement invariance (Mellenbergh, 1989), which is expressed in terms of the conditional distribution of manifest test scores Y (denoted by $f(Y|.)$). Measurement invariance with respect to v holds if:

$$f(Y \mid \eta, v) = f(Y \mid \eta), \qquad \text{(for all } Y, \eta, v\text{)}, \qquad (1)$$

where η denotes the scores on the latent variable (i.e., latent ability) underlying the manifest random variable Y (i.e., the measured variable), and v is a grouping variable, which defines the nature of groups (e.g., ethnicity, sex). Note that v may also represent groups in experimental cells such as those that differ with respect to the pressures of stereotype threat. Equality (1) holds if, and only if, Y and v are conditionally independent given the scores on the latent construct η (Lubke et al., 2003b; Meredith, 1993).

One important implication of this definition is that the expected value of Y given η and v should equal the expected value of Y given only η. In other words, if measurement invariance holds, the expected test score of a person with a certain latent ability (i.e., η) is *independent* of group membership. Thus, if two persons of a different group have exactly the same latent ability, they must have the same (expected) score on the test. Suppose v denotes sex and Y represents the scores on a test measuring mathematics ability. If measurement invariance holds, then test scores of male and female test takers depend *solely* on their latent mathematics ability (i.e., η)[15] and *not* on their sex. Then, one can conclude that measurement bias with respect to sex is absent, and that manifest test score differences in Y correctly reflect differences in latent ability between the sexes.

However, the situation changes when stereotype threat impacts test performance. Suppose v represents two groups (e.g., Blacks and Whites) that differ with respect to stereotypes that concern Y (e.g., intelligence tests). If stereotype threat directly affects (i.e., lowers) the observed scores (i.e., Y) in the Black group (or in a sub-sample of this group), then measurement invariance is violated. The reason for this is that conditioning on the latent construct (i.e., latent ability) does not remove all group differences in Y, because of the debilitating effects of stereotype threat on Y, which are limited to the Black group. This becomes particularly clear if one images a Black test taker with a particular latent ability, who, because of stereotype threat, underperforms in comparison to a White test taker with the same latent ability. Clearly, the relationship between test score and latent ability now depends on group membership and the requirements for measurement invariance no longer hold. Therefore, stereotype threat effects are *by definition* a source of measurement bias. Conversely, if measurement invariance holds in a particular group comparison, stereotype threat does not play a differential role in test score differences between those groups, because then test score differences rightly reflect group differences in the latent construct.

The definition of measurement invariance is quite general (Mellenbergh, 1989). It does not depend on the kind of test, selection variable, or the size of group differences in latent ability. Although measurement invariance may be investigated by many methods (Millsap & Everson, 1993; Raju et al., 2002) using different types of measurement models

[15] However, measurement invariance with respect to one selection variable does not necessarily imply measurement invariance with respect to another selection variable (but see Lubke et al., 2003b).

(e.g., item response models), we restrict our attention to the confirmatory factor model. We now present this model, relate it to measurement invariance, and show how stereotype threat may result in measurement bias. After that, we investigate in three studies whether experimental stereotype threat effects indeed lead to measurement bias.

3.5 Multi Group Confirmatory Factor Analysis (MGCFA)

Here we describe the measurement model (i.e., MGCFA) in a non-technical fashion, restricting our attention to the one factor case, and assuming multivariate normality throughout. Appendix A contains a more technical and more general presentation of the model (see also Bollen, 1989; Dolan, 2000; Dolan et al., 2004; Lubke et al., 2003a). The confirmatory factor model is essentially a linear regression model in which scores on several indicators (i.e., subtest scores) are regressed upon scores on the latent (i.e., unobserved) construct η. Like in ordinary regression, the model includes for each indicator the following measurement parameters: a regression weight or factor loading (expressed by the symbol λ), a residual term, and an intercept. The residual term of an indicator is expressed by the symbol ε, and contains both random measurement error and specific factors tapped by that particular indicator (i.e., all uncommon sources of variance; Meredith & Horn, 2001). In most applications of confirmatory factor analysis (e.g., one-group studies), the regression intercept is uninformative and is not modeled. However, we are also interested in studying between-group differences in means. Therefore, we add the mean structure to the analysis, which is accomplished by incorporating an intercept term for each indicator, expressed by τ (Sörbom, 1974). The extension to multiple groups enables tests of specific hypotheses concerning between-group differences in measurement parameters (i.e., measurement bias) and between-group differences in the parameters that describe the distribution of the common factor within each group (i.e., group differences in mean latent ability). The simultaneous analysis of covariance[16] and mean structure provides a test of measurement invariance, or *strict factorial invariance,* as it is denoted in this context (Meredith, 1993).

The model for subtest score Y_1 of a person j in group (or condition) i is as follows:

$$y_{1ij} = \tau_{1i} + \lambda_{1\eta i}\eta_{ij} + \varepsilon_{1ij}. \quad (2)$$

Suppose we have four subtests. Of course, the latent ability score η_{ij} of person j is the same for all subtests, so we can conveniently arrange the expressions using vector notation (e.g., Bollen, 1989):

$$\begin{bmatrix} y_{1ij} \\ y_{2ij} \\ y_{3ij} \\ y_{4ij} \end{bmatrix} = \begin{bmatrix} \tau_{1i} \\ \tau_{2i} \\ \tau_{3i} \\ \tau_{4i} \end{bmatrix} + \begin{bmatrix} \lambda_{1\eta i} \\ \lambda_{2\eta i} \\ \lambda_{3\eta i} \\ \lambda_{4\eta i} \end{bmatrix} \times \left[\eta_{ij}\right] + \begin{bmatrix} \varepsilon_{1ij} \\ \varepsilon_{2ij} \\ \varepsilon_{3ij} \\ \varepsilon_{4ij} \end{bmatrix}. \quad (3)$$

This, in turn, is more parsimoniously expressed by the following matrix notation:

$$y_{ij} = \tau_i + \Lambda_i\eta_{ij} + \varepsilon_{ij}. \quad (4)$$

[16] We are also interested in and should allow for possible differences in variances between the groups. For that reason, in MGCFA covariance matrices are analyzed instead of correlation matrices.

Except for the difference in notation Equations 3 and 4 are identical. For example, in Equation 4, τ_i is a 4 dimensional vector containing the measurement intercepts and Λ_i is a 4 x 1 matrix containing the factor loadings of group i. Equation 4 presents a model for the observations. To obtain estimates of the parameters in this model, we fit the observed covariance matrices and mean vectors to the implied (by Equation 4) covariance matrices and mean vectors (cf. Appendix A). The parameters of interest are the factor loadings (Λ_i), the vector of intercepts (τ_i), the variances of the residuals, incorporated in a matrix denoted Θ_i, and the means and variances of the common factor scores in group i, denoted by α_i and Ψ_i, respectively. In fitting the model, we introduce two types of constraints: identifying constraints, which are required in all confirmatory factor analyses (e.g., scaling; Bollen, 1989), and substantive constraints, which relate specifically to the issue of measurement invariance (Meredith, 1993). As we explain next in a two-group context, the latter concern the factor loadings, intercepts, and residual variances.

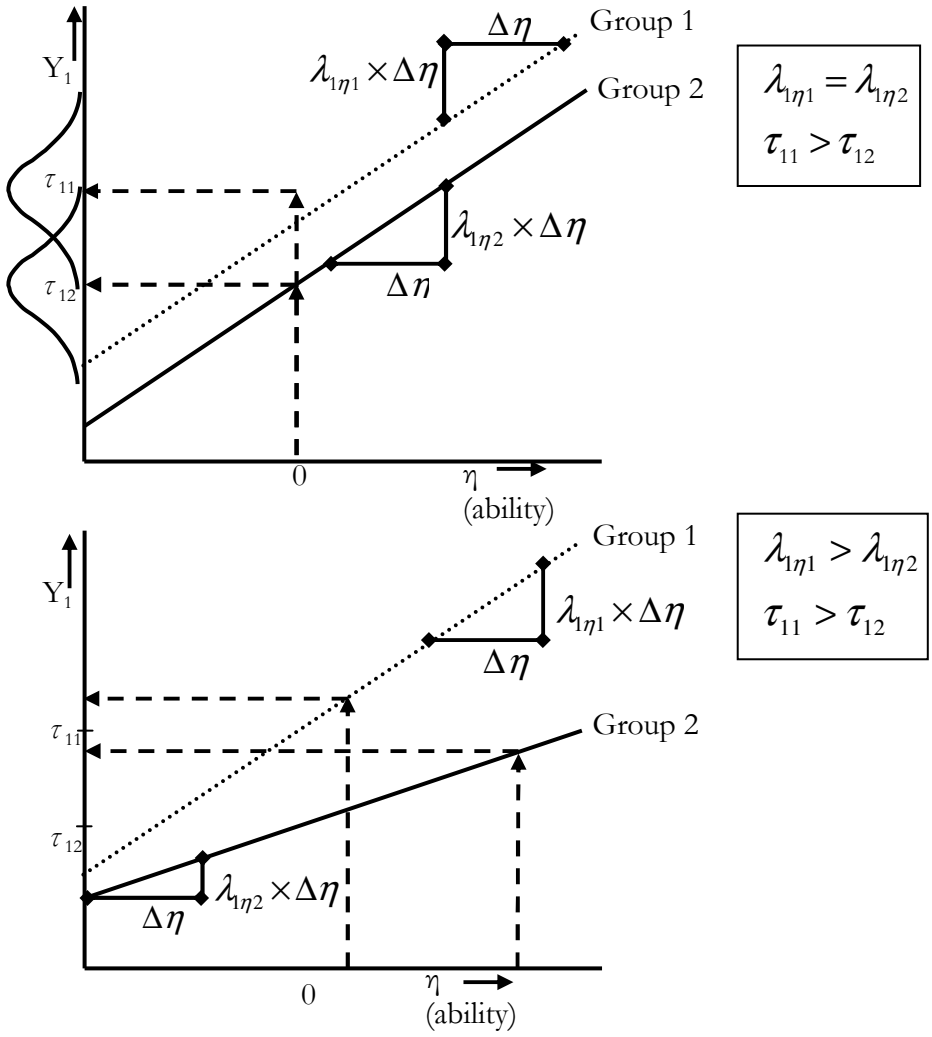

Figure 3.1 *Regression lines of test scores on latent variable in two groups with different intercepts (top) and different intercepts as well as different factor loadings bottom).*

Consider the top half of Figure 3.1. Here we see the regression lines for subtest Y_1 in two groups. The factor loading gives the slope of this line (for each increment Δ of latent ability η, the expected test score changes by $\Delta\eta$ times λ) and the intercept τ gives the point of Y_1 associated with the point $\eta = 0$. Also depicted are the normally distributed residuals in each group. Note that the residual variances appear equal in both groups. As can be seen, the regression slopes (i.e., factor loadings) are also equal in both groups. However, the intercepts differ over groups. This has serious consequences. Namely, for each possible latent factor score, the expected value on the test Y_1 is *higher* for members of Group 1 than for members of Group 2.[17] Clearly, this violates measurement invariance with respect to both groups. Hence, the equality of measurement intercepts (i.e., $\tau_{11} = \tau_{12}$) is an essential requirement for measurement invariance (cf. Meredith, 1993). The reader may have already guessed a possible source for such an intercept difference between groups: the uniform (i.e., irrespective of latent ability) depression of test scores due to stereotype threat in Group 2.

The bottom half of Figure 3.1 displays another two-group scenario. Here, the regression lines for both groups again show different intercepts. In addition, the slope of the regression line in Group 2 now differs from the slope in Group 1. Specifically, the factor loading in Group 2 is lower (i.e., $\lambda_{1\eta 1} < \lambda_{1\eta 2}$). This means that in Group 2 the test scores do not measure latent ability as well as in Group 1. Again, given a particular latent factor score, the expected test score depends on group membership. Even worse, the negative effect of "being" a Group 2 member now depends on the particular latent ability level. Higher ability scores result in more bias than lower ability scores. As is graphically depicted by the dashed arrows, it is even conceivable that a member of Group 2 with a fairly high ability score has an expected test score below that of someone in Group 1, who has a considerably lower ability. Clearly, for measurement invariance to hold between groups, factor loadings should be equal across groups (i.e., $\lambda_{1\eta 1} = \lambda_{1\eta 2}$). Note that a depressed factor loading could be due to stereotype threat affecting test performance in Group 2 in a non-uniform manner. Again, the lowering of the intercept may be viewed as a "main effect" for stereotype threat. Moreover, the lowering of a factor loading in Group 2 can be interpreted as an "interaction effect" between stereotype threat and latent ability on test performance. The latter may occur because domain identification is known to heighten stereotype threat effects, and domain identification may be strongly related to latent ability (Cullen et al., 2004; Steele, 1997). In such a scenario higher ability persons suffer more under stereotype threat, resulting in a depressive effect on the factor loading.

We have presented a graphic exposition of why factor loadings and measurement intercepts need to be invariant for measurement invariance to hold. In fact, under measurement invariance, the regression lines of each group coincide. If so, the expected value of the test scores depends solely on latent ability, regardless of group membership. An additional requirement for strict factorial invariance is that residual variances need to be invariant. This is because residual variances contain all uncommon sources of variance.

[17] Note the resemblance of this picture to what Steele (1997, p.626) called the parallel lines phenomenon when he referred to the academic underperformance of Black college students in comparison to White college students with equal standardized test scores. The differences lie in that Steele's predictor was a standardized test score and his criterion was first-year GPA, whereas our predictor is the latent ability score and the criterion is the test score.

Larger residual variance in one group means less reliable measurement. Moreover, added residual variance may also be due to stereotype threat variance. Meredith (1993) provides a rigorous statistical discussion of why group-invariant factor loadings (Λ), residual variances (Θ), and intercepts (τ) are essential requirements for strict factorial invariance. Indeed, if measurement invariance holds, as defined above (Equation 1), these equality constraints should hold to reasonable approximation (Meredith, 1993; Millsap, 1997b).

The Stereotype Threat "Factor"

In order to better understand the specific effects of stereotype threat on measurement parameters, it is convenient to imagine the presence an additional common factor (denoted by σ), which incorporates all the mediating effects of stereotype threat on test performance. Such an additional stereotype threat "factor" is neither measured nor modeled, but it still affects test performance in a manner that is restricted to the stigmatized group, resulting in group-specific changes of measurement parameters. Hence, constraining measurement parameters of a group under stereotype threat to group(s) without such effects (i.e., non-stigmatized group and/or control condition) would demonstrate a violation of strict factorial invariance. It is well-established that stereotype threat specifically affects performance on the more *difficult* tasks (Blascovich et al., 2001; O'Brien & Crandall, 2003; Quinn & Spencer, 2001; Spencer et al., 1999; Steele et al., 2002). Therefore, we expect the effects to be subtest-specific and mostly related to the most difficult subtests in a test battery.

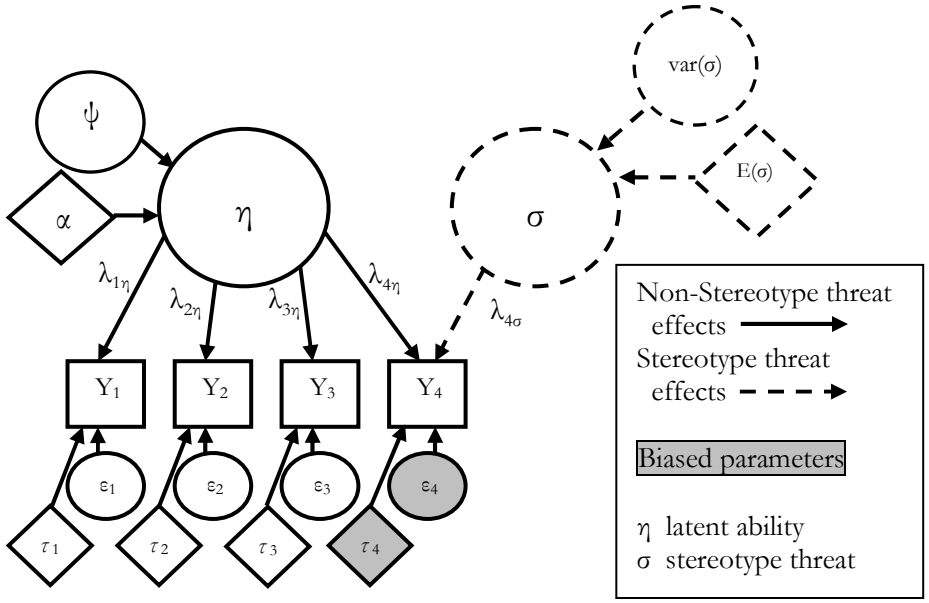

Figure 3.2 *Effects of stereotype threat on parameter estimates of affected Subtest Y_4.*

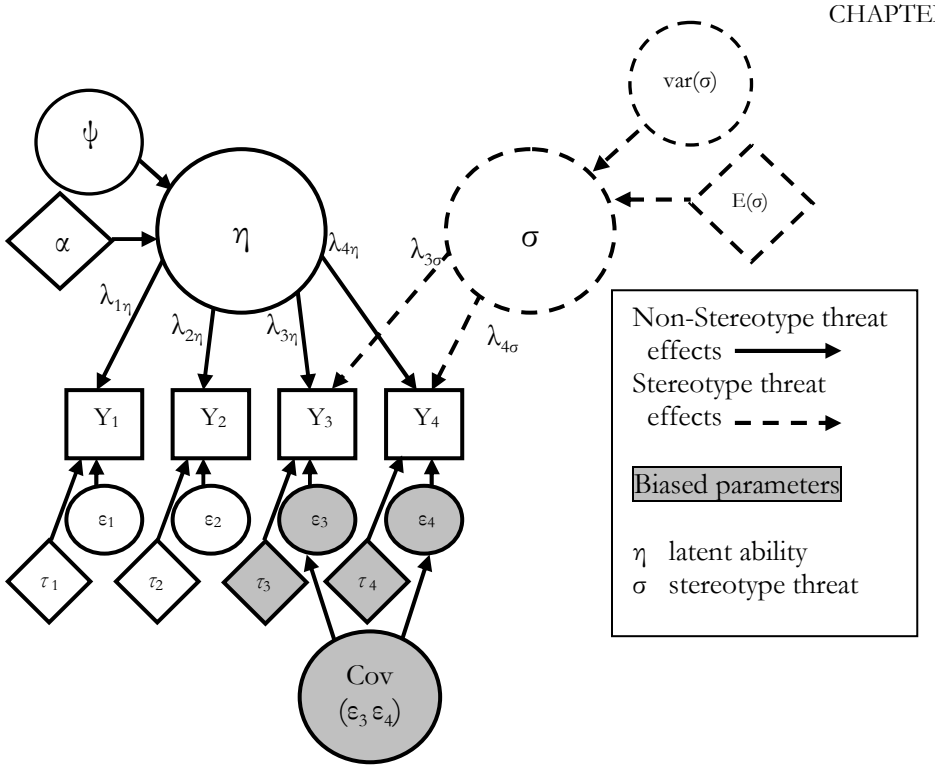

Figure 3.3 *Effects of stereotype threat on parameter estimates of affected Subtests Y_3 and Y_4.*

Figure 3.2 displays the common factor model within a group (in a particular setting), where stereotype threat affects the scores on subtest Y_4 (conceivably a particularly difficult subtest). As we show in Appendix B, such a stereotype threat effect results in a lowering of the measurement intercept of the affected subtest (cf. Figure 3.1 top half). In addition, if stereotype threat effects vary over persons within this group, perhaps because of individual differences in domain identification or group identification, the variance due to the unmeasured stereotype threat factor results in an increase of the residual variance of subtest Y_4. However, it is also conceivable that two of the four subtests are affected by stereotype threat. This situation is displayed in Figure 3.3. Again, this would result in negative effects on the measurement intercepts of these subtests (cf. Appendix B). In addition, if stereotype threat effects vary over persons, this would lead to increased residual variances of both affected subtests. Furthermore, the two affected subtests now covary more strongly than would be expected from their corresponding factor loadings on the η factor. This additional covariance due to stereotype threat constitutes a violation of the dimensionality of the factor model within this group (i.e., residual covariance), resulting in model misfit. This scenario of stereotype threat affecting the performance on two subtests can be extended to cases in which more than two (or even all) subtests are affected. Of course, in such cases, stereotype threat also violates strict factorial invariance.[18]

[18] However, if a (relatively) large number of subtests are affected by stereotype threat, model misfit due to such stereotype threat effects disperses over the model. This makes it difficult to interpret measurement bias in terms of sole parameters.

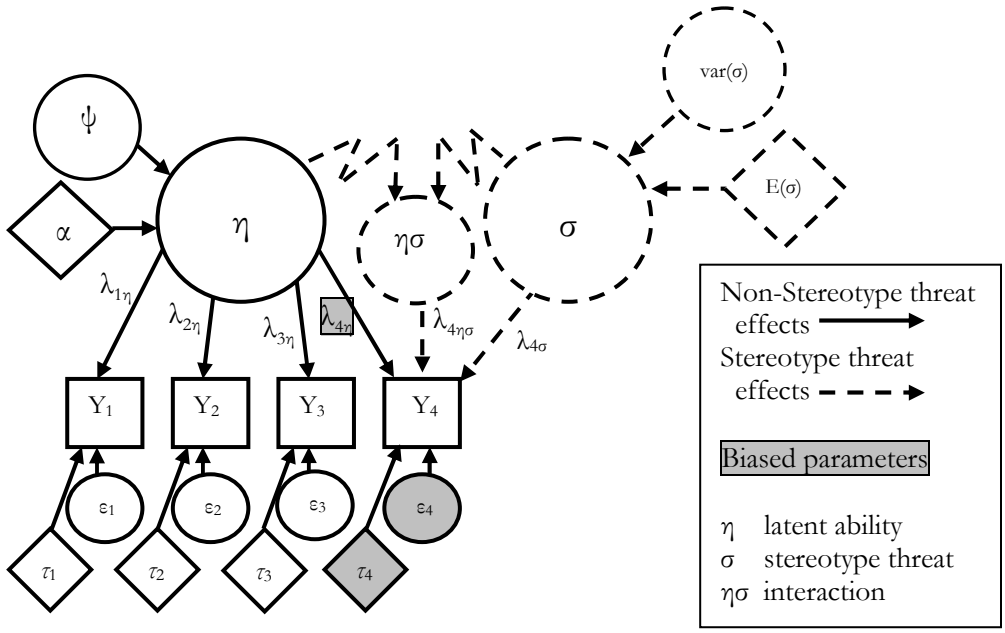

Figure 3.4 *Non-uniform effect of stereotype threat on parameter estimates of affected subtest Y_4.*

As a final scenario, consider Figure 3.4 in which a nonlinear effect on subtest Y_4 is expressed as an interaction between latent ability and stereotype threat. As is shown in Appendix B, such an effect results in a lowering of the factor loading of subtest Y_4 (cf. Figure 3.1 bottom half). Additionally, one would expect an increase of the residual variance of the affected subtest and a downward bias of the intercept.

In conclusion, the effects of stereotype threat are detectable by tests for measurement invariance using MGCFA. Possible stereotype threat effects would show up particularly in group differences in the measurement parameters of difficult subtests. We now turn to three experiments in which the amount of stereotype threat for stigmatized groups was manipulated. We thus use tests of strict factorial invariance with respect to groups and conditions to identify the effects, if any, of stereotype threat on the test scores.

3.6 Study 1: Dutch Minorities and the Differential Aptitude Test

On average, Dutch minority students attain lower educational levels and have a higher drop-out rate than Dutch majority students (Dagevos, Gijsberts, & van Praag, 2003). Several studies have indicated that Dutch high school students often view minority students as less smart (Verkuyten & Kinket, 1999), and minority groups as less well educated (Kleinpenning & Hagendoorn, 1991). Recently, Verkuyten and Thijs (2004) found that academic disidentification among Dutch minority students was moderated by the perception of being discriminated in scholastic domains. The first aim of Study 1 was to study the effects of stereotype threat on intelligence test performance in a sample of minority high school students in the Netherlands. To this end, we administered a short

intelligence test containing three subtests, and varied the amount of stereotype threat related to ethnic minorities by changing the presentation of the test and by altering the timing of an ethnicity questionnaire. The second aim of this study is to find out whether tests for measurement invariance using MGCFA can successfully highlight the effects of stereotype threat. Furthermore, we compare the results of confirmatory factor analysis to the results of analysis of variance, in order to find out whether both analyses lead to the same conclusions.

Method

Participants

Two hundred and ninety five students from nine high schools in large cities in the Netherlands participated during obligatory classes, which were aimed at counseling the students in choosing a major ("profile") in the second phase ("tweede fase") of their high school education. The students were aged between 13 and 16 (M = 14.86, SD = 0.64), and attended the third year of the HAVO education. Given that the HAVO level is the second-highest level in the Dutch high school system, the sample is expected to be heterogeneous in terms of identification with the academic domain, which is considered an important moderator of stereotype effects (cf. Aronson et al., 1999).

All 157 students in the majority group were born in the Netherlands, as were all their parents and grandparents. Of the 138 minority students, most were born in The Netherlands (76%), but all of them had one (10%) or two (90%) parents born outside The Netherlands. The (grand)parents of the minority students were immigrants from (former) Dutch colonies (Surinam/Antilles; N = 47), Turkey (N = 36), or Morocco (N = 55).[19] Because of the absence of large test score differences between these minority groups, and to increase the sample sizes, these minority groups are pooled.[20] When asked to indicate the cultural group they identified with, most (N = 93; 67%) of the minority students indicated their own minority group. Twenty-three minority students (17%) indicated the Dutch majority group, and 22 minority students (16%) indicated both the Dutch group, and their minority group as the group they identified with. The total sample consisted of 119 boys and 176 girls. Both ethnic groups did not differ in sex and age composition.[21]

Procedure and Design

Three shortened subtests of the Differential Aptitude Test or DAT (Evers & Lucassen, 1992) were administered during classes, which were attended by 17 to 27 students. Upon arrival in the classroom, students found a test booklet on their desks, and a female tester of Dutch origin told them that they would be taking a counseling test. The

---

[19] These data stem from a larger study containing 430 students (Wicherts et al., 2003). We selected only students that could be categorized unambiguously in the majority group (student, his/her parents, and grandparents are all born in The Netherlands) or in one of these three minority groups.

[20] Although there may be differences between these minority groups in terms of *general* stereotypes, in terms of *academic* stereotypes differences between these groups are quite small (see, e.g., Kleinpenning & Hagendoorn, 1991).

[21] To ensure the existence of stereotypes concerning the intellectual ability of minority groups, we conducted a pilot study in which we asked a group of 41 students in comparable schools and classes whether they believed that there existed prejudices concerning the intellectual ability of their cultural group (direction unspecified). On a scale from 1 (no prejudice) to 5 (strong prejudice), the 20 majority students (M = 2.00, SD = 1.12) scored significantly lower ($t(39) = 4.53$, $p < 0.001$) than the 21 minority students (M = 3.62, SD = 1.16), indicating that the minority students reported a strong awareness of the stereotypes concerning the intellectual abilities of their group.

tester said that the test booklet contained questions about their personal interests and abilities, and that their answers would be used for guidance in their choice of specialty. No explicit mention of intelligence was made. The tester told the students that the test booklet consisted of several sections, and that they would be told when to start and stop with a particular section. This enabled the timing of each of the following sections of the test booklet: An ethnicity questionnaire, the DAT tests, an interest inventory, an additional language test (used for exploratory purposes), and the actual profile-counseling test (administered last). After the test session, students were debriefed extensively on the purpose of the experiment. After a week, all students received written counseling on their specialty choice, which was based solely on the profile-counseling test (cf. Zand Scholten, 2003). Special care was taken to ensure that the answers on this test were not affected by the stereotype threat manipulation, or by ethnicity (Wicherts et al., 2003).

Participants were assigned to two conditions that differed in the features that elicit stereotype threat for the minority students. Assignment to conditions was achieved by randomly distributing two versions of a test booklet, which were indistinguishable by the cover. In the stereotype threat condition, this test booklet presented each DAT subtest as an "intelligence test". The test booklet of participants in the control condition made no mention of intelligence, and the tests were simply presented as a section of the test booklet. In addition, in the stereotype threat condition, an ethnicity questionnaire was administered prior to the DAT. This questionnaire consisted of 14 questions concerning ethnic and cultural background (religion, language use), and questions about place of birth of the students, their parents, and grandparents. In the control condition, the ethnicity questionnaire was administered after the DAT. While participants in the stereotype threat condition filled in the ethnicity questionnaire, participants in the control condition filled in an interest inventory containing 15 items without any connection to ethnicity. This interest inventory was administered to students in the stereotype threat condition after the intelligence tests. Thus, two stereotype threat manipulations were employed in concert to increase stereotype threat for ethnic minorities: an ethnicity prime and a manipulation of the diagnosticity of the intelligence test (cf. Steele & Aronson, 1995).

Intelligence Test

Three subtests of the Dutch DAT (Evers & Lucassen, 1992) were used as a measure of general intelligence. The subtests were shortened by selecting items with the highest item-rest correlations in the Dutch standardization sample (N = 2100). The Numerical Ability test (NA; originally 40 items, 25 min) contains 14 complicated mathematic problems. Abstract Reasoning (AR; originally 45 items, 25 min) contains 18 items with a logical sequence of diagrams, which had to be completed. Verbal Reasoning (VR; originally 50 items, 20 min) contains 16 verbal analogy items. All subtests were administered with a time limit of six minutes. All items have a 5-option multiple-choice answer format. Based on the standardization data, Numerical Ability is the most difficult subtest in terms of proportion correct of the items retained in the short version (average p-value 0.43), followed by Verbal Reasoning (average p-value 0.49) and Abstract Reasoning (average p-value 0.59). Thus, one would expect the strongest stereotype threat effects on the Numerical Ability test. The instruction pages of the subtests were slightly adapted with regard to the time limit, number of items, and the presentation of the tests as either a

section (control condition), or as an intelligence test (stereotype threat condition). To correct for possible order effects, and to avoid cheating (e.g., copying answers), two order versions of the test booklet were employed (bringing the total number of versions to 4). The order in these two versions was NA-AR-VR, and VR-NA-AR, respectively. Because none of the main or interaction effects for order reached significance (ANOVA; all Ps > 0.10), these order versions are pooled for the subsequent analyses.

Analyses

Considering previous factor analyses on the complete DAT (Evers & Lucassen, 1992; Te Nijenhuis et al., 2000; Wicherts et al., 2004), the use of a one-factor model for these three subtests is sensible. Although our primary interest lies in testing for strict factorial invariance with respect to groups, we also conduct a 2-by-2 MANOVA, with stereotype threat and ethnicity as factors, and the three subtests as dependent variables. MANOVA provides a means to interpret the experimental mean effects. We predicted a significant main effect for ethnicity, with majority students outscoring the minority students (see, e.g., Te Nijenhuis et al., 2000). In addition, we expected a significant ethnicity by condition interaction, because stereotype threat would primarily depress scores of minority students. Given the heterogeneous sample used, we also expected heterogeneity in covariances and variances over design cells. Therefore, as is common in the (M)ANOVA framework, we also conduct tests for variance and covariance heterogeneity by means of Box's M test and the univariate Levene's test.

MGCFA can be used to shed light on the nature of differences in (co)variance and mean structure between groups. Within this two-by-two experimental design, the tenability of strict factorial invariance with respect to groups and conditions (i.e., 4 groups) is investigated by fitting a series of increasingly restrictive models. These models as well as the restrictions imposed are presented in Table 3.1. In the first step, no between-group restrictions are imposed. The next steps involve restricting all factor loadings (Step 2) and all residual variances (Step 3) to be invariant over all four groups. Because of the random assignment to experimental conditions, one does not expect there to be differences on the factor level between conditions for both *existing* groups. Step 4 can be used to investigate whether factor variance of the existing groups are affected by the stereotype threat manipulation. That is, in this step, the factor variance for majority students in the stereotype threat condition is restricted to be equal to the factor variance for majority students in the control condition (and similarly for the minority students). In Step 5, the invariance of the mean structure is investigated by restricting the measurement intercepts to be equal across all groups. In the same step, factor mean differences with respect to an arbitrary baseline group are estimated. Finally, in Step 6, the means of the existing groups are restricted to be equal over condition (e.g., factor mean of majority group in control condition equal to factor mean of majority group in stereotype threat condition). This ensures that the experimental manipulation has no effect on the mean of the common factor. As can be seen, if the restrictions implemented in these six steps hold, measurement invariance holds. In that case, the differences between the existing groups are a function of the differences in the means (α) and variances (Ψ) of the common factor. However, we expected the test scores to be affected in a differential manner across groups.

Table 3.1
Equality constraints imposed in the steps towards strict factorial invariance

No.	Description	Λ factor loadings	Θ residual variances	Ψ factor variance	τ intercepts	α factor mean
1	Configural invariance	-	-	-	-	-
2	Metric invariance	all groups	-	-	-	-
3	Equal residual variances	all groups	all groups	-	-	-
4	Factor variances invariant over condition	all groups	all groups	existing groups	-	-
5	Strict factorial invariance	all groups	all groups	existing groups	all groups	-
6	Factor means invariant over condition	all groups	all groups	existing groups	all groups	existing groups

Note: Each step is nested under the previous one. Underlined restrictions are tested in each step.

The tenability of each restriction is judged by differences in fit between the restricted model and the less-restricted model. For instance, Step 2 vs. Step 1 involves the tenability of equality of factor loadings. Because of the nesting of models, a likelihood ratio test is employed to test each restriction. Besides attention for chi-squares, the CFI and the RMSEA are used in determining the absolute and relative model fit. The Comparative Fit Index (CFI; Bentler, 1990) ranges from 0 to 1, and is a measure of the relative fit of a model in relation to a null model of complete independence[22]. The Root Mean Square Error of Approximation (RMSEA; Browne & Cudeck, 1993) is a so-called close fit measure that is known to be relatively insensitive to sample size. Several rules of thumb have been proposed for these fit measures. Based on their simulation study, Hu and Bentler (1999) proposed that RMSEA values smaller than 0.06, and CFI values larger than 0.95 are indicative of good model fit.

In case a step is accompanied by a clear deterioration in model fit, the particular restriction is rejected. In such cases, modification indices can highlight the particular parameter(s) causing the misfit. A modification index (MI) is a measure of how much chi-square is expected to decrease if a constraint on a given parameter is relaxed, and the model is re-fitted (Jöreskog & Sörbom, 1993). In cases where a restriction is accompanied by a deterioration in fit, parameters with the highest modification index are freely estimated and the sequence of models is continued. We expected that stereotype threat effects on test performance would result in measurement bias expressed by high modification indices in the minority group in the stereotype threat condition. All factor analyses were carried out using LISREL 8.54[23] (Jöreskog & Sörbom, 2003).

[22] Widaman and Thompson (2003) have argued that because of the nesting of models it is inappropriate to use such a null model within a multi-group context with mean structure. Therefore, we use a model without any factor structure, in which intercepts and residual variances are restricted to be group invariant (i.e., model 0A in Widaman & Thompson, 2003) as the null model in computing the CFI values.

[23] All input files used here can be downloaded from: http://users.fmg.uva.nl/jwicherts.

Results

Table 3.2

Means and standard deviations by experimental condition and ethnic group (Study 1)

	Condition							
	Control				Stereotype threat			
	Majority		Minority		Majority		Minority	
N =	79		65		78		73	
Subtest	M	SD	M	SD	M	SD	M	SD
Numerical	5.35	2.54	4.88	2.47	5.49	2.31	4.67	2.52
Abstract Reasoning	10.42	2.96	6.80	3.33	9.24	3.34	7.34	2.83
Verbal Reasoning	7.27	3.01	5.37	2.82	6.65	3.47	5.56	2.70

The values for univariate skewness and kurtosis in the four groups are in an acceptable range (i.e., [-0.89 - 0.88]), suggesting no large deviations from normality. Therefore, the use of Maximum Likelihood for estimating the factor models is justified. Table 3.2 contains means and standard deviations of the three subtests for both ethnic groups in the two conditions. First, we provide the analysis of variance results. Box's M test suggests some covariance heterogeneity over groups ($F(18, 28483) = 1.787$, $p < .05$). The univariate Levene's test for homogeneity of variance gives a significant value for the Verbal Reasoning subtest ($F(3, 291) = 3.63$, $p < .05$). Because MANOVA is often claimed to be robust to (co)variance heterogeneity (e.g., Stevens, 1996), we do interpret the results of the MANOVA. The multivariate main effect for ethnicity is significant ($F(3, 289) = 20.36$, $p < .001$), as well as all univariate effects (Numerical Ability: $F(1, 291) = 5.07$, $p < .05$; Abstract Reasoning: $F(1, 291) = 57.47$, $p < .001$; Verbal Reasoning: $F(1, 291) = 17.83$, $p < .001$), with the majority group outscoring the minority group. Neither the multivariate, nor any of the univariate main effects for condition reach significance (all Ps > .30). The multivariate interaction effect between condition and ethnicity is significant ($F(3, 289) = 2.642$, $p = .050$). The only significant univariate interaction effect is found on the Abstract Reasoning subtest ($F(1, 291) = 5.56$, $p < .05$). However, this interaction effect is due to the *majority* group underperforming in the stereotype threat condition. Namely, the condition simple effect is significant for majorities ($F(1, 155) = 5.45$, $p < .05$), but non-significant for minorities ($F(1, 136) = 1.07$, $p > .30$). All multivariate and univariate simple effects for condition within the minority group are non-significant (all Ps > .30), which is opposite to what one would expect from stereotype threat theory. Whereas, the minority group scored significantly lower than the majority group, these ANOVA results indicate that *on average* the minority students in the stereotype threat condition did not score lower than the minority students in the control condition.

However, it is important to stress that the sample may be expected to be heterogeneous with respect to domain-identification, considered an important moderator of stereotype threat effects (e.g., Steele, 1997). For instance, Aronson et al. (1999) found that test-takers that identified strongly with the domain of interest (i.e., mathematics) were more susceptible to stereotype threat, whereas test-takers who moderately identified with the domain performed *better* under stereotype threat conditions than under control

conditions. This suggests that within heterogeneous samples that contain both highly identified and moderately identified test-takers, effects of stereotype threat may differ substantively over persons. In such samples, positive and negative effects may cancel out, resulting in no, or only a small effect on the mean. However, the absence of a mean effect does not necessarily mean the absence of an effect. To investigate the possibility that covariance structure was affected by the stereotype threat induction, we tested for measurement invariance with respect to the four groups. The results of the multi-group confirmatory factor analyses are reported in Table 3.3.

Table 3.3
Fit measures of steps towards strict factorial invariance (Study 1)

Step	Restrictions	DF	χ^2	p	ΔDF	$\Delta\chi^2$	p	RMSEA	CFI
1	-	0	0	1.000	-	-	-	0.000	1.000
2	$\underline{\Lambda}$	6	14.73*	0.023	6	14.73*	0.023	0.145	0.942
2a	$\Lambda^{\underline{1}}$	5	4.74	0.449	(-) 1	9.99**	0.002	0.000	1.000
3	$\Lambda^1, \underline{\Theta}$	14	23.68	0.050	9	18.94*	0.026	0.097	0.936
3a	Λ^1, Θ^2	13	16.45	0.226	(-) 1	7.23**	0.007	0.058	0.977
4	$\Lambda^1, \Theta^2, \underline{\Psi con}$	15	16.91	0.324	2	0.46	0.795	0.040	0.987
5	$\Lambda^1, \Theta^2, \underline{\tau}^3, \Psi con$	20	31.78*	0.046	5	14.87*	0.011	0.089	0.922
5a	$\Lambda^1, \Theta^2, \tau^{\underline{3,4}}, \Psi con$	19	27.43	0.095	(-) 1	4.35*	0.037	0.079	0.944
5b	$\Lambda^1, \Theta^2, \tau^{\underline{3,4,5}}, \Psi con$	18	23.70	0.165	(-) 1	3.73	0.053	0.065	0.962
6	$\Lambda^1, \Theta^2, \tau^{3,4,5}, \Psi con, \underline{acon}$	20	24.44	0.224	2	0.74	0.691	0.056	0.971

Note: Underlined restrictions are tested by likelihood ratio test $\Delta\chi^2$; *p<0.05; **p<0.01; (-): parameter freely estimated; 1: Factor loading Numerical Ability, minority group, stereotype threat; 2: Residual variance Numerical Ability, minority group, stereotype threat; 3: Intercept Numerical Ability, minority group, stereotype threat; 4: Intercept Abstract Reasoning, majority group, control; 5: Intercept Abstract Reasoning, minority group, control

Because a one-factor model with three indicators is saturated (i.e., equal number of input statistics and parameters), the baseline model without across-group restrictions has a chi-square of zero with zero degrees of freedom. In the second step the factor loadings are restricted to be equal over the four groups. This restriction results in a significant increase in chi-square. In addition, both the RMSEA and the CFI exceed the rule-of-thumb values for good fit. The misfit in this step is almost solely due to the factor loading of the Numerical Ability subtest of the minority group in the stereotype threat condition (MI = 11). Freeing this parameter leads to a significant improvement of model fit, as can be seen in Step 2a. In the minority group, stereotype threat condition, this (unstandardized) factor loading is not significantly different from zero (λ_1 = -0.04, SE = 0.20, Z = -0.19, p> .05), whereas in the other groups this factor loading is significantly greater than zero (λ_1 = 0.92, SE = 0.22, Z = 4.19, p< .01). In Step 3, the residual variances are restricted to be invariant over the four groups. This, again, leads to a significant deterioration in model fit, as can be seen by the significant increase in chi-square, increase in RMSEA, and lowering of CFI. Not surprisingly, the misfit in this step is mainly due to the residual variance of the Numerical Ability subtest of the minority group in the stereotype threat condition (MI = 7). Freeing this parameter leads to a significant improvement in model fit (Step 3a). The residual variance of Numerical Ability is larger in the minority group, stereotype threat condition (6.33, SE = 1.06), than in the other groups (3.47, SE = 0.61). In the fourth step,

we restrict factor variances of both ethnic groups to be invariant over condition. This leads to a relative improvement in model fit. The factor variance of the minority group is slightly smaller (Ψ = 3.32, SE = 1.08) than the factor variance of the majority group (Ψ = 4.12, SE = 1.23). In the fifth step, mean structure is modeled by restricting the intercepts to be invariant over the groups, and by freeing the factor means of three groups (cf. Table 3.1). In light of the different factor loading of the Numerical Ability subtest in the minority group, stereotype threat condition, it does not make sense to restrict this particular intercept. Hence, in Step 5, this parameter is freely estimated for this particular group. Step 5 results in a significant increase in chi-square, an increase in RMSEA, and a clear drop in CFI. The restriction on intercepts is clearly rejected. The highest modification index is related to the intercept of the Abstract Reasoning test of the majority group in the control condition. Freeing this parameter results in an improvement in model fit (Step 5a). However, as judged by RMSEA (> 0.06) and CFI (< 0.95), the model fit of Step 5a is still not very good. The highest modification index (MI = 4) in this step is related to the intercept of the Abstract Reasoning subtest of the minority group in the control condition. Freeing this parameter results in an improvement in model fit in terms of RMSEA and CFI (Step 5b). Interestingly, the intercept of the Abstract reasoning subtest is higher in the majority group, control condition (τ_2 = 8.67, SE = 0.47), than in the two ethnic groups in the stereotype threat condition (τ_2 = 7.54, SE = 0.31). This is not surprising considering the mean effect of the stereotype threat manipulation on this subtest in the majority group. In the minority group, control condition, this intercept is even lower (τ_2 = 6.72, SE = 0.37). This suggests the presence of bias with respect to ethnicity in the control condition. In the sixth and final step, we investigated whether the factor means of both groups differed over experimental condition. This step is accompanied by a relative improvement in model fit. The factor mean of the majority is significantly higher than that of the minority group (α = 1.62, SE = 0.39, Z = 4.20, p< .001).

Conclusion

Although MANOVA results indicated an absence of *mean* effects of stereotype threat on test performance of the minority group, the stereotype threat manipulation clearly resulted in measurement bias with respect to the minority group. The measurement bias due to stereotype threat was related to the most difficult Numerical Ability subtest. Interestingly, because of stereotype threat, the factor loading of this subtest did not deviate significantly from zero. This change in factor loading suggests a non-uniform effect of stereotype threat. This is consistent with the third scenario discussed above (cf. Appendix B), and with the idea that stereotype threat effects are positively associated with latent ability (cf. Cullen et al., 2004). Such a scenario could occur if latent ability and domain-identification are positively associated. This differential effect may have led low ability (i.e., moderately-identified) minority students to perform slightly *better* under stereotype threat (cf. Aronson et al., 1999), perhaps because of moderate arousal levels, whereas the more able (i.e., highly-identified) minority students performed *worse* under stereotype threat. Such a differential effect is displayed graphically in Figure 3.5. This pattern could explain the absence (i.e., canceling out) of mean-effects, the increased residual variance, and the smaller factor loading in the minority group. Another explanation for this effect may lie in

individual differences in working memory capacity (WMC). Beilock and Carr (2005) recently found that students high in WMC underperformed on a difficult arithmetic task under pressure, whereas students low in WMC showed a slight increase in performance when put under high pressure.

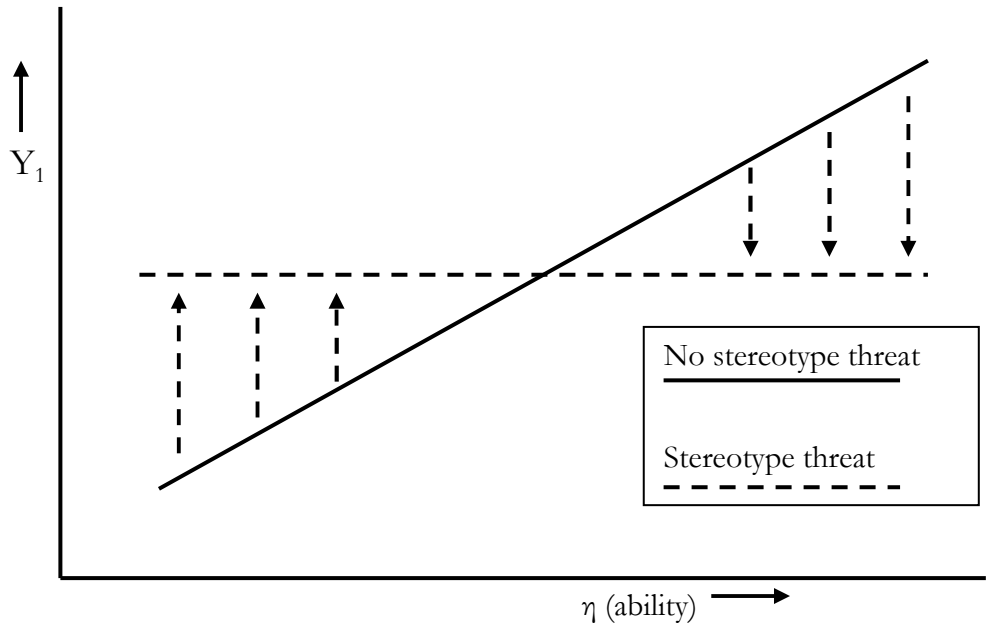

Figure 3.5 *Non-uniform effect on factor loading of Subtest Y_1 in case of an Interaction between latent ability and stereotype threat.*

The biasing effect of stereotype threat would have been completely overlooked, had we restricted ourselves to the MANOVA, and had we regarded the covariance heterogeneity as a statistical annoyance, instead of as an important source of information. The bias due to stereotype threat on test performance of the minority group is quite serious. The intelligence factor explains approximately 0.1% of the variance in the Numerical Ability subtest, as opposed to 30% in the other groups. To put it differently, due to stereotype threat, the Numerical Ability test has become completely worthless as a measure of intelligence in the minority group. Note, however, that such an effect changes our interpretation of the factor within the minority group under stereotype threat. It is also conceivable that the stereotype threat effects were present on the other two subtests. However, because of the rather small factor model, such an effect is hardly distinguishable from a non-uniform effect on the Numerical Ability test. Nevertheless, the latter subtest is the most difficult subtest, and it is apparent that stereotype threat has resulted in severe measurement bias with respect to the minority group.

In the control condition, there also appears to be measurement bias with respect to ethnicity, indicating that even in that condition test scores of minority and majority students are incomparable. It could be argued that because the test setting resembled

strongly the common practice of testing in Dutch high school, the test setting could have elicited stereotype threat even in the control condition (cf. Steele & Davies, 2003). However, because the bias in this condition was related to the easiest of the three subtests, it seems unlikely that stereotype threat has caused this bias. Further research could shed light on the issue of whether stereotype threat is also present in the control condition or if perhaps the bias is caused by something else (cf. Te Nijenhuis et al., 2000). Based on this study, we would advise great caution in the use of these DAT scales for Dutch minority students.

Surprisingly, the manipulation also had a depressing effect on the Abstract Reasoning subtest in the majority group. Perhaps this is due to a priming effect of the ethnicity questionnaire (cf. Wheeler & Petty, 2001). Further research could shed light on why the scores on this relatively easy subtest were depressed in the majority group. Nevertheless, the depressing effect of stereotype threat on this subtest became apparent in the analysis of variance, and clearly resulted in measurement bias in the factor analyses.

The presence of covariance effects in the absence of mean effects in this first study, led us to re-analyze the results of another stereotype threat study, in which a clear mean effect on test performance was also absent. In an experiment by Nguyen and colleagues (2003) the effects of stereotype threat on Black students' test performance were studied within a job-selection context. A timed short version of a cognitive ability test containing three subtests was used to assess cognitive ability. A total of 86 Blacks and 86 Whites were randomly assigned to a stereotype threat or control condition. Like in Study 1 above, stereotype threat was manipulated by both an ethnicity prime and by test diagnosticity (Nguyen et al., 2003). Using analysis of variance, Nguyen et al. (2003) found that Whites outscored the Blacks on all subtests (i.e., significant multivariate and univariate main effects for ethnicity). However, MANOVA indicated no significant interaction between stereotype threat manipulation and race, as would be expected from stereotype threat theory. Therefore, Nguyen and colleagues concluded that stereotype threat effects on test performance were absent. We submitted these data to MGCFA and our re-analysis suggested that (besides an increased residual variance for Whites in the stereotype threat condition) strict factorial invariance with respect to conditions and race was mainly tenable. Although the power may have been low, this result suggests that the race differences in test performance in either condition appear not to be caused by stereotype threat. Therefore, the argument that the stereotype threat manipulation in the Nguyen study was unsuccessful due to the fact that stereotype threat was already present in the control condition (Steele & Davies, 2003), appears implausible.

From an experimental perspective the results of the first study are unusual in the sense that experimental mean effects on test performance of the stigmatized group were absent. Hence, it is desirable to investigate the merits of our modeling approach in the presence of clear experimental mean effects.

3.7 Study 2: O'Brien and Crandall (2003) Re-Analysis

O'Brien and Crandall (2003) studied the effects of stereotype threat on performance of females on three mathematics tests, which differed in difficulty level: A

difficult test, an easy test, and a relatively easy math persistence test. Here we re-analyze these data with our modeling approach in order to investigate whether a test of strict factorial invariance can highlight the stereotype threat effects on test performance. We briefly describe the original study. For more details the reader is referred to O'Brien and Crandall (2003).

Method

Participants

A total of 164 students enrolled in a psychology class participated in this study in exchange for course credit. Because of missing data of five participants on the Math Persistence test, the current analysis is based on a sample of 58 females and 101 males.

Design and Procedure

Participants were randomly assigned to two conditions that differed in the amount of stereotype threat for women. In the control condition, the gender stereotype was made irrelevant for the test setting by a text stating that the test at hand had "NOT been shown to produce gender differences" (O'Brien & Crandall, 2003, p. 785). In the stereotype threat condition, the text indicated that the test had been shown to produce gender differences. After reading this text, participants completed a questionnaire regarding their feelings concerning test taking. After that, the three math tests were administered in a counterbalanced order.

Materials

The Easy math test had a time limit of 10 minutes, and consisted of 20 relatively easy multiplication problems. The Difficult test was administered with a time limit of 11 minutes, and consisted of 15 difficult items from the quantitative SAT. Items were in a five option multiple-choice format. The Math Persistence test contained 24 addition and subtraction problems, which were to be solved mentally (i.e., without the aid of paper and pencil) within 8 minutes (O'Brien & Crandall, 2003).

Analyses

Reasoning that the effects of heightened arousal on task performance depend on task difficulty, O'Brien and Crandall (2003) expected that stereotype threat would heighten scores of females on the Easy math test, while depressing their scores on the Difficult test. The Math Persistence test was originally used as a control for effort. However, because of quite high correlations between all three subtests, and in light of the clear mathematical nature of the three tests, the use of a one-factor model in describing these data is justified. In the male groups in both conditions and in the female group, stereotype threat condition, all inter-subtest correlations are significantly greater than zero ($p < .05$; range: 0.33 to 0.55). However, the correlation between the Easy and the Difficult test of the female group in the control condition is not significant. Furthermore, the correlation between the Easy test and the Math Persistence test in this group is negative. This appears not to be caused by any distinguishable bivariate outliers (L. T. O'Brien, personal communication, June 7, 2004). Moreover, in this group, the Math Persistence test has a platykurtotic distribution (kurtosis: - 1.3). In combination with the small sample size ($N = 30$), this makes the data of this group less suitable for Maximum Likelihood estimation. Therefore, we limited the factor analyses to three groups: the female group in the stereotype threat condition, and the Male

groups in both conditions. For our modeling approach this poses no problem. We expected measurement bias because of stereotype threat in the female group. We again use the steps given in Table 3.1 to assess the tenability of restrictions over these three groups.

Results

Except for the Math Persistence test scores of the male group in the stereotype threat condition[24], the kurtosis and skewness values are in the moderate range, making the data suitable for Maximum Likelihood estimation. The means and standard deviations of the four groups are reported in Table 3.4. Using repeated-measures ANOVA on the standardized scores of the Easy and the Difficult tests, O'Brien and Crandall (2003) found a significant main effect for gender, with males outscoring the females. More importantly, this test showed a significant three-way interaction between gender, condition, and test difficulty, which indicated that stereotype threat lowered scores of women on the Difficult test, while heightening the scores on the Easy test. In a separate two-way ANOVA on the Math Persistence scores, O'Brien and Crandall (2003) found a significant main effect for sex (males outscoring females), although the interaction between sex and condition was not significant. Thus, these ANOVA results indicate no effects of condition for males. For females, ANOVAs indicate a clear mean effect of stereotype threat on the Easy and Difficult tests, but no effect on the Math Persistence test.

Table 3.4
Means and standard deviations of males and females by experimental condition (Study 2)

	Condition							
	Control				Stereotype threat			
	Males		Females		Males		Females	
N =	50		30		51		28	
Subtest	M	SD	M	SD	M	SD	M	SD
Easy	7.50	4.34	6.37	3.91	7.80	3.93	8.18	3.98
Difficult	9.13	2.36	7.99	2.88	9.19	2.51	6.81	2.55
Persistence	18.72	5.79	15.30	6.13	19.53	4.67	16.43	6.30

Note: Descriptive statistics provided by L. T. O'Brien

The results of the factor analyses on the three groups are reported in Table 3.5. Again, the first step involves a saturated model with perfect model fit. The second step (equal factor loadings), the third step (equal residual variances), and the fourth step (equal factor variance in male groups) all result in non-significant increases in chi-square. Moreover, the CFI and RMSEA clearly indicate that these three restrictions are tenable. This is not the case for the restriction on measurement intercepts, which is tested in the fifth step. This restriction clearly results in a worsening in model fit, as is clear in the significant increase in chi-square and the clear worsening in CFI and RMSEA values. The

[24] The high kurtosis value (2.6) in this group was due to a very low scoring male. Excluding this outlier does not change the results of the factor analyses.

largest modification indices are related to the intercepts of the Difficult test (MI = 7) and the Easy test (MI = 6) of the female group in the stereotype threat condition. Freeing both parameters (Steps 5a and 5b), results in clear improvements in model fit. The intercept of the Difficult test is lower in the female group under stereotype threat (τ_2 = 7.72, SE = 0.57), than in both male groups (τ_2 = 9.05, SE = 0.30). The intercept of the Easy test is higher in the female group (τ_3 = 9.65, SE = 0.94) than in both male groups (τ_3 = 7.48, SE = 0.50). In the sixth step the factor mean of the males in both conditions is restricted to be equal. This does not result in a worsening in model fit. In this last step, the factor mean of the female group is significantly lower than that of the male group: a = 2.70, SE = 1.28, Z = 2.11, p< .05. However, because of the two freely estimated intercepts, this factor mean difference is actually a significance test of the difference between males and females on the Math Persistence test.

Table 3.5

Fit measures of steps towards strict factorial invariance (Study 2)

Step	Restrictions	DF	χ^2	p	ΔDF	$\Delta\chi^2$	p	RMSEA	CFI
1	-	0	0	1.000	-	-	-	0.000	1.000
2	$\underline{\Lambda}$	4	2.74	0.602	4	2.74	0.602	0.000	1.000
3	$\Lambda,\underline{\Theta}$	10	5.87	0.826	6	3.13	0.792	0.000	1.000
4	$\Lambda,\Theta,\underline{\Psi con}$	11	6.40	0.846	1	0.53	0.467	0.000	1.000
5	$\Lambda,\Theta,\underline{\tau},\Psi con$	15	22.78	0.089	4	16.38**	0.003	0.113	0.896
5a	$\Lambda,\Theta,\underline{\tau^1},\Psi con$	14	12.42	0.572	(-) 1	10.36**	0.001	0.000	1.000
5b	$\Lambda,\Theta,\underline{\tau^{1,2}},\Psi con$	13	6.66	0.919	(-) 1	5.76*	0.016	0.000	1.000
6	$\Lambda,\Theta,\tau^{1,2},\Psi con,\underline{\alpha con}$	14	7.02	0.934	1	0.36	0.549	0.000	1.000

Note: Underlined restrictions are tested by likelihood ratio test $\Delta\chi^2$. *p<0.05; **p<0.01; (-): Parameter freely estimated; 1: Intercept Difficult subtest in women, stereotype threat; 2: Intercept Easy subtest, women, stereotype threat

Conclusion

The re-analysis of O'Brien & Crandall's data demonstrated one drawback of the current modeling approach. Because of the platykurtotic distribution of test scores, and the negative correlation between tests in the female group, control condition, this group had to be excluded from the test for measurement invariance. Nevertheless, the factor analysis approach remained feasible. Even without the possibility to compare the female group in the stereotype threat condition to a female group without such threat effects (i.e., in the control condition), we were able to establish that test scores of males and females were incomparable. It became apparent that intercepts were not invariant across groups, and that strict factorial invariance was violated due to stereotype threat. Suppose that these data would have been non-experimental data, stemming from a real-life, or even a high-stakes, test setting. Even then, a test for strict factorial invariance would have pointed towards the measurement bias with respect to sex. The re-analysis of these data illustrates our point that because of their nature, stereotype threat effects are detectable in principle by means of tests for measurement invariance.

Of course, O'Brien and Crandall (2003) especially selected their math tests to show this pattern of effects. However, their study can contribute to future studies into stereotype threat effects within real-life test settings. A careful selection of easy and more difficult

tests, together with the current modeling approach, enables one to investigate the existence of stereotype threat effects on test performance. In sum, the results of the current re-analysis are clearly in line with the results of analysis of variance by O'Brien and Crandall (2003). Moreover, the current results support the notion that whenever stereotype threat affects test performance on a collection of tests, it does so in a way incompatible with the requirements for measurement invariance within the common factor model.

One drawback of the first two studies is the small number of subtests. In Study 3, we use a test battery consisting of four subtests measuring arithmetic/mathematic ability. In addition, we want to investigate strict factorial invariance in three conditions that differ with respect to stereotype threat related to female test takers: a control condition with no explicit reference to sex differences, a nullified condition in which gender stereotype was made irrelevant to the test, and a stereotype threat condition with explicit mention of sex differences. The latter condition is interesting because it has well-known negative effects on female test performance, while male test performance is often enhanced (i.e., a stereotype lift effect; Walton & Cohen, 2003). We expected that both this negative and this positive effect would result in measurement bias. The comparison with regard to strict factorial of three conditions that differ in stereotype threat, enables one to find a test setting where stereotype threat is absent, and where test scores of males and females are comparable.

3.8 Study 3: Sex Differences in Arithmetic Test Performance

The first aim of this third study is to replicate the effects of stereotype threat on women's test scores on a collection of arithmetic/mathematic ability tests in a sample of psychology undergraduates in the Netherlands. The second aim is to investigate whether tests for measurement invariance using MGCFA can successfully differentiate between conditions, in which stereotype threat is manipulated. To this end, we administered an arithmetic test battery to males and females, varied the amount of stereotype threat for females over conditions, and tested for strict factorial invariance with respect to groups.

Method

Participants

Two hundred and eighty-three undergraduate psychology students of the University of Amsterdam participated as part of course requirements.[25] On average, the 142 females were slightly younger (age: M = 20.40, SD = 3.76) than the 141 males (M = 21.64, SD = 4.97). The sample is highly educated, but not especially selected for good arithmetic/mathematic skills. The sample is expected to be heterogeneous with respect to identification with the arithmetic/mathematical domain.

Design and Procedure

An arithmetic test battery was administered by computer during two large mixed-sex group sessions. Participants were randomly assigned by the computer to one of three conditions, in which the introduction texts were used to manipulate the amount of stereotype threat. All three texts started by mentioning that the test of arithmetic ability

[25] Due to computer failure, three additional participants, one male and two females, were excluded from the analyses.

contained four timed subtests. The three versions differed with respect to the next section in the instruction text. In the control condition, meant to resemble the usual testing circumstances, no mention was made of sex differences. In the nullified condition, on the other hand, the instruction read (translated from Dutch): "Although on many arithmetic tests sex differences have been found, previous research has shown that on this arithmetic test, females achieve as well as males. Mean scores of males and females on the four subtests are equal." This nullified condition was created to make the gender-stereotype irrelevant for the test that participants were making, thereby hopefully reducing the effects of stereotype threat on females (cf. Brown & Pinel, 2003; O'Brien & Crandall, 2003; Smith & White, 2002; Spencer et al., 1999). In the stereotype threat condition the text was changed to (translated from Dutch): "Previous research has shown that females and males score differently on this arithmetic test. On the average females score lower than males on all four subtests." This instruction text was meant to increase stereotype threat for female test-takers in the stereotype threat condition. (cf. Keller, 2002; O'Brien & Crandall, 2003; Spencer et al., 1999). After this manipulation, the participants completed the four subtests. Each subtest consisted of a page with a specific instruction, an example item, and a test page containing the test items. The computer automatically stopped the subtests when the allocated test period had passed. Total test time was 21 minutes. After the test session, all participants were debriefed extensively on the purpose of the experiment.

Materials

We used a selection of subtests that measure arithmetic/mathematical proficiency. The four subtests differ in form and difficulty level, but are nevertheless expected to measure one single trait, which we henceforth denote by arithmetic ability. In order of presentation, these subtests are: Arithmetic, Number Series, Worded Problems, and Sums.

The *Arithmetic* test is a timed test of three minutes containing 40 items that stem from an arithmetic ability test by Elshout (1976). The latter test is part of the standard test program of psychology undergraduates at the University of Amsterdam. The original test has high internal consistency and validity (Vorst & Zand Scholten, 2000). The items have an open-ended answer format, for example: "43 x 6 =".

The *Number Series* test is a test developed to be parallel to the Number Series Test by Elshout (1976). The latter test is also part of the standard test program of the Psychology Department, and has high internal consistency and validity (Elshout, 1976; Vorst & Zand Scholten, 2000). The test used in the current study contains 20 items in a five-option multiple-choice format and has a time limit of six minutes. Example item: "-12 –11 –8 –3 4 ... (options: 7 12 13 15 9)".

The *Worded Problems* test has a time limit of four minutes, and contains 23 worded arithmetic problems. This test is based on the Arithmetic subtest of the WAIS-Dutch edition (Stinissen, Willems, Coetsier, & Hulsman, 1970), and contains some additional and comparable items from the CMS test by Elshout (1976). All items have an open-ended answer format, and were adapted to increase difficulty. Example item: "Someone has a loan at a 5% interest rate per year. After three years he has paid 225 Euros interest. What is his debt in Euros?".

The *Sums* test is the numerical ability test of the Primary Mental Abilities (T. G. Thurstone, 1958, 1962). It contains 60 items and was administered with a (adapted) time

limit of 5 minutes. The respondents are required to indicate whether a sum is correct or incorrect. E.g., "13 + 39 + 99 + 32 = 183". To correct for guessing on this subtest, the total score is computed by subtracting half the number of incorrect responses from the number of correct responses.

Although speediness increases the difficulty of all subtests, the items themselves are fairly easy to solve. The Number Series subtest is the most difficult in terms of abstractness and item difficulty. We therefore expected that stereotype threat would particularly affect scores on this subtest.

Analyses

Again, we also provide to results of a two-way MANOVA with sex and condition (3 levels) as factors and the four tests as dependent variables. Based on research in previous cohorts of psychology undergraduates (e.g., Vorst & Zand Scholten, 2000), we anticipated that males would outscore the females on all subtests. We expected that the instruction texts would particularly influence female test performance. Specifically, we expected that females in the nullified condition would outscore the females in the control and stereotype threat conditions. In addition, we predicted females in the stereotype threat condition to score lowest of all groups. We expected no negative effects for males, although stereotype lift effects (Walton & Cohen, 2003), could conceivably provide a pattern of mean differences for the males opposite to those of females.

As the four subtests were expected to load on a general arithmetic ability factor, we fitted a single common factor model in the confirmatory factor analyses. We again follow the stepwise approach given in Table 3.1, this time involving six groups. We expected to find measurement bias for females in the stereotype threat condition. This should result in the rejection of strict factorial invariance, particularly due to the induced bias in the relatively difficult Number Series subtest. Whether strict factorial invariance with respect to sex is tenable in the control and nullified conditions depends on the degree of stereotype threat. However, we expected the degree of measurement bias to be greatest in the stereotype threat condition.

Results

With two exceptions (i.e., Arithmetic subtest for males in control and stereotype threat conditions), univariate skewness and kurtosis values are moderate ([-1,1]), suggesting univariate normality of most subtests in most of the cells. Therefore, use of Maximum Likelihood in estimating the factor models seems appropriate. Means and standard deviations of the subtests for males and females in the three conditions are given in Table 3.6. The Box test shows that homogeneity of covariance matrices across conditions is rejected ($F(50, 139810) = 1.748$, $p < .01$). Levene's tests for equal variances across conditions show significant values for Arithmetic ($F(5, 277) = 4.683$, $p < .001$) and Number Series ($F(5, 277) = 4.619$, $p < .001$), but non-significant values for the other two subtests. Assuming robustness to this violation of (co)variance homogeneity, we continue with the MANOVA. The multivariate sex main effect is associated with a significant F-value ($F(4, 274) = 7.351$, $p < .001$). The univariate analyses of variance show significant sex main effects on all subtests (Arithmetic: $F(1, 277) = 12.89$, $p < .001$; Number Series: $F(1, 277) = 25.79$, $p < .001$; Worded Problems: $F(1, 277) = 19.58$, $p < .001$; Sums: $F(1, 277) = 5.43$, $p <$

.05), with males outscoring the females on all subtests. Furthermore, compared to the nullified and control conditions, there is a clear trend for females in the stereotype threat condition to score lower. For the males, the picture is less clear, with highest scores in conditions depending on the subtest used. The multivariate main effect of condition does not reach significance ($F(8, 548) = 1.708$, $p > 0.05$). Most importantly, the multivariate interaction of condition and sex is significant: $F(8, 548) = 2.366$, $p < 0.05$. None of the univariate condition main effects reach significance (All Ps > .10). As expected, the only significant univariate interaction effect between sex and condition is found on the Number Series subtest: $F(2, 277) = 4.32$, $p < .05$. Within the female group, the simple effect for condition is significant ($F(2, 139) = 7.29$, $p < .01$). Paired comparisons show that females in the stereotype threat condition scored significantly lower than females in the control condition ($p < .01$), and significantly lower than females in the nullified condition ($p < .05$), but that female scores did not differ significantly between nullified and control conditions ($p > .50$). Although male scores on the Number Series subtest are highest in the stereotype threat condition, the condition simple effect for males did not reach significance ($F(2, 138) = 0.48$, $p > .50$), nor did any of the paired comparisons for males (all Ps > .50). In other words, the stereotype lift effect for males did not reach significance using the traditional analysis of variance approach. To summarize, these ANOVA results indicate a clear suppression of scores on the Number Series subtest for females in the stereotype threat condition.

Table 3.6

Means and standard deviations of subtests per sex and condition (Study 3)

	Condition											
	Control				Nullified				Stereotype Threat			
	Males		Females		Males		Females		Males		Females	
N =	46		48		50		47		45		47	
Subtest	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
Arith-metic	13.28	7.46	10.23	4.62	14.18	7.78	11.70	3.53	12.20	5.53	9.96	6.16
Number Series	8.52	3.74	7.60	2.86	8.56	4.36	7.11	2.66	9.22	3.33	5.62	2.35
Worded Probl.	8.39	3.43	6.40	2.80	7.60	3.09	6.72	2.32	7.44	2.88	5.74	2.72
Sums	12.90	5.92	11.55	5.14	13.14	5.86	11.21	4.66	12.97	5.11	11.81	5.18

Results of factor analyses in the six groups are reported in Table 3.7. In the first step we assessed the fit of the one-factor model, which is acceptable. The second step does not result in a significant increase in chi-square. Therefore, factor loadings appear invariant over the six groups. The restriction on residual variances in the third step results in a clear deterioration in model fit. The largest modification indices are found in the male group, nullified condition, and are related to the residual variance of the Number Series subtest (MI = 23) and of the Arithmetic subtest (MI = 18). Furthermore, the residual variance of the Arithmetic test in the females in the stereotype threat condition is also partly responsible for misfit (MI = 13). Freeing these three parameters in a stepwise fashion (Steps 3a, 3b, 3c) results in clear improvements in model fit. These freely estimated residual

variances are larger in the corresponding groups than in the other groups. In the fourth step, the factor variances of the male group and of the female group are restricted to be equal over conditions. This results in a slight, but non-significant, increase in chi-square. Considering the perfect values of RMSEA and CFI in Step 4, we conclude that factor variances of the sex groups are invariant over conditions. The factor variance of the female group is smaller (ψ = 15.08, SE = 2.47) than the factor variance of the male group (ψ = 38.09, SE = 5.56).

Table 3.7
Fit measures of steps towards strict factorial invariance (Study 3)

Step	Restrictions	DF	χ^2	p	ΔDF	$\Delta\chi^2$	p	RMSEA	CFI
1	-	12	9.61	0.650	-	-	-	0.000	1.000
2	$\underline{\Lambda}$	27	18.39	0.891	15	8.78	0.889	0.000	1.000
3	$\Lambda,\underline{\Theta}$	47	64.17*	0.049	20	45.78**	0.001	0.099	0.967
3a	$\Lambda,\Theta^{\underline{1}}$	46	47.18	0.424	(-) 1	16.99**	0.000	0.031	0.998
3b	$\Lambda,\Theta^{\underline{1,2}}$	45	36.74	0.805	(-) 1	10.44**	0.001	0.000	1.000
3c	$\Lambda,\Theta^{\underline{1,2,3}}$	44	26.00	0.986	(-) 1	10.74**	0.001	0.000	1.000
4	$\Lambda,\Theta^{1,2,3}$, $\underline{\Psi con}$	48	35.39	0.912	4	9.39	0.052	0.000	1.000
5	$\Lambda,\Theta^{1,2,3}$, $\underline{\tau}$, Ψcon	63	76.73	0.115	15	41.34**	0.000	0.072	0.973
5a	$\Lambda,\Theta^{1,2,3}$, $\tau^{\underline{4}}$, Ψcon	62	63.99	0.407	(-) 1	12.74**	0.000	0.040	0.996
5b	$\Lambda,\Theta^{1,2,3}$, $\tau^{4,\underline{5}}$, Ψcon	61	55.19	0.685	(-) 1	8.80**	0.003	0.000	1.000
6	$\Lambda,\Theta^{1,2,3}$, $\tau^{4,5}$, Ψcon, $\underline{\alpha con}$	65	59.12	0.682	4	3.93	0.416	0.000	1.000

Note: Underlined restrictions are tested by likelihood ratio test $\Delta\chi^2$. * p<0.05; **p<0.01; (-): Parameter freely estimated; 1: res.var. Number Series, Males, Nullified; 2: res.var. Arithmetic, Males, Nullified; 3: res.var. Arithmetic females, stereotype threat; 4: intercept Number Series, Females, stereotype threat; 5: intercept Number Series, Males, stereotype threat

Considering the mean effects that we found by means of the MANOVA, one would expect intercept differences across groups. In the fifth step the intercepts are restricted to be invariant across groups. This clearly results in a deterioration in model fit, with a highly significant increase in chi-square, worsening in RMSEA, and drop in CFI. Inspection of the modification indices shows that this restriction is untenable because of the intercept of the Number Series subtest in the stereotype threat condition in both sex groups (females: MI = 12; males: MI = 8). Indeed, freeing both parameters results in clear improvement in model fit (i.e., Steps 5a and 5b). As expected, the intercept of this difficult subtest is lower in the female group in the stereotype threat condition (τ_2 = 5.92, SE = 0.45), than in the groups in the other conditions (τ_2 = 7.19, SE = 0.31). In the male group under stereotype threat this intercept is higher (τ_2 = 8.40, SE = 0.45), thus nicely reflecting the stereotype lift effect on this relatively difficult subtest. In the sixth step, factor means of each sex group are restricted to be equal over conditions. This restriction appears tenable. The factor mean of the male groups is significantly higher than the factor mean of the female groups (a = 2.61, SE = 0.67, Z = 3.92, p< .001). In terms of the pooled within-group standard deviation units of the latent factor, this difference in latent ability has an effect size of 0.52.

The current stepwise approach has the risk of path-dependence, in the sense that the results of later restrictions (i.e., steps in the lower part of Table 3.1) may depend on the particular parameters, which were freed in previous steps because of high modification indices. In addition, within a particular test setting one would normally test for strict factorial invariance with respect to the *existing* groups. Therefore, both as an illustration, and as a check, we also report tests for strict factorial invariance with respect to sex *within* each of the three conditions. This enables us to investigate whether these tests can differentiate between situations (i.e., conditions) in which stereotype threat is, or is not, present. Note that in this situation it does not make sense to restrict factor variances and factor means, thus Steps 4 and 6 are skipped. The results of the tests per condition are reported in Table 3.8. As can be seen, in the control condition, restricting factor loadings, residual variances, and intercepts does not result in a worsening in model fit. In this condition strict factorial invariance with respect to sex is clearly tenable. Test scores of males and females in this condition are therefore comparable, and sex differences in test performance can be explained by differences in factor mean (a = 3.16, SE = 1.28, Z = 2.47, p< .01). This sex difference in factor mean has an effect size of 0.55, which is comparable to the effect size estimate in the six-group analysis.

Table 3.8

Fit measures of stepwise test of strict factorial invariance over sex per condition (Study 3)

Step	Restrictions	DF	χ^2	p	ΔDF	$\Delta\chi^2$	p	RMSEA	CFI
				Control condition					
1	-	4	2.33	0.675			-	0.000	1.000
2	$\underline{\Lambda}$	7	4.72	0.694	3	2.39	0.495	0.000	1.000
3	$\Lambda,\underline{\Theta}$	11	6.39	0.846	4	1.67	0.796	0.000	1.000
5	$\Lambda,\Theta,\underline{\tau}$	14	10.03	0.760	3	3.64	0.303	0.000	1.000
				Nullified condition					
1	-	4	2.56	0.634				0.000	1.000
2	$\underline{\Lambda}$	7	5.04	0.655	3	2.48	0.479	0.000	1.000
3	$\Lambda,\underline{\Theta}$	11	18.69	0.067	4	13.65**	0.009	0.104	0.946
5	$\Lambda,\Theta,\underline{\tau}$	14	19.42	0.150	3	0.73	0.866	0.071	0.962
				Stereotype threat condition					
1	-	4	4.72	0.317				0.063	0.996
2	$\underline{\Lambda}$	7	7.23	0.406	3	2.51	0.473	0.000	0.999
3	$\Lambda,\underline{\Theta}$	11	17.89	0.084	4	10.66*	0.031	0.113	0.958
5	$\Lambda,\Theta,\underline{\tau}$	14	40.31**	0.000	3	22.42**	0.000	0.197	0.839

Note: Underlined restrictions are tested by likelihood ratio test $\Delta\chi^2$. *p<.05; **p<.01; Restrictions: equality constraints over sex-group

In the nullified condition, restricting the residual variances leads to a clear deterioration in fit, as is evident by the significant chi-square difference between Steps 3 and 2, increased RMSEA, and lowered CFI. With the added restriction on intercepts, model fit does not appear to worsen any further, indicating that the mean-structure is sex-invariant. The largest modification indices are related to the residual variances of the Arithmetic and the Number Series subtests.

In the condition in which the gender-stereotype was activated, we see that the baseline model (Step 1) shows sufficient fit, although RMSEA is somewhat large (i.e.,

RMSEA > .06). Here, again, the restriction on factor loadings is not accompanied by any substantial worsening in model fit. In the third step, in which residual variances are restricted to be sex-invariant, the fit does deteriorate. However, the clearest deterioration in model fit is found when mean structure is modeled (Step 5). All fit measures show that strict factorial invariance is *un*tenable in this condition. As expected, the largest modification indices are found with the intercept of the Number Series subtest and the residual variance of the Arithmetic subtest.

Conclusion

The MANOVA results indicate that stereotype threat affected the arithmetic test scores of the male and female groups in a differential manner. As expected, the clearest effect of stereotype threat was found on the difficult Number Series subtest. Females clearly underperformed on this subtest when they were reminded of the gender stereotype that females perform less well than males on arithmetic ability tests. This corroborates the typical result that stereotype threat negatively affects math performance of female test takers on difficult tests (e.g., Spencer et al., 1999).

The factor analyses showed that strict factorial invariance over sex clearly failed in the stereotype threat condition. Specifically, stereotype threat resulted in bias with respect to sex in the Number Series subtest. In the nullified condition we saw that residual variances were larger in the male group, indicating the presence of slight measurement bias with respect to males. Perhaps this is because the instruction text had a sort of stereotype threat effect on these males. Therefore, the instruction text (falsely) stressing the absence of sex differences appears not to create ideal test circumstances for males. In the control condition, strict factorial invariance with respect to sex was tenable. Thus, in that condition, test scores of males and females are comparable, and sex differences in test scores can be interpreted in terms of differences in the latent construct.

In contrast with several studies conducted in the US (Ben Zeev, Fein, & Inzlicht, 2005; Smith & White, 2002; Spencer et al., 1999), we did not find a significant mean difference on female math performance between control and nullified conditions. This may be due to a difference in test setting. In the majority of American studies participants were tested alone as opposed to in large mixed-sex groups. Such differences in setting are known to affect the strength of stereotype threat (Inzlicht & Ben Zeev, 2003; Sekaquaptewa & Thompson, 2003). Alternatively, gender stereotypes may be less strong in the Netherlands.

When test takers were reminded of gender stereotypes concerning math ability, this resulted in stereotype threat negatively affecting female performance and in stereotype lift positively affecting male performance. Interestingly, this stereotype lift effect did not reach significance in the MANOVA analysis, but was clearly detected using MGCFA. In sum, the results of the MGCFA analyses clearly indicate that tests for strict factorial invariance are capable of determining whether or not stereotype threat plays a role in a particular test-situation.

3.9 General Discussion

There is a large and still-growing body of research that supports the notion that stereotype threat can negatively affect test performance in stigmatized groups (Steele et al., 2002). The magnitude of these negative effects is often investigated in laboratory experiments, in which stereotype threat can be manipulated. However, such research within real-life settings is difficult for ethical and logistical reasons (Sackett, 2003; Steele & Davies, 2003; Steele et al., 2002). Nevertheless, viewing and modeling stereotype threat effects as a source of measurement bias, the seriousness of stereotype threat for the comparability of groups can be investigated by testing for measurement invariance with respect to groups, regardless type of group, test setting, or test under investigation, provided, of course, that a reasonable factor structure is tenable.

Stereotype Threat as a Biasing Variable

Measurement invariance with respect to groups is an essential aspect for interpreting group differences in scores of any kind of psychological measurement. Tests for measurement invariance enable one to differentiate between group differences in the latent constructs that a certain test is supposed to measure (i.e., real ability differences), and measurement artifacts related to group membership. We view stereotype threat as a source of measurement bias. Surely, no one would suggest that stereotype threat affects real (i.e., latent) abilities, at least not in the short term. Instead, stereotype threat affects the *measurements* of ability, and this is precisely what tests of measurement invariance are designed to investigate. Formally, if measurement invariance holds, and one conditions on latent ability, there should be, by definition, no group differences in (manifest) test scores. This is clearly *not* the case if stereotype threat lowers scores of members of a group that is subject to negative ability stereotypes. Therefore, measurement invariance is expected to be violated if stereotype threat differentially affects test scores of groups. Note that the same applies to stereotype lift effects (Walton & Cohen, 2003) and priming effects on test scores (e.g., Wheeler & Petty, 2001). For instance, in Study 3 we saw that the stereotype lift effect of males on the difficult subtest resulted in a heightening in the measurement intercept of this subtest. Moreover, the enhanced performance of females on the Easy test due to stereotype threat in Study 2 was also clearly detected.

Recent studies into the mediating variables of stereotype threat effects have shown that stereotype threat negatively affects working memory capacity (Schmader & Johns, 2003) or increases disruptive mental load (Croizet et al., 2004). This research suggests that the mediatory principle underlying stereotype threat effects has a strong relation to the construct of intelligence. If indeed stereotype threat affects test performance through the construct, this could result in stereotype threat effects that are completely collinear with the subtests' factor loadings. In that case, the relative strength of stereotype threat effects on each subtest correlates perfectly with the relation of each subtest with the construct. If this occurs, stereotype threat effects could conceivably be accompanied by measurement invariance with respect to groups. However, constructs such as intelligence and mathematic ability are stable characteristics, and stereotype threat effects are presumably short-lived effects, depending on factors such as test difficulty (e.g., O'Brien & Crandall, 2003; Spencer

et al., 1999). Furthermore, stereotype threat effects are often highly task-specific. For instance, Seibt and Förster (2004) found that stereotype threat leads to a more cautious and less risky test-taking style (i.e., prevention focus) the effects of which depend on whether a particular task is speeded or not, or whether a task demands creative or analytical thinking (cf. Quinn & Spencer, 2001). In light of such task-specificity, we view stereotype threat effects as test artifacts, resulting in measurement bias. Steele appears to subscribe to this view when he states that *"stereotype threat effects may be a possible source of bias in standardized tests*" (Steele, 1997, p. 622). It is an empirical question whether stereotype threat effects could ever be accompanied by measurement invariance. However, the results of the studies reported here lend support to the conceptualization of stereotype threat effects as a source of measurement bias.

It should be noted that within our empirical examples sample sizes are rather small. The power to find subtle group differences in model parameters may therefore be low. Nevertheless, the fact that bias was clearly detected in our studies indicates that MGCFA is a powerful tool in detecting measurement bias (cf. Cheung & Rensvold, 2002; Meade & Lautenschlager, 2004), even if these effects are only present at the covariance level (Study 1). In light of the fact that measurement invariance is basically a null hypothesis (Borsboom, 2006b), the failure to reject measurement invariance may always be due to a lack of power. Fortunately, power studies within MGCFA can be conducted readily (Saris & Satorra, 1993).

Using MGCFA in Experiments

Our results show that multi-group confirmatory factor analysis provides a fruitful means to investigate stereotype threat effects. It is unfortunate that many investigators do not go beyond mean differences as tested by analysis of variance or ANOVA in analyzing experimental data. Variance and covariance differences are a potential source of information. For instance, the absence of an increase in residual variance of the affected subtests in Study 2, suggests that the stereotype threat effect did not vary over women (see Appendix B, scenario 1). The effect of stereotype threat on the factor loading in the minority group in Study 1, suggests that the stereotype threat effects interacted with latent ability (see Appendix B, scenario 3). Moreover, MGCFA allows for more specific tests of experimental effects thereby increasing power. For example, the stereotype lift effect for males in Study 3 did not reach significance in the MANOVA framework, yet with MGCFA the corresponding intercept differed significantly from those in the other groups. If possible, the use of a measurement model such as multi-group confirmatory factor analysis should be preferred to analysis of variance. Moreover, the use of measurement models can add to our understanding of stereotype threat effects.

Many recent stereotype threat studies are aimed at identifying the mediating factor underlying its effects on test performance (see, e.g., Smith, 2004 for an overview). The current modeling framework may greatly contribute to this exercise, because mediators such as anxiety (e.g., Ben Zeev et al., 2005), working memory capacity (Schmader & Johns, 2003), and regulatory focus (Seibt & Förster, 2004) can be measured. Such measured mediators as well as many conceivable moderators (e.g., domain identification; Smith & White, 2001) may be incorporated in the model in a way that may eventually capture the

"stereotype threat factor" as displayed in Figures 2 through 4. Lubke and colleagues (2003a) discuss the incorporation of covariates in the MGCFA framework. When studying mediators, this method boils down to extending the factor model by adding factors, which are believed to be responsible for the depressing effect of stereotype threat. For instance, one may measure arousal (e.g., Ben Zeev et al., 2005), add to the factor model an arousal factor (besides the ability factor), and see whether this arousal factor shows an increase in factor mean (or variance) under stereotype threat. Then, in a model that takes into account latent ability, one can test whether the stereotype threat effect on test performance is mediated by arousal. Moreover, one can compare various alternative models statistically, such as whether arousal also affects the ability factor, whether arousal fully mediates the effect, whether arousal interacts with ability, etc. In comparison to traditional approaches of studying mediation (e.g., Baron & Kenny, 1986), the advantage of using MGCFA lies in the fact that MGCFA allows for a differentiation between effects on measurements of ability and effects on ability itself. This distinction is of substantive interest and may have consequences for statistical power, which is often an issue in mediation analysis (cf. MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002). The flexibility of the common factor model and structural equation modeling in general to incorporate many factors, mediators, and moderators in a linear or nonlinear fashion, opens many doors that can contribute to our understanding of stereotype threat.

Understanding Measurement Bias

Of course, measurement bias may have many causes besides stereotype threat. It is important to stress that the broad definition of measurement invariance does not suppose *anything* about the possible causes of measurement bias. Unfortunately, measurement bias has been, and still is, mostly interpreted incorrectly in terms of item content. For instance, a test item could contain a concept (e.g., a football term such as "40-yard line") that is less known to one group (say, women), resulting in increased difficulty of that item for that particular group. However, measurement bias is not a fixed characteristic of a certain test or test item, but a characteristic of how test *scores* relate to the construct that a test is supposed to measure. Although item content may be used to *interpret the causes* of measurement bias, the latter may be due to characteristics of test settings. Therefore, stereotype threat theory provides a better understanding of why measurement bias occurs. Unfortunately, the use of bias detection methods is rarely accompanied by theoretical expectations regarding why and how measurement bias occurs (but see Oort, 1992). Needless to say, understanding the sources of measurement bias can increase the chances of measurement bias being detected, either when bias is studied by MGCFA, or when bias is studied by item response models.

Stereotype Threat and Item Response Modeling

As we saw in our three studies, within multi-group confirmatory factor analysis, the effects of stereotype threat are particularly evident in the performance on the more difficult subtests. This differential aspect of stereotype threat is also relevant to the study of measurement invariance within the framework of item response theory, where item difficulty is modeled explicitly. The item level can be very informative in investigating stereotype threat effects, particularly when these are viewed as sources of measurement

bias. Within item response theory, several methods have been developed to investigate measurement bias, which in this respect is usually denoted by Differential Item Functioning or DIF (see Millsap & Everson, 1993). If only difficult items are subject to the interference of stereotype threat, this implies that easy items should be hardly affected (e.g., Spencer et al., 1999). This enables one to use easy items of tests for conditioning in testing for measurement bias with respect to stigmatized groups. In addition, only the complex or difficult items in a test would show bias in the presence of stereotype threat. Therefore, DIF analyses can also be used to investigate the effects of stereotype threat on test scores in real-life settings. In this respect recent results of a study into DIF with respect to sex on the SAT-m are of interest. Bielinski and Davison (1998; 2001) found that particularly difficult items are biased with respect to sex, which is consistent with the idea that stereotype threat has depressed scores of females on this test.

Generalizability

The generality of stereotype threat effects on test performance in real-life settings is an important issue. The number of studies investigating strict factorial invariance with respect to ethnic groups is rather small (but see Dolan, 2000; Dolan & Hamaker, 2001; Dolan et al., 2004). Clearly, there is a need for more research in this topic. If a certain test score gap is accompanied by measurement invariance (and power is not an issue), stereotype threat is not likely to play a differential role in those particular group differences. If, on the other hand, strict factorial invariance with respect to groups is violated, stereotype threat is one of the probable causes of measurement bias. Then, measures of mediators or moderators of stereotype threat could be used to model the sources of measurement bias (Lubke et al., 2003a).

As argued by Steele and colleagues (Steele et al., 2002), it depends on the test situation, domain-identification of a person, the content of the stereotype, and the kind of test, whether stereotype threat has an effect on test performance. We argue that its effects are detectable by means of tests for measurement invariance, regardless of test situation. Clearly, tests for measurement invariance can be useful to investigate the seriousness of stereotype threat on test performance, particularly in high-stakes test situations. We hope that by using the current modeling approach within an experimental context we can bridge the gap between differential psychology (with its interest in individual differences) and experimental psychology (with its interest in experimental effects), in order to gain a better understanding of when individual abilities are correctly reflected in test scores, and when they are not (cf. Cronbach, 1957).

3.10 Appendix A: General Formulation MGCFA Model

Let y_{ij} denote the observed p-dimensional random column vector of subject j in group (or experimental condition) i. We specify the following linear factor model for y_{ij}:

$$y_{ij} = \tau_i + \Lambda_i \eta_{ij} + \varepsilon_{ij} \quad (5)$$

where η_{ij} is a q-dimensional random vector of correlated common factor scores (q<p), and ε_{ij} is a p-dimensional vector of residuals that contain both random error and unique measurement effects (Meredith, 1993). The (p x q) matrix Λ_i contains factor loadings, and the (p x 1) matrix τ_i contains measurement intercepts. It is generally assumed that ε_{ij} is p-variate normally distributed with zero means and a diagonal covariance matrix Θ_i, i.e., residual terms are mutually uncorrelated. Furthermore, the vector η_{ij} is assumed to be q-variate normally distributed with mean α_i and a (q x q) positive definite covariance matrix Ψ_i. In addition, η_{ij} and ε_{ij} are assumed to be uncorrelated. Given these assumptions, the observed variables are normally distributed $y_{ij} \sim N_p(\mu_i, \Sigma_i)$, where,

$$\mu_i = \tau_i + \Lambda_i \alpha_i \,, \quad (6)$$

$$\Sigma_i = \Lambda_i \Psi_i \Lambda_i^t + \Theta_i \,, \quad (7)$$

where the superscript t denotes transposition. Equations 6 and 7 represent the implied mean vector and implied covariance matrix, respectively. In case of several correlated common factors, a sufficient number of elements in Λ_i should be fixed to zero to avoid rotational indeterminacy (Bollen, 1989; Jöreskog, 1971). In the same matrix Λ_i, q elements should be fixed to equal 1 to identify the variances of the common factors. Similarly, for reasons of identification, latent group differences in means instead of latent means themselves are modeled (Sörbom, 1974).

3.11 Appendix B

Here we present three scenarios where measurement bias due to stereotype threat (ST) is present. We use the one factor model presented in Equations 2 - 4 and the assumptions given above. We assume the presence of an *unmeasured* ST factor that incorporates all the mediating variables of ST. The scores on this ST factor are represented by σ. We assume that ST effects are uncorrelated with latent ability (i.e., $\mathrm{Cov}(\eta, \sigma) = 0$). For clarity, we leave out person and group indices and restrict our attention to the group that is affected by ST (i.e., stigmatized group). Our aim is to highlight the effects of ST on the measurement parameters of the manifest variables. For an extensive discussion of the implications of strict factorial invariance, see Lubke et al. (2003b).

Scenario 1. ST Effects on Subtest L (Figure 3.2)

Let Y_l denote the scores on a biased subtest L, and let Y_k denote the scores on a subtest K that is not affected by ST. In that case, the linear model for Y_k is given by:

$$Y_k = \tau_k + \lambda_{k\eta}\eta + \varepsilon_k, \quad (8)$$

where $\lambda_{k\eta}$ represents the factor loading of Y_k on the latent ability factor η. The linear model for Y_l (i.e., scores on the affected subtest) is given by:

$$Y_l = \tau_l + \lambda_{l\eta}\eta + \lambda_{l\sigma}\sigma + \varepsilon_l, \quad (9)$$

where $\lambda_{l\sigma}$ denotes the factor loading of Y_l on the ST factor. Note that $\lambda_{l\sigma}$ has a negative value by definition, indicating the debilitating effect of ST on test performance on subtest L. From this model, one can derive (see, e.g., Bollen, 1989) the following expressions for the implied variance (Var), and the expected value (E) of the affected subtest scores Y_l, as well as the implied covariance (Cov) of Y_l with the unaffected scores Y_k:

$$\mathrm{Var}(Y_k) = \lambda^2_{k\eta}\mathrm{Var}(\eta) + \mathrm{Var}(\varepsilon_k), \quad (10)$$

$$\mathrm{Var}(Y_l) = \lambda^2_{l\eta}\mathrm{Var}(\eta) + \lambda^2_{l\sigma}\mathrm{Var}(\sigma) + \mathrm{Var}(\varepsilon_l), \quad (11)$$

$$\mathrm{Cov}(Y_k, Y_l) = \lambda_{k\eta}\lambda_{l\eta}\mathrm{Var}(\eta), \quad (12)$$

$$\mathrm{E}(Y_k) = \tau_k + \lambda_{k\eta}\mathrm{E}(\eta), \quad (13)$$

$$\mathrm{E}(Y_l) = \tau_l + \lambda_{l\eta}\mathrm{E}(\eta) + \lambda_{l\sigma}\mathrm{E}(\sigma), \quad (14)$$

where $\mathrm{E}(\sigma)$ is greater than 0. Because the effects of ST (i.e., σ) are unknown and not modeled, the effects of the ST factor on Y_l are incorporated in the measurement parameters of this subtest on the latent factor (η). This leads to measurement bias in the corresponding parameters. The residual variance of the affected subtest is larger in the stigmatized group due to the added variance of ST: $\mathrm{Var}(\varepsilon_l)^* = \lambda^2_{l\sigma}\mathrm{Var}(\sigma) + \mathrm{Var}(\varepsilon_l)$. In addition, the intercept (τ_l) in the stigmatized group would be lower due to the ST effects: $\tau_l^* = \tau_l + \lambda_{l\sigma}\mathrm{E}(\sigma)$, reflecting increased difficulty and lowered scores of the affected subtest. Note that, since the covariance between the scores on the affected subtest and the scores on any *un*affected subtest (such as Y_k) is unrelated to σ, the factor loading of the biased subtest L (i.e., $\lambda_{l\eta}$) remains unchanged. In homogeneous samples, ST effects may not vary over persons (i.e., $\mathrm{Var}(\sigma) = 0$). This would result in the absence of added variance, while intercept bias is still present. Furthermore, it is conceivable that the mean of the ST effect is zero (i.e., $\mathrm{E}(\sigma) = 0$), resulting in the absence of intercept bias. Finally, if the mean the ST factor is negative (i.e., $\mathrm{E}(\sigma) < 0$), σ may be viewed as a stereotype lift effect (Walton & Cohen, 2003).

Scenario 2. ST Effects on Subtests L and M (Figure 3.3)

Suppose that subtests L *and* M are affected by ST. Let Y_l and Y_m denote the scores on these two affected subtests. Suppose again that scores Y_k on subtest K are unaffected by ST. The linear model for Y_k is given by (8), whereas those for Y_l and Y_m are:

$$Y_l = \tau_l + \lambda_{l\eta}\eta + \lambda_{l\sigma}\sigma + \varepsilon_l, \quad (15)$$

$$Y_m = \tau_m + \lambda_{m\eta}\eta + \lambda_{m\sigma}\sigma + \varepsilon_m. \quad (16)$$

The expected value and implied variance of Y_k are given by (10) and (13), respectively. We derive the following expressions for implied variances, implied covariances, and expected values of Y_l and Y_m:

$$\mathrm{Var}(Y_m) = \lambda^2_{m\eta}\mathrm{Var}(\eta) + \lambda^2_{m\sigma}\mathrm{Var}(\sigma) + \mathrm{Var}(\varepsilon_m), \quad (17)$$

$$\mathrm{Cov}(Y_k, Y_m) = \lambda_{k\eta}\lambda_{m\eta}\mathrm{Var}(\eta), \quad (18)$$

$$\mathrm{Cov}(Y_l, Y_m) = \lambda_{l\eta}\lambda_{m\eta}\mathrm{Var}(\eta) + \lambda_{k\sigma}\lambda_{l\sigma}\mathrm{Var}(\sigma), \quad (19)$$

$$\mathrm{E}(Y_l) = \tau_l + \lambda_{l\eta}\mathrm{E}(\eta) + \lambda_{l\sigma}\mathrm{E}(\sigma), \quad (20)$$

$$\mathrm{E}(Y_m) = \tau_m + \lambda_{m\eta}\mathrm{E}(\eta) + \lambda_{m\sigma}\mathrm{E}(\sigma), \quad (21)$$

Var(Y_l) is given by (10), and Cov(Y_k,Y_l) is given by (12). The effects on residual variances and intercepts for both the affected subtests are parallel to the effects in the first scenario. Thus, the residual variances of L and M are increased, and the intercepts of L and M are lowered due to ST. In addition, the covariance between Y_l and Y_m is now increased by the effect due to the ST factor: $\lambda_{l\sigma}\lambda_{m\sigma}Var(\sigma)$. This added covariance shows up as a subdiagonal element in the residual covariance matrix. Specifically, this results in an additional covariance between the residuals of subtest L and M: Cov$(\varepsilon_l,\varepsilon_m) = \lambda_{k\sigma}\lambda_{l\sigma}Var(\sigma)$. However, if the effects of ST do not vary over persons (i.e., Var(σ) = 0), the bias due to stereotype threat is only apparent in between-group differences of the intercepts of the affected subtests L and M, and the residual variances and residual covariance are unbiased.

Scenario 3. Non-Uniform ST Effects on Subtest L (Figure 3.4)

Non-uniform effects of ST can occur if ST effects depend on the level of latent ability. This may occur, for instance, if domain identification and latent ability are positively correlated, with higher ability reflecting stronger identification with the domain and hence stronger ST effects. Suppose subtest L is non-uniformly affected by ST, and subtest K is again unaffected by ST. Let Y_k and Y_l represent the scores on subtests K and L. The usual linear model for subtest K is given by (8). Non-uniform ST effects on Y_l can be modeled by adding an interaction factor $\eta\sigma$, resulting in this non-linear expression for the affected subtest:

$$Y_l = \tau_l + \lambda_{l\eta}\eta + \lambda_{l\sigma}\sigma + \lambda_{l\sigma\eta}\eta\sigma + \varepsilon_l\,, \tag{22}$$

where $\lambda_{l\sigma\eta}$ represents the negative factor loading of the interaction term on Y_l. This model gives rise to the following expressions for Y_l:

$$\text{Var}(Y_l) = \lambda^2_{l\eta}\text{Var}(\eta) + \lambda^2_{l\sigma}\text{Var}(\sigma) + \lambda^2_{l\eta\sigma}\text{Var}(\eta\sigma) + 2\lambda_{l\eta}\lambda_{l\eta\sigma}\text{Cov}(\eta, \eta\sigma) + 2\lambda_{l\sigma}\lambda_{l\eta\sigma}\text{Cov}(\sigma, \eta\sigma) + \text{Var}(\varepsilon_l), \tag{23}$$

$$\text{Cov}(Y_k,Y_l) = \lambda_{k\eta}\lambda_{l\eta}\text{Var}(\eta) + \lambda_{k\eta}\lambda_{l\eta\sigma}\text{Cov}(\eta, \eta\sigma) \tag{24}$$

$$\text{E}(Y_l) = \tau_l + \lambda_{l\eta}\text{E}(\eta) + \lambda_{l\sigma}\text{E}(\sigma) + \lambda_{l\eta\sigma}\text{E}(\eta\sigma). \tag{25}$$

As can be seen, this scenario leads to an increased residual variance:

$$\text{Var}(\varepsilon_l)^* = \text{Var}(\varepsilon_l) + \lambda^2_{l\sigma}\text{Var}(\sigma) + \lambda^2_{l\eta\sigma}\text{Var}(\eta\sigma) + 2\lambda_{l\eta}\lambda_{l\eta\sigma}\text{Cov}(\eta, \eta\sigma) + 2\lambda_{l\sigma}\lambda_{l\eta\sigma}\text{Cov}(\sigma, \eta\sigma)), \tag{26}$$

where $2\lambda_{l\eta}\lambda_{l\eta\sigma}Cov(\eta, \eta\sigma)$ is negative, while the other terms increase the variance. Furthermore, the ST effect depresses the intercept of the affected subtest: $\tau_l^* = \tau_l + \lambda_{l\sigma}E(\sigma) + \lambda_{l\eta\sigma}E(\eta\sigma)$. What most clearly characterizes the interaction effect, however, is the fact that the value of the factor loading of subtest L is lowered due to the non-uniform effect. This effect is due to the fact that the covariance of Y_l with all other unaffected subtests, such as Y_k, is lowered by the negative term $\lambda_{k\eta}\lambda_{l\eta\sigma}Cov(\eta, \eta\sigma)$ (provided that the mean of η is different from zero). If the mean of the biasing factor E(σ) is zero, this can account for the absence of mean effects (i.e., $\lambda_{l\sigma}$E$(\sigma) = \lambda_{l\eta\sigma}E(\eta\sigma) = 0$), and for the fact that the direction of the effect changes for low and high ability persons (cf. Figure 3.5). Finally, whereas the factors η and σ can have a normal distribution, the nonlinear effects lead to non-normal distribution of Y_l. Therefore, besides the fact that kurtosis and skewness values can point towards such nonlinear effects, such non-normality leads the normal-theory Maximum Likelihood estimator to show an upward bias in terms of model fit.

3.12 Appendix C: Stereotype Threat Research and the Assumptions Underlying Analysis of Covariance

In this appendix, we argue that the use of analysis of covariance (ANCOVA) in stereotype threat (ST) experiments is problematic, because ST theory implies violations of the assumptions underlying ANCOVA. Such violations could result in incorrect type-one error rates, and distortions in the adjustment of means.

Besides the usual analysis of variance assumptions, the assumptions underlying ANCOVA are as follows (e.g., Wildt & Ahtola, 1978). First, the relationship of the dependent variable and the covariate is linear. Second, the regression weights of the dependent variable on the covariate are equal for all design cells (i.e., regression weight homogeneity). Third, the variance of residuals is equal over cells (i.e., homogeneity of residual variance). Fourth, the covariate is measured without error and is independent of the experimental manipulation. For theoretical reasons, the tenability of these assumptions within ST experiments is at least questionable.

In a typical ST experiment (e.g., Steele & Aronson, 1995, study 2) the effects of a ST manipulation (e.g., non-diagnostic vs. diagnostic condition) on the test scores (i.e., dependent variable) of two groups (e.g., Blacks and Whites) are investigated. If a covariate (e.g., SAT scores) is used to adjust the dependent variable for pre-existing group differences, an (2x2) ANCOVA appears suitable. However, the tenability of assumptions underlying this analysis appears unlikely, especially when one compares the ST cell (i.e., stigmatized group, diagnostic condition) with the other cells in the design.

Stereotype threat theory states that ST effects particularly influence test scores of people for whom the ability of interest is important or self-relevant (Steele, 1997). It is likely that within each cell there are individual differences in domain-identification. Therefore, the manipulation triggering ST would result not only in mean effects (i.e., ST effects identical for each subject), but also in (co)variance effects (i.e., ST effects differing for subjects) on the dependent variable. Furthermore, if we suppose the presence of a positive correlation between (latent) ability and domain-identification (see Steele, 1997, p. 617), this would result in an interaction between the covariate (i.e., ability as measured by the SAT) and the experimental manipulation (i.e., ST effects on dependent variable). Higher SAT scores would imply higher domain-identification, and therefore stronger ST effects. This would result not only in a curvilinear relation between covariate and dependent variable in the affected cell (i.e., ST condition), but also in differences in regression weights over the cells. Admittedly, most ST research has used homogeneous samples, but even if individual differences in domain-identification within cells are absent, there are other reasons to expect a violation of homogeneity of regression weights.

If mediators such as heightened anxiety or lowered motivation are the causes of lowered test scores within ST conditions, it is likely that these mediators will also affect the regression of the dependent variable on the covariate. Again, such mediators can result in a violation of homogeneity of regression weights. In addition, added mediator variance (e.g., anxiety variance) could result in differences in error variances between design cells, which would violate the homogeneity of variance assumption.

Finally, the assumption that the covariate is error-free seems to be untenable because such measures are not perfectly reliable. Error in the covariate lowers the precision of the analysis. More importantly, the covariates themselves (e.g., SAT) are possibly affected by ST. It could be argued that for the high-ability participants used in most ST studies, the SAT is fairly easy and hence would not be stereotype threatening (Spencer et al., 1999). However, there are several reasons (e.g., SAT is by definition self-relevant and diagnostic of ability) to expect that the SAT scores are affected by ST. Either way, from a theoretical point of view, use of a covariate that may already be affected by the phenomenon under investigation is potentially tautological. Technically, if the covariate is affected by ST then this implies that the covariate and the manipulation are not independent, which may obscure the effects of the manipulation or even produce effects that are spurious (Wildt & Ahtola, 1978, p. 90).

In conclusion, ST theory explicitly predicts violations of practically all assumptions underlying ANCOVA. Therefore, ANCOVA appears to be unsuitable for investigating ST effects in quasi-experimental settings. In light of ST theory's emphasis on individual differences, it seems unlikely that ST only affects the means of the dependent variable (i.e., effects are identical for each subject within a cell) and leaves the covariance structure unaffected. Therefore, measurement models in which such effects are explicitly modeled (e.g., MGCFA) appear more suitable in analyzing ST effects.

4

Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn Effect

The gains of scores on standardized intelligence tests (i.e., Flynn Effect) have been subject of extensive debate concerning their nature, causes, and implications. The aim of the present chapter is to investigate whether five intelligence tests are measurement invariant with respect to cohort. Measurement invariance implies that gains over the years can be attributed to increases in the latent variables that the tests purport to measure. The studies reported contain original data of Dutch WAIS gains from 1967 to 1999, Dutch DAT gains from 1984 to 1995, gains on a Dutch children intelligence test (RAKIT) from 1982 to 1993 and re-analyses of results from Must et al. (2003) and from Teasdale and Owen (2000). The results of multi-group confirmatory factor analyses clearly indicate that measurement invariance with respect to cohorts is untenable. Uniform measurement bias is observed in some, but not all subtests. The implications of these findings are discussed.

4.1 Introduction

Ever since Flynn (1984; 1987) documented worldwide increases in scores on standardized intelligence tests, there has been extensive debate about the nature, the causes, and the implications of these increases (e.g., Neisser, 1998). There are several unresolved issues concerning the *nature* of these increases, now commonly denoted the Flynn Effect. One issue concerns the exact cognitive abilities that have increased over the years. The rise of scores is usually found to be greater on tests of fluid intelligence (e.g., Raven Progressive Matrices) than on tests of crystallized intelligence, especially on verbal IQ tests (Colom, Andres-Pueyo, & Juan-Espinosa, 1998; Emanuelsson, Reuterberg, & Svensson, 1993; Emanuelsson & Svensson, 1990; Flynn, 1987, 1998b; Lynn & Hampson, 1986, 1989; Teasdale & Owen, 2000). Differential increases have raised the question whether the gains can be related to an increase in general intelligence, or *g* (Colom & García-López, 2003; Colom, Juan Espinosa, & Garcia, 2001; Flynn, 1999a, 1999b, 2000a; Jensen, 1998; Must, Must, & Raudik, 2003; Rushton, 1999, 2000a).

A second, more fundamental, issue is whether the increases are genuine increases in cognitive ability, or that they merely reflect measurement artifacts, such as heightened test sophistication or altered test taking strategies (Brand, 1987; 1990; Brand, Freshwater, &

Dockrell, 1989; Flynn, 1990; Jensen, 1996; Rodgers, 1998). The proponents of the view that the intelligence gains are genuine have searched for real-world signs of the increase (e.g., Howard, 1999, 2001). They have offered several explanations, including improved nutrition (Lynn, 1989, 1990; Martorell, 1998), a trend towards smaller families (Zajonc & Mullally, 1997), better education (Ceci, 1991; Husén & Tuijnman, 1991; Teasdale & Owen, 1989; Tuddenham, 1948), greater environmental complexity (Schooler, 1998), and heterosis (Mingroni, 2004).

If, on the other hand, the increases are due to a measurement artifact, this obviously complicates the comparison of cohorts with respect to intelligence test scores. In addition, this may possibly have implications for the comparisons of other groups (e.g., Blacks and Whites in the US). Based on his results, Flynn (1987) questioned the validity of IQ tests, and suggested that other between-group differences on IQ tests may not reflect true intelligence differences (p.189). Furthermore, Flynn states that: "Massive IQ gains add viability to an environmental hypothesis about the IQ gap between Black and White Americans*"* (1998a, p. 40). High heritability estimates of IQ are supposedly incompatible with the hypothesized environmental causes of the secular increases (but see Mingroni, 2004). Dickens and Flynn (2001) have recently proposed a formal model that can account for this paradox. This extensive model offers an explanation of the Flynn Effect in the presence of high heritability. However, the model does not address the issue of the *nature* of the score gain since it is primarily concerned with measured intelligence or IQ.

The purpose of the present chapter is to consider the nature of the Flynn Effect. Our specific aim is to investigate whether secular gains found on five different multivariate intelligence tests reflect gains in the common factors, or hypothetical constructs, that these test are supposed to measure. These common factors are typically identified by means of factor analyses of test scores obtained within a group (cohort). To this end, we investigate whether these tests are factorially invariant with respect to cohort. Factorial invariance implies that the same constructs are measured in different cohorts, and that the observed gains in scores can be accounted for by gains on these latent constructs (Lubke et al., 2003a; Meredith, 1993). In addition, factorial invariance implies measurement invariance with respect to cohort (Meredith, 1993), which in turn means the intelligence test is unbiased with respect to cohort (Mellenbergh, 1989). We use Multi-Group Confirmatory Factor Analysis (MGCFA) to investigate factorial invariance between cohorts. An explicit technical discussion of this approach may be found in Meredith (1993). Discussions in more conceptual and applied terms are provided by Lubke, et al. (2003a)(2003a) and Little (1997). MGCFA addresses within-group differences (i.e., the covariances between cognitive subtests within a cohort) and between-group differences (i.e., the mean difference between cohorts on these tests) simultaneously. If factorial invariance is tenable, this supports the notion that (within-group) individual and (between-group) cohort differences are differences on the same underlying constructs (Lubke et al., 2003a). Conversely, if factorial invariance is untenable, the between-group differences cannot be interpreted in terms of differences in the latent factors supposed to underlie the scores within a group or cohort. This implies that the intelligence test does not measure the same constructs in the two cohorts, or stated otherwise, that the test is biased with respect to cohort. If factorial invariance is not tenable, this does not necessarily mean that all the constituent IQ subtests

are biased. MGCFA provides detailed results concerning the individual subtests, and allows one to consider partial factorial invariance (Byrne, Shavelson, & Muthen, 1989). Measurement bias between cohorts could be due to a variety of factors, which require further research to identify (Lubke et al., 2003a).

Several studies have addressed the issue whether differential gains on intelligence subtests are positively correlated with the *g* loadings of these subtests (Colom et al., 2001; Flynn, 1999a; Jensen, 1998; Must et al., 2003; Rushton, 1999, 2000a). This issue concerns the question whether between-cohort differences are attributable to the hypothetical construct *g*. As such these studies address the same question as we do here. However, we do not limit ourselves to *g*, and we employ MGCFA, rather than the method of correlated vectors (i.e., correlating differences in means on a subtest and the subtest's loading on common factor interpreted as *g*). Using the method of correlated vectors, Jensen (1998, pp. 320-321), Rushton (1999), and Must et al. (2003) found low or negative correlations, and conclude that the Flynn Effect is not due to increases in *g*. However, Flynn (1999a; 1999b; 1999c; 2000a), in a critique of Rushton's conclusions concerning Black-White differences, obtained contradictory results. In addition, Colom, et al. (2001) report high positive correlations using standardization data of the Spanish DAT. Thus, it remains unclear whether the Flynn Effect is due to increases in *g*. It may be argued that the contradictory findings are the result of differences in the tests' emphases on crystallized or fluid intelligence (Colom & García-López, 2003; Colom et al., 2001). However, of more immediate concern is the method of correlated vectors. This method has been criticized extensively by Dolan (2000) and Dolan and Hamaker (2001). One problem is that the correlation, which forms the crux of this method (i.e., the correlation between the differences in means and the loadings on what is interpreted as the g factor), may assume quite large values, even when *g* is not the major source of between group differences (Dolan & Lubke, 2001; Lubke et al., 2001). Indeed this correlation may assume values which are interpreted in support of the importance of *g*, while in fact MGCFA indicates that factorial invariance is not tenable (Dolan et al., 2004). MGCFA may be viewed as a comprehensive model based approach, which includes explicit testing of the various aspects of factorial invariance, and which includes, but is not limited to, the hypothesis that g is the dominant source of group differences. Note that in the investigation of black-white differences in intelligence test scores, this hypothesis (i.e., the importance of *g*) is referred to as "Spearman's hypothesis". The emphasis of the present analyses is on establishing factorial invariance in common factor models. Due to the nature of the available data sets, our focus is on first order common factor models, rather than on the (first or second order) *g* model.

4.2 Testing Factorial Invariance with MGCFA

Multi-Group Confirmatory Factor Analysis can be applied to address the question whether differences in IQ test score between groups reflect true, i.e., latent differences in ability (Lubke et al., 2003a). We now present in detail the confirmatory factor model which can be used to this end (c.f. Bollen, 1989; Lubke et al., 2003a; Sörbom, 1974).

Let y_{ij} denote the observed p-dimensional random column vector of subject j in population i. We specify the following model for y_{ij}:

$$y_{ij} = \tau_i + \Lambda_i \eta_{ij} + \varepsilon_{ij}, \quad (1)$$

where η_{ij} is a q-dimensional random vector of correlated common factor scores (q < p), and ε_{ij} is a p-dimensional vector of residuals that contain both random error and unique measurement effects. The (p x q) matrix Λ_i contains factor loadings, and the (p x 1) matrix τ_i contains measurement intercepts. It is generally assumed that ε_{ij} is p-variate normally distributed with zero means and a diagonal covariance matrix Θ_i, i.e., residual terms are mutually uncorrelated. Furthermore, the vector η_{ij} is assumed to be q-variate normally distributed with mean a_i and (q x q) positive definite covariance matrix Ψ_i. Given these assumptions, the observed variables are normally distributed $y_{ij} \sim N_p(\mu_i, \Sigma_i)$, where, assuming the covariance between η_{ij} and ε_{ij} is zero:

$$\mu_i = \tau_i + \Lambda_i a_i \quad (2)$$

$$\Sigma_i = \Lambda_i \Psi_i \Lambda_i' + \Theta_i. \quad (3)$$

Note that superscript t denotes transposition.

We identify a sufficient number of fixed zeroes in Λ_i to avoid rotational indeterminacy given correlated common factors. In the same matrix Λ_i, we fix certain elements to equal 1 to identify the variances of the common factors. Similarly, for reasons of identification, we model latent differences in means instead of latent means themselves (Sörbom, 1974: see below).

Table 4.1

Summary of models in case of the two cohorts 1 and 2

No.	Description	Σ_1=	Σ_2=	μ_1=	μ_2=
1	Configural invariance	$\Lambda_1\Psi_1\Lambda_1'+\Theta_1$	$\Lambda_2\Psi_2\Lambda_2'+\Theta_2$	τ_1	τ_2
2	Metric invariance	$\Lambda\Psi_1\Lambda'+\Theta_1$	$\Lambda\Psi_2\Lambda'+\Theta_2$	τ_1	τ_2
3	Equal residual variances	$\Lambda\Psi_1\Lambda'+\Theta$	$\Lambda\Psi_2\Lambda'+\Theta$	τ_1	τ_2
4a	Strict factorial invariance	$\Lambda\Psi_1\Lambda'+\Theta$	$\Lambda\Psi_2\Lambda'+\Theta$	τ	$\tau+\Lambda\delta$
4b	Strong factorial invariance	$\Lambda\Psi_1\Lambda'+\Theta_1$	$\Lambda\Psi_2\Lambda'+\Theta_2$	τ	$\tau+\Lambda\delta$

Note: Except for step 4b (nested under 2) each model is nested under the previous one; Between-cohort differences in common factor means are expressed by δ (i.e., $\delta = a_2 - a_1$).

Factorial invariance can be investigated by fitting a series of increasingly restrictive models. These are presented in Table 4.1. We fit three models without mean restrictions, namely configural invariance (Model 1; equal pattern of factor loadings; Horn & McArdle, 1992), metric invariance (Model 2; $\Lambda_1 = \Lambda_2$; factor loadings equal across cohorts; Horn & McArdle, 1992), and a model with equal factor loadings and equal residual variances (Model 3; $\Lambda_1 = \Lambda_2$ & $\Theta_1 = \Theta_2$). In the next two steps we impose additional restrictions on the mean structure, and fit two models that are denoted strong factorial invariance (Model 4b) and strict factorial invariance (Model 4a; Meredith, 1993).[26] The latter involves the equality

[26] Note that these models go by different names. Model 2 is also known as Weak Factorial Invariance (Widaman & Reise, 1997) or Pattern Invariance (Millsap, 1997a), whereas Steenkamp and Baumgartner (1998) denote step 4b by Scalar Invariance.

of intercepts ($\tau_1 = \tau_2$), in addition to equality of factor loadings and residual variances. Observed mean differences are then due to common factor mean differences: $m_2 - m_1 = \Lambda(a_2 - a_1)$. Strong factorial invariance does not include the equality constraint on the residual variances ($\Theta_1 \neq \Theta_2$). Meredith (1993) has shown that for normally distributed data, strict factorial invariance within a factor model is required to demonstrate measurement invariance with respect to groups. As mentioned above, measurement invariance implies unbiasedness with respect to groups, or cohorts (Dolan et al., 2004; Lubke et al., 2003a; Mellenbergh, 1989). Strong factorial invariance is less restrictive in the sense that it allows unique/error-variances to differ between cohorts. One may argue that strong factorial invariance is sufficient in comparisons made between groups (Little, 1997). However, we fit both models and view the strong version as a minimal requirement for measurement invariance. Strict factorial invariance enables one to draw clearer conclusions concerning group differences (Lubke & Dolan, 2003).

In the context of the Flynn Effect we consider carefully the restriction on measurement intercepts ($\tau_1 = \tau_2$), necessary for both strong and strict factorial invariance. Note that the mean of a given subtest within the later cohort is a function of both the intercept and the common factor mean multiplied by the corresponding factor loadings (see Eq. 2). Intercept differences between groups imply uniform bias with respect to groups (Mellenbergh, 1989). In the present context, this may occur, if, say, one group has higher test sophistication or different test taking strategies that raise the scores in ways unrelated to latent intelligence (Brand, 1987). Therefore we define true intelligence differences between cohorts as factor score differences within a strict or strong factorially invariant factor model, and consequently we define true intelligence differences between cohorts as differences in the means (and possibly (co)variances) of these common factors.

We assume that the data are approximately normally distributed and fit models in the LISREL program (LISREL 8.54; Jöreskog & Sörbom, 2003) using maximum likelihood estimation. We assess model fit by the χ^2 in relation to Degrees of Freedom (DF), and by other fit indices such as the RMSEA (Browne & Cudeck, 1993), the CFI (Bentler, 1990), and the AIC and CAIC (cf. Jöreskog & Sörbom, 2003). The relative fit of the models in Table 4.1 can be assessed with these indices, with lower values of AIC and CAIC indicating better fit. By rule of thumb, a given model is judged to be a reasonable approximation if RMSEA is about .05 or lower, and CFI is greater than 0.95. We view the χ^2 in relation to degrees of freedom as a measure of badness of fit, rather than a formal test of exact fit (Jöreskog, 1993). The Comparative Fit Index (CFI) gives the relative fit of a model in relation to a null model of complete independence. Widaman and Thompson (2003) have argued that because of the nesting of models it is inappropriate to use such a null model within a multi-group context. Therefore, we use a model without any factor structure, in which intercepts and residual variances are restricted to be group invariant (i.e., model 0A in Widaman & Thompson, 2003) as the null model in computing the CFI values.

We use a stepwise approach, in which increasingly more across-cohort constraints are introduced. If a given equality constraint leads to a clear deterioration in fit (i.e., difference in χ^2, in relation to difference in DF), we conclude that the particular constraint is untenable. If so, modification indices can pinpoint the source, in terms of parameters, responsible for misfit. Modification Indices (MIs) are measures of how much chi-square is

expected to decrease if a constraint on a given set of parameters is relaxed, and the model is re-fitted (Jöreskog & Sörbom, 2003). We now turn to the confirmatory factor analyses of the five datasets.

4.3 Study 1: Dutch Adults 1967/1968 and 1998/1999: WAIS

Samples

The Wechsler Adult Intelligence Scale (WAIS) was translated in Dutch more than thirty-five years ago (Stinissen, Willems, Coetsier, & Hulsman, 1970). Here we compare the 1967/1968 standardization sample of the Dutch WAIS (N = 2100) with 77 Dutch subjects who completed the WAIS during standardization of the WAIS-III in 1998 and 1999 (Wechsler, 2000). Mean age of the nineties sample is 40.3 years (SD = 14.0). In terms of the WAIS-III scores, this sample appears representative, with a mean WAIS-III IQ of 100.6 and a standard deviation of 14.8 (Wechsler, 2000). However, it should be noted that the original Dutch WAIS-III standardization sample is slightly underrepresented with respect to subjects from low-educational backgrounds (Swets & Zeitlinger, 2003; Tellegen, 2002). Therefore these WAIS-III IQ's are an underestimation of approximately 2 IQ points (Swets & Zeitlinger, 2003).

In the 1998/1999 sample, the WAIS administration followed between two and twelve weeks after administration of the WAIS-III. This quasi-retest could have resulted in an increase in WAIS subtest scores. However, the subtests of the WAIS-III have been altered and the percentage of overlapping items of the WAIS-III and WAIS (mean per subtest: 50%) is smaller than that found in comparisons of for example the WAIS versus the WAIS-R (84%) in the US. Furthermore, the differential gains of the subtests reported below do not seem to reflect those that show the largest retest-effect (e.g., Catron & Thompson, 1979; Matarazzo, Wiens, Matarazzo, & Manaugh, 1973). Nevertheless, a test of factorial invariance of these data sets is considered relevant since Flynn (1984; 1998c) has used data sets where administrations of an older version were preceded by the administration of a new one, or vice versa. Our focus is primarily on factorial invariance between the cohorts, more representative samples without the possible retest-effect should be used to investigate WAIS-IQ gains of the general Dutch population.

Measures

The WAIS contains eleven subtests: Information (INF), Comprehension (COM), Arithmetic (ARI), (SIM), Digit Span (DSP), Vocabulary (VOC), Digit Symbol (DSY), Picture Completion (PCO), Block Design (BDE), Picture Arrangement (PAR), Object Assembly (OAS). Appendix A contains a brief description of all subtests (c.f. Stinissen, 1977; Stinissen et al., 1970; Wechsler, 1955). The confirmatory factor analyses are based on an oblique three-factor model, which includes the common factors: verbal comprehension (INF, VOC, COM, SIM), perceptual organization (PCO, PAR, BDE, OAS, DSY), and memory/freedom from distractibility (DSP, ARI, DSY). This factor model is displayed in Figure 4.1.

Results and Discussion

Correlations between subtests as well as means and standard deviations of both cohorts are reported in Table 4.2[27]. As can be seen from the mean differences between cohorts, the Flynn Effect is present on all subtests, with effect sizes (in 1967/1968 SD-units) varying from 0.51 (Digit Span) to 1.48 (Similarities). This results in IQ-increases of 15.5, 22.4, and 19.8 for Verbal-IQ, Performance-IQ and Total-IQ, respectively. These IQ gains are in line with gains on the WAIS(-R) found in the US and in Germany (Flynn, 1998c; Satzger, Dragon, & Engel, 1996).

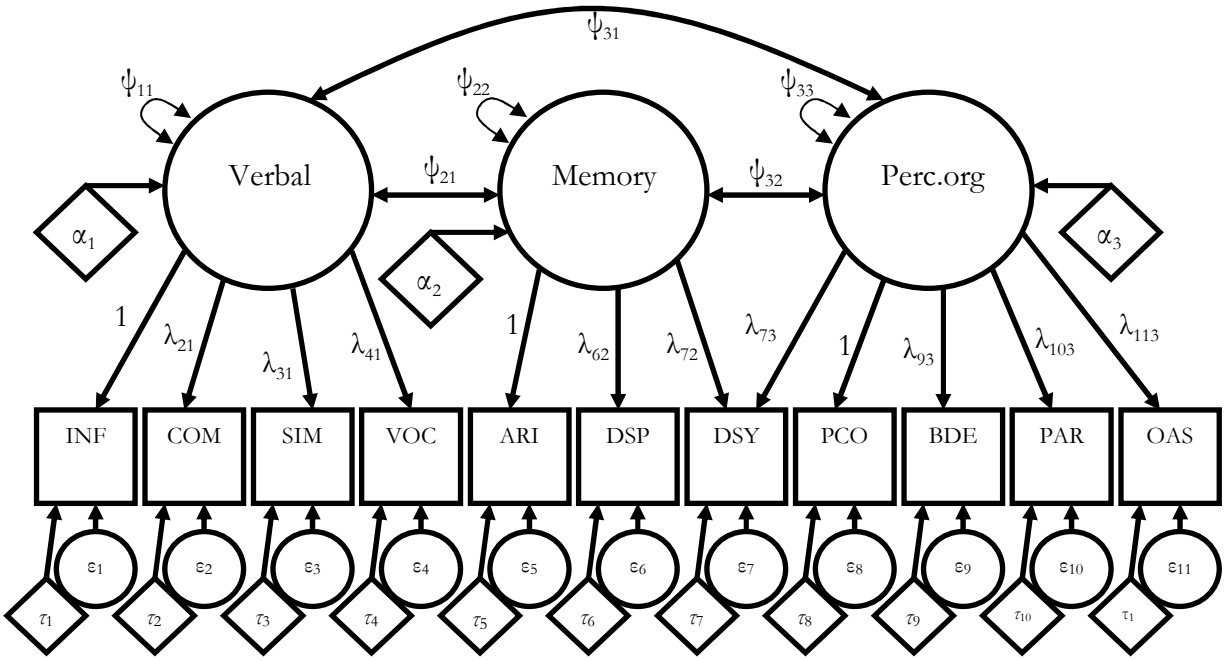

Figure 4.1 *WAIS factor model.*

[27] We include summary statistics in this paper, so that the interested reader may investigate factorial invariance using alternative (factor) models. The LISREL input files for all analyses carried out here can be downloaded from http://users.fmg.uva.nl/jwicherts/.

Table 4.2
Correlations and descriptive statistics of WAIS 1967/1968-1998/1999

	INF	COM	ARI	SIM	DSP	VOC	DSY	PCO	BDE	PAR	OAS
INF		.63	.57	.67	.35	.76	.33	.49	.32	.35	.05
COM	.66		.60	.67	.34	.73	.29	.41	.32	.33	.18
ARI	.57	.52		.56	.43	.52	.43	.42	.34	.48	.10
SIM	.67	.67	.53		.36	.75	.36	.49	.42	.40	.09
DSP	.43	.40	.48	.41		.35	.51	.27	.16	.32	.05
VOC	.75	.71	.55	.72	.44		.34	.43	.33	.31	.05
DSY	.45	.41	.44	.43	.39	.49		.47	.39	.55	.29
PCO	.50	.44	.39	.47	.34	.50	.44		.55	.58	.42
BDE	.41	.42	.43	.44	.37	.44	.45	.46		.60	.42
PAR	.41	.35	.31	.39	.26	.44	.39	.49	.43		.37
OAS	.34	.33	.28	.36	.21	.36	.38	.46	.49	.41	
	INF	COM	ARI	SIM	DSP	VOC	DSY	PCO	BDE	PAR	OAS
M '67/68	9.10	14.13	7.60	10.93	11.17	27.63	47.47	9.53	13.27	10.73	36.00
SD '67/68	5.44	5.52	3.80	5.55	3.33	12.10	12.83	3.54	6.42	4.45	14.88
M'98/99	13.78	20.84	11.10	19.14	12.88	40.22	58.58	13.51	20.41	14.38	44.65
SD '98/99	4.82	4.21	2.96	4.35	3.66	10.21	13.20	3.03	5.93	3.74	15.10
Effect Size	0.86	1.22	0.92	1.48	0.51	1.04	0.87	1.13	1.11	0.82	0.58

Note. Correlations of 1967/1968 sample (N = 1100) below diagonal and 1998/1999 sample (N = 77) above diagonal. Effect sizes in 1967/1968 SD units.

Table 4.3
Fit indices test for factorial invariance WAIS 1967/1968-1998/1999

Model	Equality constraints	χ^2	DF	Compare	$\Delta\chi^2$	ΔDF	RMSEA	CFI	AIC	CAIC
1	-	274.9	80				0.047	0.994	418	913
2	Λ	279.7	89	2 vs 1	4.8	9	0.044	0.994	406	841
3	Λ & Θ	332.3	100	3 vs 2	52.6	11	0.044	0.993	421	782
4a	Λ & Θ & τ	408.5	108	4a vs 3	76.2	8	0.050	0.990	494	801
4b	Λ & τ	368.1	97	4b vs 2	88.4	8	0.051	0.991	484	865

The fit indices of the factor models differing with respect to between-cohort equality constraints are reported in Table 4.3. The model with identical configuration of factor loadings in both cohorts (Model 1; configural invariance) fits poorly in terms of Chi-square. However, the large χ^2 is due to the large standardization sample (Bollen & Long, 1993), and RSMEA and the CFI indicate that this baseline model fits sufficiently. In the second model (Model 2; metric invariance) we restrict factor loadings to be equal across both cohorts (i.e., $\Lambda_1 = \Lambda_2$). All fit indices indicate that this is does not result in an appreciable deterioration in model fit, and therefore this constraint seems tenable. However, the restriction imposed on the residual variances (Model 3; $\Theta_1 = \Theta_2$) is not completely tenable, since AIC and $\Delta\chi^2$ indicate a clear deterioration in fit as compared to

the metric invariance model. However, RMSEA, CFI, and CAIC indicate that this restriction is tenable. In a formal sense, residual variances are unequal across groups, although the misfit due to this restriction is not large. More importantly, both models (4a and 4b) with equality constraints on the measurement intercepts ($\tau_1 = \tau_2$) show insufficient fit. The RMSEA-values are larger than the rule-of-thumb-value of 0.05, and (C)AICs show larger values in comparison to the values of the third model. Although the CAIC values of models 4a and 4b are somewhat lower than the CAIC of the unrestricted model (reflecting CAIC's strong preference for parsimonious models), the difference in chi-square comparing models 4a and 4b to less restricted models is very large.[28] Both strong factorial invariance and strict factorial invariance therefore appear to be untenable. This means that measurement intercepts of the cohorts are unequal and consequently that mean differences in test scores (the Flynn Effect) on this Dutch WAIS-test cannot be explained by latent (i.e., common factor mean) differences between the 1967/1968 and 1998/1999 samples.

However, using MGCFA it is possible to relax selected constraints in an ill-fitting model, to investigate the source of misfit, and, perhaps to arrive at an interpretable modified model. We now turn to a modification of the strong factorial invariance model that we denote by *partial* strong factorial invariance (Byrne et al., 1989). In this model we free the parameters with the highest modification indices (in Model 4b), namely the intercepts of Similarities (MI = 33) and Comprehension (MI = 20). By allowing these parameters to differ between the cohorts, we attain a model with acceptable fit ($\chi^2 = 301.3$, DF = 95, RMSEA = 0.044, CFI = 0.993, AIC = 414, CAIC = 809). This enables a cautionary interpretation of the factor mean gains ($a_2 - a_1$) thus found. The parameter estimates of the gains in this partial invariance model are: memory/freedom from distractibility: 2.34 (SE = 0.25, Z = 9.20, p< .01); verbal comprehension: 5.11 (SE = 0.49, Z = 10.53, p< .01); perceptual organization 3.69 (SE = 0.33, Z = 11.27, p< .01). Thus, all three common factors show significant gains. It should be noted that this model must be seen as a post hoc (exploratory) analysis, and that mean differences on the Similarities and Comprehension subtests are now *unexplained* by the factor on which these load.

This partial strong invariance model has three correlated (oblique) first-order factors, which interrelatedness can be explained by a second-order factor, which can be denoted by *g* or general intelligence. This enables a test of the hypothesis that the score gain found in the current comparison could be solely due to increases in this higher order factor. Note that this second-order model with additional constraints is nested under the partial strong factorial invariance model above (without such a higher-order factor). We found that the second-order model has group-invariant second order factor loadings (invariance test: $\Delta\chi^2 = 1.0$, ΔDF = 2), and group-invariant first order factor variances (invariance test of $\Psi_1 = \Psi_2$; $\Delta\chi^2 = 1.5$, ΔDF = 3). In the second order model with invariant second-order factor loadings and invariant first-order factor variances, we allow only second-order factor mean and second order factor variance differences. This second-order model has the following fit indices: $\chi^2 = 321.5$, DF = 102, RMSEA = 0.044, CFI = 0.993, AIC = 423,

[28] Note that the CFI does not differentiate well between the models. This is primarily due to the fact that inter-subtest correlations are high and therefore the null-model has a very large chi-square. Even if, say, the chi-square of Model 4a would have been 1500, the CFI still would assume a value well above 0.95. This renders the CFI less suitable for investigating between-group restrictions in this data set.

CAIC = 771. It appears that this model fits reasonably, although the high modification index (MI = 17) of the factor mean difference in the perceptual organization (first order) factor suggests that the gains are not solely due to general intelligence.

In conclusion, although the overall gains found in this comparison are unexplained by the factor mean differences, a cautionary conclusion would be that part of the gains (excluding the subtests Similarities and Comprehension) could be explained by genuine increases in intelligence.

4.4 Study 2: Danish Draftees 1988 and 1998: Børge Prien's Prøve

Samples

The data in this comparison stem from Teasdale and Owen (2000) who compared several cohorts of Danish draftees, tested in the year they turn 18. The data includes all Danish draftees of 1988 (N = 33.833) and 1998 (N = 25.020), comprising about 90 to 95 percent of the Danish male population of 18 year-olds of those years (Teasdale & Owen, 1989, 2000).

Measures

All draftees completed a group test of cognitive abilities named Børge Prien's Prøve (BPP), which includes four subtests: Letter Matrices (LEM), Verbal Analogies Test (VAT), Number Series Test (NST) and Geometric Figures Test (GFT). These subtests are characterized by fluid and abstract (Teasdale & Owen, 1987, 1989, 2000). A short description of the subtests is given in Appendix B. The factor model used has one factor with four indicators. Although this is a small number of subtests for a factor model, this single factor model is consistent with the common use of a total test score based on these subtests (see e.g., Teasdale & Owen, 1987). More practically, the tenability of this model should be judged by its fit. We use (normal theory) maximum likelihood estimation even though the data are slightly negatively skewed (Teasdale & Owen, 2000), since maximum likelihood (ML) estimation is quite robust to mild skewness.

Table 4.4
Correlations and descriptive statistics of Børge Prien's Prøve 1988-1998

	LEM	VAT	NST	GFT
LEM		.56	.61	.47
VAT	.57		.59	.47
NST	.62	.61		.43
GFT	.48	.49	.45	
	LEM	VAT	NST	GFT
M 1988	9.99	12.27	9.61	10.06
SD 1988	2.59	4.02	3.11	3.18
M 1998	10.18	12.53	9.80	10.57
SD 1998	2.46	3.93	3.04	3.18
Effect size	0.07	0.06	0.06	0.16

Note. Correlations of 1988 sample (N = 33833) below diagonal and of 1998 sample (N = 25020) above diagonal. Effect sizes in 1988 SD units.

Results and Discussion

Descriptive statistics of both cohorts are reported in Table 4.4. As previously described by Teasdale and Owen (2000), the largest increase between 1988 and 1998 is found on the Geometric Figures Test. It is also apparent that the overall gain is small in terms of 1988 SD-units. Furthermore it is noteworthy that standard deviations of all subtests but the Geometric Figures Test have decreased in the ten-year period. Teasdale and Owen (2000) show that the overall standard deviation decline is mostly caused by the fact that gain is strongest in the lower end of the distribution. In addition, they conclude that this is probably not caused by a ceiling effect.

Table 4.5
Fit indices test for factorial invariance of Børge Prien's Prøve 1988-1998

Model	Equality constraints	χ^2	DF	Compare	$\Delta\chi^2$	ΔDF	RMSEA	CFI	AIC	CAIC
1	-	471.9	4				0.062	0.955	507	746
2	Λ	475.5	7	2 vs 1	3.7	3	0.047	0.955	504	714
3	Λ & Θ	547.1	11	3 vs 2	71.6	4	0.040	0.949	565	735
4a	Λ & Θ & τ	782.4	14	4a vs 3	235.3	3	0.043	0.926	797	936
4b	Λ & τ	710.1	10	4b vs 2	234.6	3	0.048	0.932	734	913

Teasdale and Owen (2000) state that the similarity of test inter-correlations across both cohorts is striking (p. 117). We now use these data to test whether factorial invariance with respect to cohorts is tenable. This enables us to unravel whether or not the Danish gains reflect true (i.e., latent) gains in intelligence. Table 4.5 contains the fit indices of the different factor models used to this end. As can be seen, the model without across-cohort equality constraints (Model 1; configural invariance) has a very large χ^2. However, the sample sizes are again large and both the RMSEA, and the CFI indicate that the fit of the baseline model is sufficient. In the second model (metric invariance) factor loadings are constraint to be cohort-invariant (i.e., $\Lambda_1 = \Lambda_2$). This step is accompanied by a relative improvement in fit, with all fit indices having better values in Model 2 than in Model 1. Therefore we conclude that metric invariance is tenable. The various fit indices with which we can judge the tenability of the next restriction on the residual variances (Model 3; $\Theta_1 = \Theta_2$) are somewhat inconsistent. The RMSEA indicates an improvement in model fit from Model 2 to Model 3, while $\Delta\chi^2$, CFI, AIC and CAIC show deterioration in fit. The highest modification index in this step is found on the parameter of the residual variance of the Letter Matrices Test (MI = 67). Regardless of the conclusion about the equality of the unique/error-variances, the subsequent restriction of cohort-invariant measurement intercepts (i.e., $\tau_1 = \tau_2$) leads to a clear deterioration in fit, with all fit indices assigning poorer values in Models 4a and 4b as opposed to the Models 3 and 2 (i.e., models without this mean restriction). Therefore, both strong factorial invariance (Model 4a), and strict factorial invariance (Model 4b) are rejected. Thus, we conclude that the Flynn Effect found in this Danish comparison cannot be explained by an increase in latent intelligence (i.e., factor mean differences between cohorts).

We should note the sample size is accompanied by great power to reject models. This power issue can be investigated using simulation studies. A pragmatic alternative could

be to treat the data as if it was composed of a smaller number of cases (see Muthén, 1989). We have used the number of cases command of LISREL to this end, and found that in case of 1000 subjects in each cohort the results are similar to those found with the original number of cases. Therefore, a reasonable number of cases would have led to the same results and power appears not to be the main reason for the rejection of the factorial invariance models.

As shown by the modification indices of Model 4b, the rejection of the intercept-restriction is primarily caused by the intercept of Geometric Figures (MI = 231). As noted, this subtest shows greater increase than the other subtests. We could again free this intercept parameter, together with the aforementioned residual variance-parameter of the Letter Matrices Test. The model found by allowing these two parameters to differ between cohorts shows sufficient fit (χ^2= 483.2, DF = 12, RMSEA = 0.036, CFI = 0.955 AIC = 502, CAIC = 662). In this *partial* strict factorial model the factor mean of the 1998 cohort differs significantly from the factor mean of the 1988 cohort: the parameter estimate of $a_2 - a_1$ is 0.17 (SE = 0.018, Z = 9.49, p< .01). Again, a careful conclusion would be that some, but apparently not all, mean differences between the cohorts could be explained by a latent increases in intelligence. Furthermore, the partial strict factorial invariance model shows that the (latent) factor variance in the second cohort is smaller (3.67, SE = 0.047) than the factor variance of the first cohort (3.96, SE = 0.046). The latter is consistent with earlier findings (Teasdale and Owen, 1989) and the results in Teasdale and Owen (2000). They noted that the gains over the cohorts appear to be larger at the lower end of the distribution. In their 1989 paper, Teasdale and Owen have put some effort into finding out whether this differential gain is caused by a ceiling effect of the test itself. Their simulation of data suggested that a ceiling effect is not the reason for the diminishing test score variance until 1987. However, in the current comparison of the 1988 and 1998 cohorts, not only the factor variance, but also the residual variance of LMT is smaller. The possibility of a ceiling effect on this subtest in the current comparison can therefore not be ruled out.

A shortcoming of the current data set is the small number of subtests and as a result the simple factor structure. It remains unclear whether the results would haven been similar in case the test consisted of more scales and factors. However, the fit indices show sufficient fit of the one-factor model.

In conclusion, it appears that gains found on Børge Prien's Prøve from 1988 to 1998 could not be fully explained by latent increases in the factor model. Especially the large gains on Geometric Figures Test need further explanation as well as the diminishing residual variance of the Letter Matrices Test. The latter implies that ceiling effects may play a role in decreasing test score variance in this valuable Danish data set.

4.5 Study 3: Dutch High School Students 1984 and 1994/1995: DAT'83

Samples

During the standardization of the Dutch version of the DAT, Evers and Lucassen (1992) collected data from 3300 third-year high school students at the three major Dutch educational levels, namely MAVO (medium-low level), HAVO (medium-high) and VWO (high). Here we compare the standardization samples of these three levels (with 1100 cases

each) acquired from 1982 to 1986 (median in 1984) with high school students on the corresponding levels in 1994 and 1995 (from Oosterveld, 1996). Whereas the 1984 standardization samples are selected to be representative for Dutch children at their respective educational levels (Evers & Lucassen, 1992), the 1994/1995 subjects were not sampled to be representative. Nevertheless, the latter data stem from ten different schools in different parts of The Netherlands. These (regional) high schools are located in middle-sized towns and therefore the students are from both rural and urban areas. The 1994/1995 samples contain a total 922 subjects, of which 490 females (of eleven subjects gender was unknown). Because Evers and Lucassen (1992) found large sex differences on the DAT, we randomly selected 93% of the females in order to equal the gender-proportion of the three nineties cohorts to the gender-proportion (50% female) of the 1984 standardization samples. The remaining numbers of cases for the 1994/1995 cohorts are: 397 for MAVO, 272 for HAVO, and 188 for VWO. Information on the social economic background of individual students is missing, although information on the schools indicates that the ethnic composition of the schools does not greatly deviate from that of the overall Dutch population. As a matter of fact, seven of the ten schools in the nineties cohort also participated in the 1984 standardization. Thus, the representativeness of the nineties samples seems mainly to be compromised by the omission of subjects from large-sized towns such as Amsterdam. Precise age of the subjects during testing is unknown, but the mean would normally lie around 14½ years. Importantly, there is no reason to expect differences in age composition of the 1984 and 1994/1995 cohorts. In addition, some changes in the composition of the levels could have occurred, although the Dutch high school system did not undergo any systematic change between 1982 and 1995.

Measures

The Dutch Differential Aptitude Test (DAT '83; Evers & Lucassen, 1992) is a group intelligence test containing nine subtests with a time limit. The Dutch DAT is largely an adaptation of the American DAT (form S&T) with one additional vocabulary scale (Evers & Lucassen, 1992). Since two subtests were not deemed informative by the school authorities, a significant part of the nineties sample was not administered the Mechanical Reasoning (MR) (40% missing), and/or the Speed & Accuracy (SA) subtest (64% missing). This resulted in a shortening of the testing session for these subjects, but this appears not have resulted in higher scores on the remaining subtests. Probably because of the breaks in between subtests, the scores of these subjects on the subtests that would have followed MR and SA did not significantly differ from the corresponding scores of subjects that were administered both subtests. Therefore, we pool both groups and leave the two missing subtests out of the current comparison. The seven remaining subtests are: Vocabulary (VO), Spelling (SP), Language Use (LU), Verbal Reasoning (VR), Abstract Reasoning (AR), Spatial Relations (SR), and Numerical Ability (NA). Appendix C contains a description of these subtests. Throughout we apply an oblique two-factor model, roughly similar to the first two factors of the factor solution described in the manual (Evers & Lucassen, 1992). These factors can be denoted by a verbal factor (VO, SP, LU, VR, NA) and an abstract factor (VR, AR, SR, NA).

Table 4.6
Correlations and descriptive statistics of DAT '83 1984-1995 medium-low level (MAVO)

	VO	SP	LU	VR	AR	SR	NA
VO		0.13	0.51	0.33	0.13	0.18	0.11
SP	0.23		0.19	0.11	0.02	-0.04	0.10
LU	0.55	0.32		0.26	0.12	0.19	0.09
VR	0.36	0.17	0.35		0.34	0.38	0.24
AR	0.27	0.08	0.27	0.40		0.58	0.44
SR	0.28	-0.04	0.21	0.39	0.52		0.35
NA	0.25	0.16	0.18	0.32	0.42	0.38	
	VO	SP	LU	VR	AR	SR	NA
M 1984	42.5	59.2	29.4	18.7	33.3	30.7	17.7
SD 1984	10.2	8.4	6.9	7.2	7.3	9.3	6.0
M 1994/1995	39.89	58.97	27.05	17.32	32.61	30.76	15.14
SD 1994/1995	9.07	7.95	5.82	7.87	7.30	10.56	5.49
Effect size	-0.26	-0.03	-0.34	-0.19	-0.09	0.01	-0.43

Note. Correlations of 1984 sample (N = 1100) below diagonal and of 1994/1995 sample (N = 397) above diagonal. Effect sizes in 1984 SD units.

Results and Discussion

We now present results for each educational level separately, beginning with the lowest level. Means, standard deviations and inter-subtest correlations of both MAVO-cohorts are reported in Table 4.6. As can be seen from the effect sizes, there is no Flynn Effect in this subgroup. All but one subtest (Spatial Relations) show a decrease in scores from 1984 to 1994/1995. A further breakdown on gender shows no clear gender differences. These declining scores could have been the result of imperfect sampling of the nineties cohort, such as the aforementioned lack of subjects from large cities or perhaps by a changing composition of the low level educational group. Whatever the reasons for the decline, it is reassuring to see the similarity to the pattern of gains found on the Spanish DAT between 1979 and 1995 (Colom et al., 1998; 2001). Since four of the current DAT subtests (SR, AR, VR and NA) are also present in the Spanish DAT, we can compare effect sizes (i.e., gains/losses) on subtests in both countries. These four effect sizes of the MAVO- comparison correlate highly (pmcc = 0.90; spearman = 0.80) with the Spanish effect sizes found by Colom and colleagues (Colom et al., 1998; 2001).

Since our main interest is in whether the Flynn Effect is accompanied by factorial invariance, we leave out our findings on factorial invariance in this MAVO group. However, results with respect to the tenability of factorial invariance of the MAVO cohorts are in line with the following results of the HAVO cohorts.

Table 4.7
Correlations and descriptive statistics of DAT '83 1984-1995 medium-high level (HAVO)

	VO	SP	LU	VR	AR	SR	NA
VO		0.29	0.53	0.32	0.10	0.13	0.02
SP	0.31		0.35	0.09	0.03	-0.09	0.04
LU	0.51	0.36		0.39	0.20	0.18	0.03
VR	0.28	0.16	0.33		0.43	0.38	0.19
AR	0.10	0.04	0.17	0.36		0.61	0.42
SR	0.18	0.00	0.13	0.37	0.53		0.32
NA	0.12	0.13	0.13	0.23	0.35	0.29	
	VO	SP	LU	VR	AR	SR	NA
M 1984	49.8	64.5	34.5	23.6	37.5	35.9	22.2
SD 1984	9.7	8.3	6.5	8.6	6.2	10.3	6.0
M 1994/1995	48.78	66.68	35.55	23.87	37.80	37.05	20.07
SD 1994/1995	9.37	8.91	7.31	8.37	6.08	10.51	6.04
effect size	-0.11	0.26	0.16	0.03	0.05	0.11	-0.36

Note. Correlations of 1984 sample (N = 1100) below diagonal and of 1994/1995 sample (N = 272) above diagonal. Effect sizes in 1984 SD units.

The subtest correlations, as well as the descriptives of both medium-high level (HAVO) cohorts are reported in Table 4.7. In these data, a Flynn Effect is present, with the highest increase on the subtest Spelling. Nevertheless, the Numerical Ability and the Vocabulary subtests show a decrease from 1984 to 1994/1995. Again, the relative gain of the four corresponding DAT scales shows striking similarity to gains found in Spain (Colom et al., 1998), with a correlation (pmcc) between the effect sizes in both countries of 0.82 (spearman = 0.80). Since it has been suggested that the Spanish DAT gains are compatible with increases in *g,* i.e., with a "Jensen effect"[29] (see Colom et al., 2001), it is interesting to check whether the HAVO gains can be considered factorially invariant with respect to cohort, since factorial invariance is a crucial aspect of the hypothesis that the manifest gains are due to gains in *g*.

Table 4.8
Fit indices test for factorial invariance of DAT '83 1984-1995 medium-high level (HAVO)

Model	Equality constraints	χ^2	DF	Compare	$\Delta\chi^2$	ΔDF	RMSEA	CFI	AIC	CAIC
1	-	61.9	22				0.051	0.983	158	456
2	Λ	70.0	29	2 vs 1	8.1	7	0.045	0.982	152	407
3	Λ & Θ	86.2	36	3 vs 2	16.2	7	0.045	0.978	154	366
4a	Λ & Θ & τ	153.3	41	4a vs 3	67.1	5	0.063	0.952	210	390
4b	Λ & τ	136.2	34	4b vs 2	66.2	5	0.065	0.958	204	428

[29] A Jensen Effect occurs when *g* loadings of (sub)tests correlates significantly with the (sub)tests' correlations with other variables

Fit indices of the models leading up to factorial invariance in the HAVO-comparison are reported in Table 4.8. The first model fits sufficiently as judged by RMSEA and CFI. The step from the configural invariance model to the metric invariance model (Model 2; $\Lambda_1 = \Lambda_2$) is accompanied by a very slight decrease in CFI, but all other fit measures improve and therefore factor loadings appear invariant over cohort. With respect to the next restriction of equal residual variances (Model 3; $\Theta_1 = \Theta_2$), the AIC shows a small increase, and the CFI drops slightly. The other fit indices indicate that residual variances are cohort-invariant. More importantly, in comparison to Models 3 and 2, both factorial invariance models (4a and 4b) show a clear decline in all fit indices (although CAICs in steps 4a and 4b are still lower than the CAIC of Model 1). Considering the large $\Delta\chi^2$, the drop in CFI, and the clear increase in RMSEA, we conclude that the equality-restriction on the measurement intercepts ($\tau_1 = \tau_2$) is untenable and therefore that the Dutch increase in DAT test scores at this educational level cannot be explained by increases in latent intelligence.

Here we again consider the partial strong factorial invariance model, and relax the intercepts associated with the largest modification indices. The measurement intercepts of Numerical Ability (MI = 36) and Vocabulary (MI = 18) seem to be the cause of the poor fit of the factorial invariance model. Note that both scales showed a decline from 1984 to 1994/1995. When the intercepts of both tests are freed we obtain an acceptable model fit (χ^2 = 79.58, DF = 32, RMSEA = 0.046, CFI = 0.980, AIC = 112, CAIC = 390). In this partial strong factorial invariance model the factor mean of the verbal factor is significantly higher in the 1994/1995 sample as opposed to the 1984 sample (1.46, SE = 0.56, Z = 2.60, p< .01), whereas the abstract factor does not show a significant gain from 1984 to 1994/1995 (parameter estimate 0.37, SE = 0.38, Z = 0.97, p> .05).

Table 4.9
Correlations and descriptive statistics of DAT '83 1984-1995 high level (VWO)

	VO	SP	LU	VR	AR	SR	NA
VO		0.36	0.57	0.40	0.11	0.22	0.16
SP	0.39		0.37	0.25	0.19	0.12	0.24
LU	0.56	0.45		0.38	0.15	0.11	0.15
VR	0.38	0.32	0.44		0.45	0.42	0.41
AR	0.15	0.14	0.22	0.35		0.54	0.38
SR	0.19	0.08	0.15	0.39	0.53		0.33
NA	0.20	0.22	0.19	0.29	0.31	0.33	
	VO	SP	LU	VR	AR	SR	NA
M 1984	56	70.9	39.9	30.1	40	40	26.3
SD 1984	9.4	8.7	7.2	8.8	5.3	9.7	5.8
M 1994/1995	51.12	69.37	37.18	24.45	40.03	38.93	23.86
SD 1994/1995	9.81	8.84	6.93	9.59	5.18	9.85	6.22
effect size	-0.52	-0.18	-0.38	-0.64	0.01	-0.11	-0.42

Note. Correlations of 1984 sample (N = 1100) below diagonal and of 1994/1995 sample (N = 188) above diagonal. Effect sizes in 1984 SD units.

Next, we turn to the highest educational level, denoted VWO. Descriptive statistics and subtest-correlations of both VWO-cohorts are reported in Table 4.9. As was the case in the medium-low educational level (MAVO) above, Flynn Effect seems absent at this educational level. Again, this could be due to sampling or to changing composition of the educational levels. Like the MAVO-comparison, we skip the test for factorial invariance, although we should note that results indicate that again factorial invariance is untenable. In addition, the effect sizes of the four overlapping subtests (AR, SR, NA and VR) show similarity with the Spanish DAT-gains (pmcc = 0.79, spearman = 0.80).

In conclusion, the DAT shows clear gains in scores only at the medium-high educational level (HAVO), whereas the medium-low (MAVO) and high (VWO) levels show no increase. It is interesting that this result agrees with the pattern of gains that Spitz (1989) reported on the WAIS and WAIS-R. Further research based on better sampling could clear up the issue of Dutch DAT-gains. Irrespective of the causes of these conflicting findings, we found that the DAT is biased with respect to cohort. The gains found at the HAVO level and the losses found at the other levels can thus not be explained by latent (i.e., factor mean) differences in intelligence. This conclusion runs counter to the finding that the Spanish DAT gains are related to the g factor (Colom et al., 2001). Nevertheless, the effect sizes on all three levels show clear similarity with Spanish DAT gains. Finally, a partial factorial invariance model in the HAVO-group reveals that some of the observed gains can be attributed to gains in the verbal common factor, but not in the abstract factor. The Numerical Ability and Vocabulary subtests show a decrease that could not be explained by latent differences between the cohorts.

4.6 Study 4: Dutch Children 1981/1982 and 1992/1993: RAKIT

Samples

In this study we compare 5-year-olds from the 1981/1982 standardization sample of the RAKIT (Bleichrodt et al., 1984) with a sample of 5-year-old twins (210 males and 205 females) that were tested in 1992 and 1993 (Rietveld, van Baal, Dolan, & Boomsma, 2000). The standardization sample (N=207) is representative of Dutch 5-year-olds in 1982 (Bleichrodt et al., 1984). The representativeness of the second cohort may be evaluated in the light of data on Socio-Economic Status (SES) as measured by the occupational status of the fathers. The 208 twin-pairs appear to be of somewhat higher SES (low 24%, middle 48%, high: 28%; Rietveld et al., 2000) than the overall 1993 Dutch population (32%, 44%, 24% respectively; Statistics Netherlands, 2003). Nevertheless, the nineties cohort is clearly composed of a broad sample of social backgrounds.

The raw test scores of both cohorts are normalized with respect to age. Because both cohorts contain cases out of two standardization age groups (i.e., 59 to 62 months, and 63 to 71 months; Bleichrodt et al., 1984), we also conducted analyses in each age group separately. However this produces similar results as those reported below. Although some information is lost by the normalization, the scores appear comparable across cohorts. Since the 1992/1993 cohort contains twin-pairs, the individual cases are not independent. For that reason, we conduct two sets of analyses, one for each twin. Each first twin is randomly assigned to twin Sample 1 or twin Sample 2, the second twin then is assigned to

the other twin sample. The twin data provides a useful opportunity to cross-validate the results of model fitting, in which the 1982 cohort is compared to both twin Sample 1, and twin Sample 2. Finally, we note that because of a missing subtest, we deleted one twin case in the second sample, whose monozygotic brother had an IQ of 84.

Measures

The RAKIT (Bleichrodt et al., 1984) is an individually administered Dutch intelligence test for children (aged 4 to 11 years) comprising 12 subtests. RAKIT-IQ has been shown to correlate 0.86 with IQ from the WISC-R (Bleichrodt et al., 1984). In the 1992/1993 cohort the shortened version of the RAKIT was administered. The IQ of this version has been shown to correlate 0.93 with the IQ of the total scale (Bleichrodt et al., 1984). The subtests of the shortened version are: Exclusion (EX), Discs (DI), Hidden Figures (HF), Verbal meaning (VM), Learning Names (LN), and Idea Production (IP). A description of these subtests is provided in Appendix D. Throughout, we use the oblique two-factor model presented by Rietveld et al. (2000), with three subtests loading on a nonverbal factor (EX, DI and HF) and three subtests loading on a verbal factor (LN, VM and IP).

Table 4.10a
Correlations and descriptive statistics of RAKIT 1982-1992/1993 (twin Sample 1)

	EX	VM	DI	LN	HF	IP
EX		0.34	0.40	0.34	0.30	0.14
VM	0.34		0.12	0.52	0.24	0.30
DI	0.39	0.28		0.15	0.30	0.10
LN	0.24	0.40	0.06		0.19	0.33
HF	0.39	0.30	0.28	0.26		0.08
IP	0.13	0.36	0.19	0.31	0.24	
	EX	VM	DI	LN	HF	IP
M 1982	15.01	15.17	14.95	14.97	15.37	14.94
SD 1982	5.02	5.10	4.99	4.97	5.06	4.99
M 1992-1	15.50	16.00	13.60	16.63	16.30	15.36
SD 1992-1	4.38	4.24	5.28	4.58	4.61	4.23
effect size-1	0.10	0.16	-0.27	0.33	0.18	0.08

Note. Correlations of 1982 sample (N = 207) blow diagonal and of 1992/1993 sample (N = 208) above diagonal. Effect sizes in 1982 SD units.

Results and Discussion

Descriptive statistics of the standardization sample and twin Sample 1 are reported in Table 4.10a, and descriptive statistics of twin Sample 2 are given in Table 4.10b. As can be seen from the effect sizes, all but the Discs subtest show higher scores in the 1992/1993 sample, with the highest gain on the Learning Names subtest. Furthermore, there are some differences between both twin-samples, but these are trivial. Average IQ in 1982 is 100 by definition. The increase of scores to 1992/1993 is reflected in average IQs of 102.6 (SD = 13.7) and 103.0 (SD = 12.6) in twin Sample 1 and 2, respectively. Considering the

somewhat higher SES of the nineties sample, these gains appear small in comparison to gains found on the WISC-R in the US (i.e., 5.3 IQ points from 1972 to 1989; Flynn, 1998c) and on the German WISC (20 IQ points from 1956 to 1983; Schallberger, 1987).

Table 4.10b

Correlations and descriptive statistics of RAKIT 1992/1993 (twin Sample 2)

	EX	VM	DI	LN	HF	IP
EX		.35	.32	.19	.29	.15
VM			.10	.46	.26	.20
DI				.07	.18	.14
LN					.24	.26
HF						.10
	EX	VM	DI	LN	HF	IP
M 1992-2	15.68	15.76	14.34	16.68	16.37	15.13
SD 1992-2	4.21	4.48	4.65	4.66	4.37	4.09
effect size-2	0.13	0.12	-0.12	0.34	0.20	0.04

Second sample (N = 207)

Table 4.11

Fit indices test for factorial invariance of RAKIT 1982 – 1993/1994

1st sample										
Model	Equality constraints	χ^2	DF	Compare	$\Delta\chi^2$	ΔDF	RMSEA	CFI	AIC	CAIC
1	-	23.2	16				0.043	0.988	98	289
2	Λ	30.0	20	2 vs 1	6.8	4	0.046	0.983	97	268
3	Λ & Θ	47.4	26	3 vs 2	17.3	6	0.062	0.961	102	243
4a	Λ & Θ & τ	68.5	30	4a vs 3	21.2	4	0.078	0.929	116	236
4b	Λ & τ	50.7	24	4b vs 2	20.7	4	0.072	0.952	109	260
2nd sample										
Model	Equality constraints	χ^2	DF	Compare	$\Delta\chi^2$	ΔDF	RMSEA	CFI	AIC	CAIC
1	-	25.1	16				0.052	0.982	101	292
2	Λ	30.1	20	2 vs 1	5.0	4	0.049	0.979	98	269
3	Λ & Θ	38.7	26	3 vs 2	8.6	6	0.049	0.973	95	236
4a	Λ & Θ & τ	54.2	30	4a vs 3	15.6	4	0.064	0.947	104	224
4b	Λ & τ	45.5	24	4b vs 2	15.4	4	0.068	0.953	107	257

Fit indices of the various models for both twin-samples are reported in Table 4.11. The first model (i.e., configural invariance) fits well in both the comparison containing twin Sample 1 and the comparison containing twin Sample 2. With the exceptions of a minor decrease in CFI values of both samples, and a small increase in RMSEA in the first twin sample, the fit indices of the metric invariance model (Model 2) indicate that the across-cohort restriction on factor-loadings (i.e., $\Lambda_1 = \Lambda_2$) is tenable. The restriction of invariant residual variances (Model 3; $\Theta_1 = \Theta_2$) is accompanied by some decrease in fit in twin sample 1: CFI, RMSEA and AIC of Model 3 are worse than those of Model 2 and the $\Delta\chi^2$ is rather large. In the second twin sample this restriction seems tenable, despite the small

drop in CFI value. However, a clear deterioration in fit in both twin samples is found when the factorial invariance models are fitted (Models 4a and 4b). In both samples the CAIC is the only fit index with a smaller value in these models as opposed to models 1 through 3. All other fit indices indicate that the restriction of invariant measurement intercepts (i.e., $\tau_1 = \tau_2$) is untenable. Again it appears that mean differences between both cohorts cannot be explained by latent (i.e., factor mean) differences in intelligence.

The rejection of factorial invariance (Models 4a and 4b) is caused mainly by the intercepts of the Discs and Learning Names subtests. That is, in both twin samples, these parameters have the largest modification index in Model 4b (DI: MI = 15 and LN: MI = 4 in twin sample 1; DI: MI = 7 and LN: MI = 8 in twin sample 2). Relaxing the equality constraints on these parameters, resulted in a partial strong factorial invariance model with the following fit indices: $\chi^2 = 31.43$, DF = 22, RMSEA = 0.042, CFI = 0.985, AIC = 94, CAIC = 255, and $\chi^2 = 31.09$, DF = 22, RMSEA = 0.045, CFI = 0.982, AIC = 95, CAIC = 256 in twin Samples 1 and 2, respectively. Thus, this partial strong factorial invariance model appears to have sufficient fit. A further look at the factor mean differences between the 1982 cohort and both 1992/1993 twin cohorts indicates that the factor means in the first twin sample are not significantly larger than those of the standardization sample: 0.69, SE = 0.43, Z = 1.61, p> .05 and 0.80, SE = 0.44, Z = 1.82, p> .05 for the nonverbal and the verbal factor, respectively. However, in the second twin sample the factor mean of the nonverbal factor is significantly higher than the standardization sample (0.85, SE = 0.42, Z = 2.04, p< .01), whereas the factor mean of the verbal factor in this second twin sample is not significantly higher (0.56, SE = 0.46, Z = 1.22, p > .05) than the corresponding factor mean of the 1982 cohort.

Again, we conclude that factorial invariance with respect to cohort is rejected. Hence, mean gains on the RAKIT between the 1982 and the 1992/1993 cohorts could not be explained fully by latent (i.e., factor mean) differences in intelligence. Only in the second twin sample a small part of the gains can be explained by a significant latent gain in the abstract factor. Especially the decline in scores on the Discs subtest and the gain in scores on Learning Names subtest require further investigation.

4.7 Study 5: Estonian Children 1934/1936 and 1997/1998: National Intelligence Test

Samples

The data from this last comparison stems from Olev Must and colleagues (Must et al., 2003), who compared two Estonian datasets covering a period of 60 years, from 1934/1936 to 1997/1998. The two cohorts contain 12- to 14-year-old schoolchildren who completed the Estonian National Intelligence Test. Must et al. (2003) found gains on most of the subtests, which were not consistent with a "Jensen effect". It is interesting to submit these Estonian data to the MGCFA approach since MGCFA has been found to lead to different conclusions then those found with Jensen's method of correlated vectors (e.g., Dolan, 2000; Dolan & Hamaker, 2001; Dolan et al., 2004). In addition, MGCFA can pinpoint subtests that manifest the gains in this Estonian data set. For the analyses we have pooled both age groups, we thus have 307 and 381 cases in the thirties and nineties

cohorts, respectively. For further information on the samples the reader is referred to Must et al. (2003).

Measures

The Estonian version of the National Intelligence Test is a group-administered intelligence test containing 10 subtests: Arithmetic (AR), Computation (CT), Sentence Completion (SC), Information (IN), Concepts (CC), Vocabulary (VO), Synonyms-Antonyms (SA), Analogies (AN), Symbol-Number (SN), and Comparisons (CP) (c.f., Must et al., 2003). These subtests are described shortly in Appendix E.

In order to obtain a reasonable factor structure, we have conducted exploratory factor analyses on both cohorts, using promax rotation. This resulted in an oblique two-factor model with factors denoted abstract (AR, CT, AN, SN, and CP) and verbal (AR, SC, IN, CC, VO, SA, AN). This model is used in the fitting of the subsequent models.

Results and Discussion

Table 4.12 provides the subtest correlations, as well as the means and standard deviations of both cohorts, computed by pooling the data over both age groups. As can be seen by the effect sizes, highest increase is found on the symbol number subtest. Counter to the expected Flynn Effect, four subtests show a decline, namely: Arithmetic, Computation, Vocabulary, and especially Information. Since this decline may also be due to a decrease in the latent factor(s), we proceed with the analyses.

Table 4.12

Correlations and descriptive statistics of National Intelligence Test 1934/1936 – 1997/1998

	AR	CT	SC	IN	CC	VO	SA	AN	SN	CP
AR		.41	.49	.48	.23	.40	.38	.45	.23	.24
CT	.49		.36	.48	.27	.46	.35	.53	.34	.48
SC	.65	.43		.60	.44	.53	.47	.50	.25	.30
IN	.68	.48	.76		.47	.63	.41	.62	.26	.42
CC	.47	.32	.65	.61		.35	.34	.42	.31	.30
VO	.53	.40	.66	.73	.56		.39	.52	.27	.39
SA	.50	.34	.51	.55	.43	.46		.45	.31	.33
AN	.57	.48	.64	.67	.57	.58	.48		.31	.40
SN	.48	.44	.48	.52	.45	.40	.33	.53		.44
CP	.43	.40	.43	.53	.38	.43	.44	.49	.44	
	AR	CT	SC	IN	CC	VO	SA	AN	SN	CP
M 1934/1937	16.92	24.45	27.21	25.26	35.60	25.65	26.62	13.86	24.28	27.10
SD 1934/1937	4.47	5.27	6.45	6.70	8.27	5.51	12.92	5.78	6.63	8.42
M 1997/1998	14.53	22.26	29.83	19.20	39.14	24.84	29.52	17.28	30.04	33.00
SD 1997/1998	4.50	5.36	6.02	5.45	7.00	6.50	8.20	5.99	5.62	8.64
Effect size	-0.53	-0.42	0.41	-0.90	0.43	-0.15	0.22	0.59	0.87	0.71

Note. Correlations of 1934/1936 sample (N = 307) below diagonal and of 1997/1998 sample (N = 381) above diagonal. Effect sizes in 1934/1946 SD units.

Table 4.13
Fit indices test for factorial invariance of NIT 1934/1936 – 1997/1998

Model	Equality constraints	χ^2	DF	Compare	$\Delta\chi^2$	ΔDF	RMSEA	CFI	AIC	CAIC
1	-	150.7	64				0.063	0.987	282	648
2	Λ	209.6	74	2 vs 1	58.9	10	0.074	0.978	324	634
3	Λ & Θ	316.2	84	3 vs 2	106.5	10	0.088	0.964	400	655
4a	Λ & Θ & τ	1147.5	92	4a vs 3	831.3	8	0.185	0.831	1250	1460
4b	Λ & τ	1029.1	82	4b vs 2	819.5	8	0.183	0.853	1120	1386

Table 4.13 provides the fit indices of the various factor models. The baseline model (Model 1; configural invariance) fits sufficiently as judged by the CFI, although RMSEA is somewhat on the high side. Moreover, it is apparent that the metric invariance model (Model 2) fits worse than the configural invariance model. All fit measures but the CAIC show deteriorating fit. Therefore, factor loadings cannot be considered cohort-invariant (i.e., $\Lambda_1 \neq \Lambda_2$). Note that this is in stark contrast with the high congruence coefficient of the first principal component found by Must, et al. (2003). This is due to the different natures of principal component analysis (PCA) and confirmatory factor analysis. PCA is an exploratory analysis that does not involve explicit hypothesis testing as is the case with MGCFA. In addition, the congruence coefficient has been criticized for sometimes giving unjustifiably high values (Davenport, 1990). The rejection of the metric invariance model is caused by several subtests, but most clearly by Vocabulary (MI = 20) and Symbol-Number (MI = 18). The failure of metric invariance is probably the worst possible outcome, as it implies non-uniform bias with respect to cohorts (Lubke et al., 2003a). Consequently, we present the next steps for illustrative reasons only. In fitting Model 3 ($\Theta_1 = \Theta_2$) the fit deteriorated still further. The fit indices of the factorial invariance models (4a and 4b) all indicate a clear deterioration in fit. Clearly the measurement intercepts are not invariant over cohorts (i.e., $\tau_1 \neq \tau_2$). The latter is primarily caused by the Information subtest. Because of the large number of parameters that show large modification indices in all non-fitting invariance models, we do not attempt to fit a partial factorial invariance model. The conclusion regarding the Estonian comparison is clearly that factorial invariance does not hold, and that the gains (either increases or decreases) found could not be explained by latent (i.e., factor mean) differences between the cohorts. Overall, the greatest modification index is found with the intercept of the Information subtest.

Again, factorial invariance between cohorts most clearly fails at the intercept level. This result is in line with the results from the Jensen test conducted by Must et al. (2003). The most notable difference between the analyses in that study and ours is the finding concerning the factor structure.

4.8 General Discussion

The present aim was to determine whether observed between-cohort differences are attributable to mean differences on the common factors that the intelligence tests are supposed to measure. Stated otherwise, we wished to establish whether the Flynn Effect is

characterized by factorial invariance. To this end, we conducted five studies comprising a broad array of intelligence tests and samples. The results of the MGCFAs indicated that the present intelligence tests are not factorially invariant with respect to cohort. This implies that the gains in intelligence test scores are not simply manifestations of increases in the constructs that the tests purport to measure (i.e., the common factors). Generally we found that the introduction of equal intercept terms ($\tau_1 = \tau_2$; Models 4a and 4b, see Table 4.1) resulted in appreciable decreases in goodness of fit. This is interpreted to mean that the intelligence tests display uniform measurement bias (e.g., Mellenbergh, 1989) with respect to cohort. The content of the subtests, which display uniform bias, differs from test to test. On most biased subtests, the scores in the recent cohort exceeded those expected on basis of the common factor means. This means that increases on these subtests were too large to be accounted for by common factor gains. This applies to the Similarities and Comprehension subtests of the WAIS, the Geometric Figures Test of the BPP, and the Learning Names subtest of the RAKIT. However, some subtests showed bias in the opposite direction, with lower scores in the second cohorts than would be expected from common factor means. This applies to the DAT subtests Arithmetic and Vocabulary, the Discs subtest of the RAKIT, and several subtests of the Estonian NIT. Although some of these subtests rely heavily on learned content (e.g., Information subtest), the Discs subtest does not.

Once we accommodated the biased subtests, we found that in four of the five studies the partial factorial invariance models fitted reasonably well. The common factors mean differences between cohorts in these four analyses were quite diverse. In the WAIS, all common factors displayed an increase in mean. In the RAKIT, it was the nonverbal factor that showed gain. In the DAT, the verbal common factor displayed the greatest gain. However, the verbal factor of the RAKIT, and the abstract factor of the DAT showed no clear gains. In the BPP, the single common factor, which presumably would be called a (possibly poor) measure of *g* showed some gain. Also in the second order factor model fit to the WAIS, the second order factor (again presumably a measure of g) showed gains. However in this model, results indicated that the first order perceptual organization factor also contributed to the mean differences.

It could be argued that the current results depend to a large extend to the choice of factor models. We put considerable effort in finding the best fitting models as the baseline models. In addition, we have tested for factorial invariance using alternative models, and found similar results to those reported here. Nevertheless, the interested reader is invited to replicate results with other factor models. The samples used in the studies differ substantively in size, resulting in differences in power to reject across-cohort equality constraints. However, we considered several fit measures that differ in their sensitiveness to sample size. Since those fit measures show a similar pattern, differences in statistical power, although important, do not seem to be a critical issue.

Here we investigated factorial invariance at the subscale level. Measurement invariance can also be investigated at the item level. Flieller (1988) compared two cohorts of French eight-year-olds that were administered the "Gille Mosaïque Test" in 1944 and 1984. Using a Rasch model to describe item responses in both cohorts, Flieller (1988) found that two-thirds of the 64 items were biased with respect to cohort. That is, the

majority of item parameters (i.e., item difficulty of the logistic item response function) in the 1984 cohort differed from the item parameters in the 1944 cohort. This uniform bias explained a large part of the test score increase on this Binet-type test over this 40-year period (Flieller, 1988). Thus, like we did in the analysis of subtest scores, Flieller, in an analysis of item scores, detected uniform measurement bias with respect to cohort.

With MGCFA it is possible to identify the subtests that display measurement bias. Similarly, by means of analyses based on item response theory (IRT), such as Rasch modeling, one can identify the individual items that are biased with respect to cohort (Flieller, 1988). Knowing which subtests or items are biased enables one to formulate testable hypothesis regarding the causes of the bias. Lubke et al. (2003a) have discussed how covariates can be incorporated in a multi-group factor model to investigate the sources of measurement bias. To do this, however, one has to identify covariates or "nuisance variables" (Millsap & Everson, 1993) that can account for the bias. At the item-level several approaches also have been proposed (Mellenbergh & Kok, 1991), such as correlational research, quasi-experimental research, and experimental research. Research on the effects of video games on intelligence test performance as described by Greenfield (1998) could be seen as an example of the latter.

Generally speaking, there are a number of psychometric tools that may be used to distinguish true latent differences from bias. It is notable that with the exception of Flieller (1988), little effort has been spent to establish measurement invariance (or bias) using appropriate statistical modeling. The issue whether the Flynn Effect is caused by measurement artifacts (e.g., Brand, 1987; Rodgers, 1998), or by cultural bias (e.g., Greenfield, 1998) may be addressed using methods that can detect measurement bias, and with which it is possible to test specific hypothesis from a modeling perspective. Consider the famous Brand hypothesis (Brand, 1987; Brand et al., 1989), that test taking strategies have affected scores on intelligence tests. Suppose that subjects nowadays more readily resort to guessing than subjects in earlier times, and that this strategy results in higher scores on multiple-choice tests. A three-parameter logistic model that describes item responses is perfectly capable of investigating this hypothesis, since this model has a guessing parameter (i.e., lower asymptote in the item response function) that is meant to accommodate guessing. Changes in this guessing parameter due to evolving test taking strategies would lead to the rejection of measurement invariance between cohorts. Currently available statistical modeling is perfectly capable of testing such hypotheses.

MGCFA is greatly preferred above the method of correlated vectors. In view of its established lack in specificity (Dolan et al., 2004; Lubke et al., 2001) it is not surprising that the method of correlated vectors gives contradictory results when it is applied to the Flynn Effect (Colom et al., 2001; Flynn, 1999b; Must et al., 2003). For instance, following Jensen's method, we computed the correlations between the g-loadings and the standardized increases in subtest means in the Dutch WAIS and RAKIT data. This resulted in correlations of 0.60 (WAIS data) and 0.58 (RAKIT data). We know that in both datasets factorial invariance is not tenable. Yet correlations of about 0.60 are invariably interpreted in support of the importance of g. For instance, the repeated application of the correlated vectors method to Black-White differences in intelligence test scores are resulted in a mean correlation of about 0.60 (Jensen, 1998).

The recent applications of method of correlated vectors to intelligence score gains (e.g., Colom et al., 2001; Flynn, 2000b; Must et al., 2003) followed Flynn's critique on the conclusions that Jensen and particularly Rushton (2000a) based on this method (Flynn, 1999c, 2000a, 2000b). From its beginning the Flynn Effect has been regarded to have large implications for the comparison of these B-W differences (e.g., Flynn, 1987, 1999c). Since the current approach (MGCFA) was previously applied in US B-W comparisons, we have the opportunity to compare those B-W analyses to the current analyses of different cohorts. Here we use results from Dolan (2000) and Dolan and Hamaker (2001), who investigated the nature of racial differences on the WISC-R and the K-ABC scales. We standardized the AIC-values of the Models 1 to 4a within each of the seven data sets, in order to compare the results of tests of factorial invariance on the Flynn Effects and the racial groups. These standardized AIC values are reported in Figure 4.2.

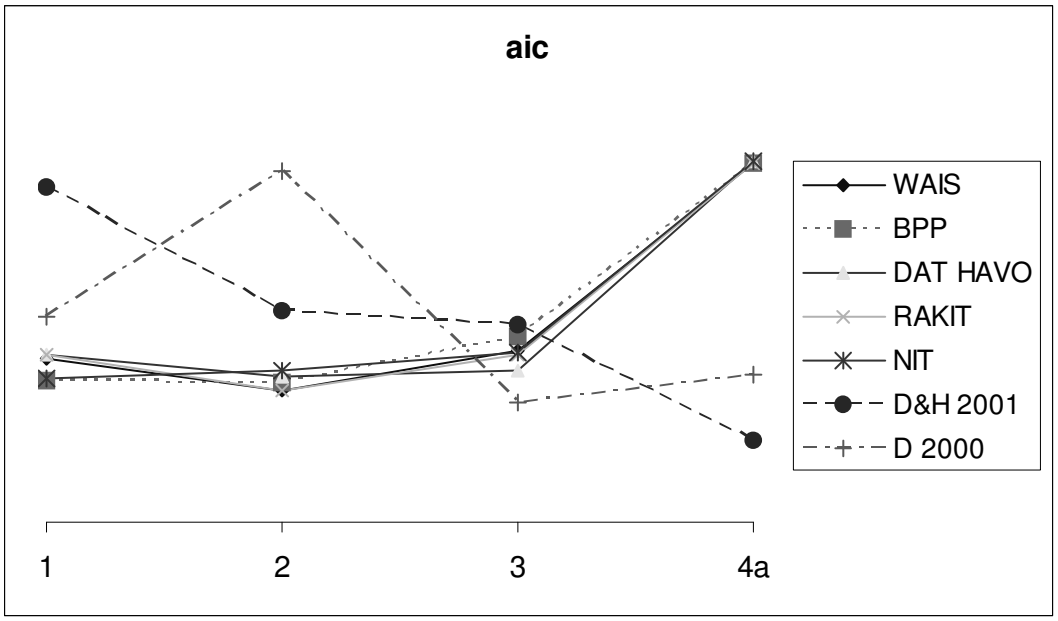

Figure 4.2 *Plot of standardized AIC values of data sets by stepwise models to achieve strict factorial invariance.*

As can be seen, the relative AIC-values of the five Flynn comparisons show a strikingly similar pattern. In these cohort comparisons, models one and two have approximately similar standardized AICs, which indicates that equality of factor loadings is generally tenable. A small increase is seen in the third step, which indicates that residual variances are not always equal over cohorts. However, a large increase in AICs is seen in the step to Model 4a, the model in which measurement intercepts are cohort-invariant (i.e., the strict factorial invariance model). The two lines representing the standardized AICs from both B-W studies clearly do not fit this pattern. More importantly, in both B-W studies it is concluded that measurement invariance between Blacks and Whites is tenable, since the lowest AIC values are found with the factorial invariance models (Dolan, 2000; Dolan & Hamaker, 2001). This clearly contrasts with our current findings on the Flynn

Effect. It appears therefore that the nature of the Flynn Effect is qualitatively different from the nature of Black-White differences in the US. Each comparison of groups should be investigated separately. IQ gaps between cohorts do not teach us anything about IQ gaps between contemporary groups, except that each IQ gap should not be confused with real (i.e., latent) differences in intelligence. Only after a proper analysis of measurement invariance of these IQ gaps is conducted, can anything be concluded concerning true differences between groups.

Whereas implications of the Flynn Effect for B-W differences appear small, the implications for intelligence testing in general are large. That is, the Flynn Effect implies that test norms become obsolete quite quickly (Flynn, 1987). More importantly however, the rejection of factorial invariance within a time period of only a decade implies that even subtest score *interpretations* become obsolete. Differential gains resulting in measurement bias for example imply that an overall test score (i.e., IQ) changes in composition. The effects on validity of intelligence tests are unknown, but one can easily imagine that the factors that cause bias over the years also influence within-cohort differences. Further research on the causes of the artifactual gains is clearly needed.

The overall conclusion of the present chapter is that factorial invariance with respect to cohorts is not tenable. Clearly this finding requires replication in other datasets. However, if this finding proves to be consistent, it should have implications for explanations of the Flynn Effect. The fact that the gains cannot be explained solely by increases at the level of the latent variables (common factors), which IQ tests purport to measure, should not sit well with explanations that appeal solely to changes at the level of the latent variables.

4.9 Appendix A: Description of the WAIS Subtests

Source: Wechsler (1955; 2000) & Stinissen et al. (1970)

Information (INF) contains 22 open-ended questions measuring general knowledge concerning events, objects, people and place names.
Comprehension (COM) contains 14 daily-life or societal problems, that the subject has to understand, explain, or solve. For this the subject needs to comprehend social rules and concepts.
Arithmetic (ARI) contains 16 arithmetic items that the subject has to solve without the use of paper and pencil.
Similarities (SIM) contains 13 word pairs about daily objects and concepts. The subject has to explain the similarities of the words.
Digit Span (DSP) contains 14 series of digits that subjects has to recall verbally forwards (12 items) or backwards (2 items).
Vocabulary (VOC) contains 30 words of which the subject has to give the meaning.
Digit Symbol (DSY) contains 115 items containing pairs of numbers and symbols. The subject uses a key to write down the symbol related to a number.
Picture Completion (PCO) contains 20 incomplete pictures of everyday events and objects about which the subject has to name the missing parts.

Block Design (BDE) contains 13 two-dimensional geometric figures that the subject has to copy by arranging two-colored blocks.
Picture Arrangement (PAR) contains 10 items in which pictures have to be arranged in a logical order.
Object Assembly (OAS) contains 5 puzzles of everyday objects that the subject has to assemble.

4.10 Appendix B: Description of the Subtests from Børge Prien's Prøve

Source: Teasdale & Owen (1987; 1989; 2000)

Letter Matrices (LEM) contains 19 items (15 min) in a 3 x 3 matrix format, with cells containing series of letters conforming to a pattern. The subject has to give the letter series that conforms to this pattern.
Verbal Analogies Test (VAT) contains 24 verbal analogies that the subject has to complement (5 min). The answers have to be chosen from a two lists of 100 possible responses.
Number Series Test (NST) contains 17 series of four numbers, that the subject has to complement (15 min).
Geometric Figures Test (GFT) contains 18 items (10 min) with complex geometric figures that have to be composed by five simple figures.

4.11 Appendix C: Description of the DAT'83 Subtests

Source: Evers and Lucassen (1992)

Vocabulary (VO) contains 75 items (20 min) in which out five words the respondent has to choose the word with the same meaning as the target word, measures lexical knowledge.

Spelling (SP) contains 100 words (20 min) of which the respondent has to judge the correctness of spelling, measures spelling ability.

Language Use (LU) contains 60 sentences (25 min) in which the respondent has to look for grammatical errors, measures grammatical sensitivity.

Verbal Reasoning (VR) contains 50 verbal analogies (20 min) that the respondent has to complement, measures lexical knowledge and inductive ability.

Abstract Reasoning (AR) contains 50 items (25 min) containing series of four diagrams. The respondent has to choose the diagram that logically follows these series. Measures inductive ability.

Space Relations (SR) contains 60 items (25 min) in which the respondent has to imagine unfolding and rotating objects, measures visualization.

Numerical Ability (NA) contains 40 arithmetic problems (25 min) that the respondent has to solve, measures quantitative reasoning.

4.12 Appendix D: Description of the RAKIT Subtests

Source: Bleichrodt, et al. (1984)

Exclusion (EX) contains 30 items in which the child has to choose one out of four figures that is deviant. This subtest measures inductive reasoning.
Discs (DI) contains 12 items in which the child has to put discs with holes on sticks. This subtest measures spatial orientation and speed of spatial visualization.
Hidden Figures (HF) contains 30 items in which the child has to recognize two concrete figures in a complex drawing. This subtest measures transformation of a visual field.
Verbal Meaning (VM) contains 40 words, which meaning the child has to denote by pointing out one out of four pictures. This subtest measures passive verbal learning.
Learning Names (LN) contains 10 pictures of animals whose names the child has to learn. This subtest measures active learning.
Idea Production (IP) contains 5 items in which the child has to produce names of objects and situations that belong to a broadly described category. This subtest measures verbal fluency.

4.13 Appendix E: Description of the National Intelligence Test Subtests

Source: Must et al. (2003)

Arithmetic (AR) contains 16 arithmetic problems that require a solution for an unknown quantity.
Computation (CT) contains 22 items requiring addition, subtraction, multiplication, and division of both integers and fractions.
Sentence Completion (SC) contains 20 items requiring filling in missing words to make sentences understandable and correct.
Information (IN) contains 40 items about general knowledge.
Concepts (CC) contains 24 items requiring selecting two characteristic features from among those given.
Vocabulary (VO) contains 40 items requiring knowledge about the qualities of different objects.
Synonyms-Antonyms (SA) contains 40 items requiring evaluation of whether the words presented mean the same or opposite.
Analogies (AN) contains 32 items requiring transferring the relation between two given words to other presented words.
Symbol-Number (SN) contains 120 items in which the correct digit must be assigned to a presented symbol from a key.
Comparisons (CP) contains 50 items requiring same or different judgments about sets of numbers, family names, and graphic symbols presented in two columns.

5

The dark past, obscure present, and bright future of African IQ

On the basis of extensive reviews of the literature, Lynn concluded that average IQ of the Black population of sub-Saharan Africa lies below 70. In this chapter, the authors evaluate published empirical data on this issue. Focus is on average scores of African samples on Raven's Standard Progressive Matrices (SPM), Coloured Progressive Matrices, Goodenough-Harris Draw-a-Man test, and several other IQ tests. Validity of IQ tests in African samples is evaluated critically. Because of a general lack of rigorous measurement invariance studies, it is uncertain to what degree IQ scores in Africa reflect levels of general intelligence. Results show that average IQ in Africa lies somewhere around 80 when compared to US norms, and that SPM scores among African adults have shown a secular increase over the years. Variables representing health, fertility, nutrition, educational attainment, modernization, and urbanization are shown to correlate highly with national IQ over the world. It is concluded that the Flynn Effect is in its infancy in Africa. Implications for genetic theories of race differences in intelligence are discussed.

5.1 Introduction

On the basis of several extensive reviews of the literature, Lynn concluded that the average IQ of the Black population of sub-Saharan Africa lies below 70 (Lynn, 1978, 1991, 1997, 2003, 2006; Lynn & Vanhanen, 2002; cf. Rushton & Jensen, 2005a). In a critique on Lynn's 1978 and 1991 reviews, Kamin (1995) accused Lynn of distortions and misrepresentations of data, which, according to Kamin, constituted "a truly venomous racism" (p. 86). Lynn (2006, p. 244), in turn, accused anyone who might disagree with his review of IQ in Africa of ignorance and/or political correctness (cf. Rushton, 1996). Clearly, the topic of IQ of Africans is highly controversial.

Ad hominem arguments, poor research (followed by simplistic conclusions), or shying away from this subject (for whatever reason), will certainly not advance our understanding. We view the study of group differences in IQ test scores as a valid scientific undertaking, regardless of the nature of the groups. Our understanding of ethnic or racial group differences depends on rigorous and careful research. The aim of the present chapter is to present a balanced and critical evaluation of the present body of results of IQ testing in Africa. The specific aims of our study are threefold. First, we want to arrive at an estimate of the average performance of the Black population of sub-Saharan Africa (henceforth Africans) on three non-verbal tests of general cognitive ability. We express the

average test performance in terms of IQ based on western norms because we want to compare our estimate of average test performance to the only presently available results, namely Lynn's. The reader should be aware, however, that observed IQ scores do *not* equal particular levels of general intelligence or *g*. Whether or not these observed test scores actually reflect relative positions on the latent dimension of *g*, depends on many conditions, which relate to our second aim. The second aim is to arrive at a better understanding of the meaning of IQ test scores in Africa by focusing on the validity and the psychometric properties of western IQ tests when applied to Africans. The notion of IQ testing in Africa seems to elicit either a knee jerk rejection of the possibility of obtaining a valid measure (e.g., Berry, 1974), or a blithe acceptance of this possibility (e.g., Herrnstein & Murray, 1994; Lynn, 2006). In our review, we attempt to determine whether the conditions for good psychometric measurement have been met in African studies. The third aim of our study is to evaluate the environmental correlates of mean IQ test scores, which are often proffered in support of some causal interpretation of mean group differences in IQ. Here we concentrate on environmental variables that are suspected to have caused gains in IQ levels of western populations over the years (i.e., the Flynn Effect; Flynn, 2006; Neisser, 1998). This part of our review suggests that it is very difficult to arrive at rigorous causal claims concerning the nature of group differences in mean IQ test scores. However, when viewed in the light of common explanations of the Flynn Effect (e.g., Barber, 2005; Blair, Gamson, Thorne, & Baker, 2005; Ceci, 1991; Lynn, 1990; W. M. Williams, 1998; Zajonc & Mullally, 1997), there is reason to be optimistic about the future of average IQ in sub-Saharan Africa.

5.2 Is Average IQ in Africa Really Below 70?

To estimate average IQ of countries or racial groups all over the world, Lynn draws mainly on published data from cognitive ability tests such as Raven's Coloured Progressive Matrices (CPM; J. C. Raven, 1956) or the Standard Progressive Matrices (SPM; J. C. Raven, 1960). These tests are generally considered to be excellent non-verbal indicators of general intelligence or *g* (Carroll, 1993; Jensen, 1998), and have been administered often in Africa. For instance, Fahrmeier (1975) collected CPM data of schooled and unschooled Nigerian children. Lynn compared their CPM scores to British norms,[30] which resulted in an average IQ of about 69 (Lynn, 2006; Lynn & Vanhanen, 2002). In another study conducted in Nigeria, Wober (1969) administered the SPM twice to a group of male factory workers. Lynn compared their pretest scores to British norms and concluded that their average IQ was below 65. On the basis of these two convenience samples Lynn claims that average IQ

[30] Throughout this chapter, we assume that the work on IQ in Lynn and Vanhanen's book is by Lynn. The estimation of IQ is described as follows: "Around 1973, data for the Coloured Progressive Matrices for a sample of 375 6-13 year-olds were collected by Fahrmeier (1975). In relation to the 1979 British standardization of the Standard Progressive Matrices, the mean IQ is 70. Because of the 6-year interval between the two data collections, this needs to be reduced to 69" (Lynn & Vanhanen, 2002, p. 215). Lynn probably used a table provided on page 60 of the SPM manual (J. C. Raven, Court, & Raven, 1996) to convert raw CPM scores to raw SPM scores, to compare these CPM scores to British SPM norms of 1979. Note that his downward correction for outdated norms is an error because the norms are more recent than the test scores in Fahrmeier's sample. Hence, according to the appropriate use of this correction (i.e., 2 IQ points per decade), the IQ should have been raised by one point, not lowered by one.

in Nigeria is below 70 (Lynn & Vanhanen, 2002). Additional published data of over 50 samples from various sub-Saharan countries have led him to conclude that the average IQ of Africans is around 67 (Lynn, 1978, 1991, 1997, 2003, 2006; Lynn & Vanhanen, 2002; cf. Rushton & Jensen, 2005a). This low IQ level is rather implausible, because by western standards (cf. DSM-IV; American Psychiatric Association, 1994), it would imply that more than half of the African population suffers from mental retardation. This raises the question whether Lynn's estimate is accurate.

Several aspects of Lynn's work on African IQ have been criticized (Barnett & Williams, 2004; Dambrun & Taylor, 2005; Hunt & Sternberg, 2006; Kamin, 1995; Lane, 1994), although none of Lynn's critics have brought new data to bear on the issue. One point of critique is that Lynn's estimate of average IQ among Africans is primarily based on convenience samples, and not on samples carefully selected to be representative of a particular population (Barnett & Williams, 2004; Hunt & Sternberg, 2006). For example, the samples of Fahrmeier (N = 375) and Wober (N = 86) neither were intended to be, nor could be considered to be representative of the entire population of Nigeria, a country with over 130 million inhabitants. Moreover, despite his objective of providing a "fully comprehensive review [...] of the evidence on […] differences in intelligence worldwide" (Lynn, 2006, p. 2), in his review of IQ in Africa, Lynn does not consider a sizeable portion of the literature. For instance, Lynn did not consider several studies with the SPM in Nigeria (Maqsud, 1980a, 1980b; Okunrotifa, 1976) that clearly indicated that average IQ in this country is considerably higher than 70. In the current study, we tried to locate additional published data of western IQ tests that are most commonly used throughout Africa, namely the SPM, CPM, and the Goodenough-Harris Draw-a-Man test (DAM; Goodenough, 1926; Harris, 1963). In addition, we review and discuss all sources of data given by Lynn in his two latest books (Lynn, 2006; Lynn & Vanhanen, 2002). These additional IQ data are based on the Kaufman-Assessment Battery for Children (Kaufman & Kaufman, 1983), the Wechsler scales (Wechsler, 1974, 1981), and several other IQ tests.

Implications

Lynn's work on African IQ is often taken at face value (e.g., Abdel-Khalek & Raven, 2006; Campbell, 1996; Herrnstein & Murray, 1994; Kanazawa, 2004; Miller, 1992; Reeve & Hakel, 2002; Rindermann, 2006; Rushton & Skuy, 2000; Rushton, Skuy, & Bons, 2004; Rushton, Skuy, & Fridjhon, 2002, 2003; Sarich & Miele, 2004; Skuy et al., 2002; Te Nijenhuis, De Jong, Evers, & van der Flier, 2004; Teasdale & Owen, 2005), even by his critics (e.g., MacEachern, 2006). Moreover, Lynn's estimates of national IQ are used as data in several studies (Barber, 2005; Dickerson, 2006; Jones & Schneider, 2006; Kirkcaldy, Furnham, & Siefen, 2004; Meisenberg, 2004; Morse, 2006; Templer & Arikawa, 2006; Voracek, 2004; Weede & Kampf, 2002; Whetzel & McDaniel, 2006), which were mainly concerned with predicting national differences in economic development. In addition, Lynn's reviews of low average IQ in sub-Saharan Africa are accorded a central role in theories, which state that race differences in intelligence test scores have a substantial genetic component (Jensen, 1998; Levin, 1997; Lynn, 2006; Miller, 1995; Rushton, 2000b; Rushton & Jensen, 2005a; Templer & Arikawa, 2006).

The essence of these theories is that lower intelligence test scores of Africans and African Americans compared to people of European or Asian descent have evolutionary

causes (Lynn, 2006; Rushton, 2000b; Rushton & Jensen, 2005a). These theories further state that African Americans have a certain degree of genetic European-African admixture, which should raise their intelligence levels as compared to Africans (Lynn, 1991; Rushton & Jensen, 2005b). The implicit assumptions underlying this reasoning (Loehlin, 2000) are: (1) that Africans and people from European descent differ in the frequencies of genes affecting intelligence, (2) that these genes act in an additive fashion, and (3) that the African and European gene pools among African Americans are representative of the ancestral African and European gene pools, respectively. If this is the case, it follows that African Americans should have considerably higher IQ levels than Africans. Average IQ of African Americans is usually estimated to be 85 (Gottfredson, 2005; Jensen, 1998; Rushton & Jensen, 2005a) or somewhat higher (Dickens & Flynn, 2006). On the basis of his genetic theory of race differences in intelligence and the genetic Black-White admixture of African Americans, Lynn asserts that if the environmental circumstances of Africans would be as good as those of African Americans, average IQ of Africans should be around 80. He states that adverse environmental circumstances in Africa depress African IQ levels considerably below 80[31] (Lynn, 2006, p. 71). Thus, our estimate of average IQ in Africa provides as an empirical test of Lynn's theory.

The genetic or evolutionary theories of race differences in intelligence presuppose that IQ test scores are valid indicators of general intelligence throughout the world. The question arises whether these scores are valid and comparable to scores in western samples in terms of general intelligence (Barnett & Williams, 2004; Ervik, 2003; Hunt & Sternberg, 2006; Lane, 1994).

5.3 Measurement Problems and Psychometric Comparability

A person's IQ score and a person's level of latent general intelligence or *g* can not simply be equated for the simple reason that IQ tests are fallible instruments. Often in Africa, IQ tests are not administered in conditions resembling those in developed countries. For instance, in Fahrmeier's study with the CPM in Nigeria, "children were tested on porches, in entrance rooms, or under trees" (Fahrmeier, 1975, p. 282) by *untrained* personnel. This does not compare very well with the official guidelines as formulated in the test manual: "The person to be tested is seated comfortably opposite the psychologist at a table about 2 feet wide" (J. C. Raven, 1956, p. 13). Often more than not, test administration in Africa occurs on the ground, on veranda's, under trees, or in overcrowded and sparsely furnished classrooms (e.g., Berry, 1983; Fahrmeier, 1975; Hunkin, 1950). Such non-standard test settings, combined perhaps with harsh climatic circumstances (cf. Sternberg, 2004), are likely to depress performance.

Moreover, the claim that non-verbal IQ tests are "devoid of cultural content" (e.g., Templer & Arikawa, 2006, p. 122) does not sit very well with the following measurement problems. Several items in the CPM and SPM contain geometric shapes which have no names in many African languages (Bakare, 1972). It is not uncommon in (rural) Africa to

[31] Based on his estimate of an average IQ in Africa, Lynn asserts that adverse environmental circumstances lower average IQ in Africa by 13 points.

come across test takers who are unfamiliar with color-printed material (Giordani, Boivin, Opel, Dia Nseyila, & Lauer, 1996), or who are inexperienced with using a pencil (Badri, 1965b). Giving such test takers a paper-and-pencil test with unknown colored geometric shapes (e.g., CPM) is not likely to produce test scores that accurately measure general intelligence. Unfamiliarity with the stimulus material in western IQ tests are only the tip of the iceberg of the many possible cultural effects that may affect performance of African test takers when diagrammatic non-verbal intelligence tests such as the CPM or SPM are used to assess general cognitive ability. Other IQ tests include similar or alternative formats that may be equally unfamiliar to African test takers. For instance, for some African children photographs are an entirely new phenomenon (Fahmy, 1964). Besides, Africans may not think that acting fast represents intelligent behavior (Mpofu, 2004; Wober, 1974), the idea of responding to a multiple choice format may be entirely new to some African test takers (Irvine, 1966), and it cannot be assumed that a standard instruction enables test takers to fully comprehend what is expected from them (e.g., Kendall, Verster, & Von Mollendorf, 1988; MacArthur, Irvine, & Brimble, 1964). Such problems have led several authors to conclude that intelligence testing is strongly culturally determined (Berry, 1974, 1976; Greenfield, 1997; Irvine, 1969b; Nell, 2000; Sternberg, 2004). Clearly, measurement problems associated with IQ testing in Africa should not be ignored.

Others have claimed that IQ test scores are nevertheless comparable across cultures (Lynn, 2006; Rushton, 2000b). However, before one can interpret IQ test scores differences across individuals or groups in terms of some latent cognitive ability (e.g., *g*), several conditions have to be met. Necessary but *insufficient* conditions for such an interpretation concern reliability and validity of tests. That is, the test scores must show some level of consistency, either internally, or in repeated testing. In addition, the test scores should show merit in their correlation with other cognitive ability test scores (i.e., convergent validity). Ideally, structural equation modeling is employed to shed some light on factors involved in test performance. Test scores' validity may be substantiated by their prediction of criteria such as school grades (i.e., predictive validity). Predictive regression lines can be compared across groups, but these do not establish conclusively the absence of measurement bias (Millsap, 1997a). For a comparison across diverse cultural groups to be truly valid, tests and items should function equivalently in all groups of test takers to be compared. Specifically, tests and items should display measurement invariance with respect to groups (Mellenbergh, 1989; Millsap & Everson, 1993).

Measurement invariance across groups implies that the relation between test scores and latent traits, which are supposed to underlie those scores, is identical across groups. Measurement invariance can be tested by employing a measurement model in which this relation between test scores and latent trait(s) is explicitly modeled (Holland & Wainer, 1993; Meredith, 1993; Millsap & Everson, 1993). The relation between test scores and latent traits is central to the question of cross-cultural comparability of IQ test scores (e.g., Little, 1997; Poortinga & van der Flier, 1988). Within Item Response Theory (IRT) models, measurement bias is called Differential Item Functioning (DIF). DIF is said to be absent when, in a sufficiently restrictive measurement model (e.g., an unidimensional three parameter logistic item response model), measurement parameters linking ability to tests scores are approximately equal across groups (i.e., not-significantly different). The absence

of DIF (i.e., measurement invariance) provides strong support for the claim that the test score differences across groups reflect group differences on the latent trait that is supposed to underlie those scores. Put differently, if measurement invariance is supported, this implies that we are measuring the same thing in different groups. However, as long as measurement invariance has not been established, one cannot simply conclude that the measurement problems with IQ testing in Africa are irrelevant.

IQ scores may or may not reflect accurately levels of general intelligence or *g*. For instance, in factor analytic studies (Carroll, 1993) the SPM has often shown a high *g* loading (i.e., strong correlation with *g*), but that does not mean this test does not measure additional traits (e.g., Carpenter, Just, & Shell, 1990; Dillon, Pohlmann, & Lohman, 1981; Mackintosh & Bennett, 2005; van der Ven & Ellis, 2000). If groups differ on such an additional trait besides *g*, a group difference in SPM scores does not solely reflect a group difference in *g*. In one of the few factor analyses in which SPM scores of Africans were factor-analyzed with additional cognitive ability tests, the SPM did appear in some samples to load on additional factors besides *g* (Irvine, 1969b). Therefore, group differences in SPM or CPM scores can not simply be interpreted as group differences in *g*. In our review, we consider psychometric properties, measurement invariance, and factorial nature of the SPM, CPM, and DAM tests in Africa.

Misunderstanding instructions, measurement bias, and suboptimal testing conditions may all lead to an underestimation of cognitive ability of IQ of Africans. Therefore, we exclude from our review of average IQ in Africa those samples in which such effects were obvious. However, measurement invariance studies involving Africa samples are very sparse, and not all data sources include sufficient information to establish whether testing conditions were acceptable (e.g., whether test takers understood the context and the instructions). Therefore, we stress the importance of care in interpreting the IQ scores in Africa, which we will provide below: IQ test scores and general intelligence are distinct entities (Bartholomew, 2004).

5.4 The Flynn Effect

Besides psychometric problems, there are several possible reasons that average IQ scores among Africans are often lower than average IQ scores in western populations. Lynn (2006) and Rushton and Jensen (2005a) have claimed that genes play an important role, while environmental circumstances are less important. However, the prenatal, postnatal, and childhood circumstances of many African children are not as good as those in the developed world (e.g., Mung'ala Odera, Snow, & Newton, 2004; Sigman, Neumann, Jansen, & Bwibo, 1989). Moreover, variables related to economic and social development are known to have a strong positive effect on average IQ scores. In the western world, average IQ scores have shown remarkable gains over the course of the twentieth century (Flynn, 1984, 1987, 2006). These gains have been largest for non-verbal tests once considered relatively unaffected by cultural factors. For instance, in The Netherlands an unaltered version of Raven's SPM test was administered to male military draftees from 1952 to 1982. The 1982 cohort scored approximately 20 IQ points higher than the 1952 cohort (Flynn, 1987). Proposed causes of this so-called Flynn Effect include gains in test

sophistication (Brand, 1987) and improvements in test specific skills (Greenfield, 1998; Wicherts et al., 2004). Other proposed causes are related to gains in latent cognitive ability, such as improvements in nutrition (Lynn, 1989, 1990), urbanization (Barber, 2005), improvements in health care (W. M. Williams, 1998), a trend towards smaller families (Zajonc & Mullally, 1997), increases in educational attainment (Ceci, 1991; Husén & Tuijnman, 1991; Tuddenham, 1948), improvement of educational practices (Blair et al., 2005), greater environmental complexity (Schooler, 1998), the working of gene by environment correlation in the increasing presence of more intelligent others (Dickens & Flynn, 2001), and the genetic effect of heterosis (Mingroni, 2004). Although there is some indication of a similar secular trend in IQ scores on the CPM in Kenya in recent years (Daley, Whaley, Sigman, Espinosa, & Neumann, 2003), little is known about the Flynn Effect in Africa. However, in developing countries south of the Sahara, most of the environmental variables assumed to be responsible for the Flynn Effect, have not been subject to the improvements that the developed world has enjoyed over the last century. As far as the data permit, we will also focus on possible secular trends in IQ test scores in Africa. In addition, in the last part of our study we relate estimates of national IQs of sub-Saharan African countries and other countries around the world to variables that have been proposed as causes for the Flynn Effect in the developed world. The results of this exercise may contribute to our understanding of the current status of African IQ levels and of the potential of the Flynn Effect in sub-Saharan Africa.

5.5 Average IQ in Africa

Our review focuses primarily on the SPM and CPM. These tests are commonly used in Africa, and Lynn's review of African IQ draws mainly on SPM and CPM data. In addition, the SPM and CPM are non-verbal tests that are often claimed to be the best indicators of *g*. According to Spearman's hypothesis (Jensen, 1998), which states that *g* is the main locus of mean differences in IQ scores, these tests should show the largest difference between African and European samples. We also consider the DAM test because it is used commonly in Africa. Because the DAM test is less highly *g* loaded (Jensen, 1980), Spearman's hypothesis implies that African IQ on the DAM should be higher than IQ on the CPM and SPM tests.

Method

Selection bias

It is well known that the use of convenience sampling may result in highly inaccurate estimates of the characteristics of a population. In his attempts to estimate average scores of the population of (countries in) sub-Saharan Africa, Lynn uses published studies, which often employed convenience sampling. For instance, Fahrmeier (1975) did not intend his sample to be representative for the entire population of Nigerian school-aged children. If so, he would not have sampled children solely from one of the many ethnic groups in Nigeria. More importantly, he would not have restricted his sampling scheme to children in one town in North-Nigeria, a part of Nigeria where primary school attendance was considerably below the national average in the 1970s (i.e., under 30% as

opposed to 71% nationwide; Maduagwu, 2003). Nevertheless, some samples used by Lynn to arrive at population estimates of average test scores in sub-Saharan Africa were in fact sampled carefully to be representative of a (sub)population of a particular country (e.g., Costenbader & Ngari, 2001; MacArthur et al., 1964). Unfortunately, Lynn ignores a sizeable portion of studies in which IQ tests were administered in Africa. Besides, he gives small unrepresentative samples as much weight as large representative ones in his estimates of average IQ of Africans. Because in most cases representative samples are much larger than convenience samples, one straightforward, albeit partial, solution to the issue of selection bias is to weight average IQ scores by sample size. An additional reason to do so, is that the effect of sampling variability decreases as sample size increases.

Note that our review of the literature and our estimates of average IQ are concerned with the overall Black population of sub-Saharan Africa. This represents a crude generalization that does not do justice to the wide cultural, social, and economic differences between the many peoples of sub-Saharan Africa. However, the data are generally insufficient to arrive at an acceptable estimate of average IQ per country or cultural group.

Selection of Studies

We did not limit our attention to studies identified by Lynn. Instead, we tried to locate studies in which the IQ tests most commonly used in Africa (SPM, CPM, DAM) were administered to samples of Africans. To this end, we used Psychinfo and a combination of various search terms. The search terms we used were "Raven", "IQ", "progressive matrices", "Draw", combined with the words "Africa", “African”, and the names of all countries in the continent (e.g., "Nigeria" or "Nigerian"). We located additional papers while scanning the reference lists of the papers we found. In addition, we collected in Web of Science all articles (from 1988 onwards) referring to the various manuals of the SPM, CPM, and the DAM. This resulted in about 2500 papers for the SPM/CPM and 300 papers for the DAM. The titles of all these papers were scanned for relevance. We used only books, papers, or reports that were available through the IBL system in the Netherlands, a system to which 400 Dutch libraries are connected. Although our approach resulted in a large sample of studies of African IQ, it is conceivable that we missed other studies.

The following criteria were employed in the selection of studies. First, the condition of administration of the tests should reasonably approximate those stipulated in the test manual. For instance, we excluded the SPM scores of Zindi's (1994a) sample of Zimbabwean school children, because the SPM was not administered in its entirety (i.e., only 36 of the 60 items were given), and because it is not clear how Zindi arrived at his IQ estimate of 70. We also disregarded Klingelhofer's large sample of secondary school students from Tanzania (average IQ of 78 according to Lynn), because Klingelhofer imposed a time limit on the SPM (a nonstandard condition). He did so to "[preclude] some of the kinds of comparisons that have marked the literature" (Klingelhofer, 1967, p. 206). Whenever tests were administered twice, we used the pretest scores. We did not assign IQ values to studies in which the SPM, CPM, or DAM did not meet basic psychometric standards, as will be discussed in the results section. For instance, the test-retest reliability in Wober’s sample of Nigerian factory workers was 0.59, i.e., lower than the 0.80 typically

found with the SPM (J. C. Raven et al., 1996), and the correlation between pretest SPM scores and educational attainment did not deviate significantly from zero (see also Wober, 1966). Therefore, we did not consider this sample in our estimation of average IQ.

We used only data sets from sources which included sufficient descriptive statistics. This excluded a large number of studies, in which raw means or percentile scores were not given. Whenever medians were given, we took the median as an estimate of the mean. In one source (Morakinyo, 1985), percentile scores were reported, and we translated these to approximate raw means to compare the scores to a more recent norm table. Unless stated otherwise, all IQs are standardized IQs normed in Great Britain (CPM, SPM) or the US (DAM).

The last criterion is concerned with norms. We excluded CPM data of age ranges for which no British norms exist. This criterion resulted in the exclusion of several studies in which the CPM was administered to adolescents and adults (Berlioz, 1955; Berry, 1966; Binnie Dawson, 1984; Boissiere, Knight, & Sabot, 1985; Kendall, 1976; Sternberg et al., 2001). Lynn assigns average IQs below 70 to these samples (cf. Herrnstein & Murray, 1994). However, there are no (British) CPM norms above the age of 11. Lynn (R. Lynn, personal communication, June 22, 2006) employs a table from the SPM manual (J. C. Raven et al., 1996) with which CPM scores can be converted to SPM scores (cf. Lynn, 1997). These approximate SPM scores can be compared to norms for adults, allowing a rough estimate of IQ. However, this method does not result in accurate estimates of IQ, because the CPM is too easy for healthy test-takers above the age of 11. This results in a problematic ceiling effect. Because of this ceiling effect, it is very hard to get an above-average SPM-norms IQ on the CPM. For instance, the only possible CPM score equivalent to an above average IQ for a twenty-year-old would be a perfect score on this 36-item test. That is, an CPM score of 36 is equivalent to an SPM score of 57 (J. C. Raven et al., 1996), which corresponds to an IQ of 115. Likewise, a CPM score of 34 corresponds to an IQ of 93. However, some unreliability would normally lead an above average intelligent adult to make a few mistakes on the CPM. With 4 errors, the adult's IQ score drops to 84. This effect virtually guarantees an underestimation of IQ with the CPM in samples above the age of 11, particularly in adults. The drawbacks of this conversion method are evidenced by the fact that a carefully selected norm sample of 894 normal healthy adults from Italy and San Marino (Measso et al., 1993) would have had an average IQ of 75 on the CPM according to this conversion method. Even in the subsample of those with at least 14 years of schooling (N = 89), average IQ based on this method would be as low as 84. Because this conversion method does not result in reasonable IQ estimates, we do not consider IQ scores based on CPM scores of adults and adolescents.

Converting Raw Scores to IQ

While IQ scores on the DAM are usually provided in the original source, the SPM and CPM raw scores need to be converted to percentile scores given in norm tables. These percentile score are then translated to IQ scores on the basis of a normal distribution with M = 100 and SD = 15. For the SPM scores collected during the 1950s and 1960s, we use British norms of 1938 for children and 1948 norms for adults (both of which are given in: J. C. Raven, 1960). For the data collected after 1965 (i.e., the midpoint of various

SPM/CPM standardizations), we used British SPM norms of the 1979 standardization for children (J. Raven, 2000) and the 1992 standardization for adults (J. C. Raven et al., 1996). Likewise, we used the 1956 British norms (J. C. Raven, 1956) of the CPM for samples in the 1950s and the first half of the 1960s. For samples beginning 1966, we employed the British CPM norms of 1982 (J. C. Raven, Court, & Raven, 1990). In contrast to Lynn's approach, we did not correct for the Flynn Effect. The primary reason is that the secular trend has not been documented in African countries, and so we cannot reasonably correct for an effect, which has not been established. In addition, we chose the year of 1965 as a midpoint, so the upward and downward corrections would be approximately[32] balanced.

It is important to note that Raven never intended to use the Progressive Matrices tests to be used as an IQ test. There are indeed several reasons for not converting Raven's scores to IQs. First, the Progressive Matrices are limited to a single test-format. If the test taker is unfamiliar with this format, or the stimulus material in it, intelligence will be underestimated. In addition, in comparison to IQ batteries, such as the Wechsler scales (Wechsler, 1974, 1997), the number of items in Raven's scales is quite small in the SPM (i.e., 60 items), and smaller still in the CPM (i.e., 36 items). An additional problem arises in the translation from SPM/CPM raw scores to IQ scores, particularly in the extreme score ranges. For instance, a raw SPM score of 9 for a 7 year-old equals the first percentile of the British 1979 norms of this age group. Given a normal distribution with a mean of 100 and an SD of 15, this first percentile is equivalent to an IQ of 65. However, suppose that by chance our 7-year-old were to guess one additional item correctly. This would raise the raw score to 10, which is equivalent to the fifth percentile in the 1979 SPM norms, which corresponds to an IQ of 75. In the extremes of the distribution, norm tables include large leaps, and a single item that functions differentially between groups (i.e., is biased) might mean a 10-point IQ effect.

In our calculation of IQs based on raw SPM and CPM scores, we tried to be as careful as possible. Because most of these norms tables do not give percentiles for all raw scores, some inter- and extrapolation was necessary to arrive at percentile scores. In the few cases in which a raw mean was below the 1st percentile, we assigned an IQ of 64 (similar to the approach employed by Lynn & Vanhanen, 2002). In studies where scores were reported for subsamples, we first estimated IQs for the separate groups, and then computed an *N*-weighted average of IQ scores. Whenever scores were given for a particular age range (e.g., 7-8 years), the average IQ was compared to the norms for the corresponding age groups (e.g., 7, 7½, 8, and 8½ year-olds). The average IQ was an average of these age-norms after weighing for sample size. If a data source indicated that the age distribution was peaked at a certain age, we adjusted our estimates accordingly. Our approach almost always resulted in an IQ estimate that was equal to the estimate that was based on the overall average raw score of the sample, when compared to the norm that corresponded to the average age of the entire sample. If not, we took the average IQ of both approaches. All steps in the

[32] Nevertheless, we did check whether the Flynn Effect correction made a difference. The correction used in our check is based on Lynn's approach in which British norms on the SPM and CPM should be lowered by 2 IQ points per decade when data is more recent than standardization, or 2 points per decade upwards whenever the data is older than the standardization (Lynn & Hampson, 1986). Similarly, for the DAM, we also follow Lynn's adjustments of 3 IQ points per decade.

estimation of IQ scores are available upon request by the first author. We note that in determining IQ, it is conceivable that the aggregation of raw scores from different test takers with varying ages does not necessarily match the average of IQ scores, when these are computed at the individual level. In sum, the assignment of IQ values for the SPM and the CPM is problematic, and the values we provide are only given in order to arrive at a rough estimate of average IQ that can be compared to the average IQs estimated by Lynn.

Results

The overview of studies into IQ of samples sub-Saharan African is now provided for each of the most commonly used tests separately. Next, we will focus on some of the additional data provided by Lynn to back up his claim that average IQ in Africa lies around 67. After that, we will discuss the issue of measurement invariance, and the data employed by Lynn to validate his estimates of national IQ.

Raven's Standard Progressive Matrices

IQ estimate. Table 5.1 gives average SPM scores and corresponding IQ scores for 38 samples in sub-Saharan Africa, totaling 13,880 cases, of which 8,808 cases (63%) were included in Lynn's (2006) latest literature review. The table reports the country of origin, a short description of the sample, the sample size, the approximate or given year of administration, the age range or average age, the percentage of formally schooled (i.e., more than 3 years of education) persons in each sample, reliability of the SPM (where given), the average raw score, range of raw SD values per subsample (where given), our IQ estimates, and the IQ estimates provided by Lynn (2006). Of the 36 samples to which we did assign an IQ (N = 13,727), the average IQ varies between 69 and 97 compared to an average IQ of 100 in Great Britain. Combining these averages results in an N-weighted mean IQ of 78 (median 78, SD = 5.8).[33] Average IQ on the SPM in the United States is approximately 2 points lower than the average in Great Britain (Lynn & Vanhanen, 2002; J. C. Raven et al., 1996). If we choose to compare the African SPM scores to an IQ of 100 for the United States, average SPM IQ in sub-Saharan Africa would be 80 (median 80) on the basis of the present samples.

The samples that were considered by Lynn, but to which we did not assign an average IQ are Wober's (1969) sample of factory workers, and Verhaegen's (1956) sample of uneducated adults from a primitive tribe in then Belgian Congo in the 1950s. Verhaegen indicated that the SPM test format was rather confusing to these test takers and that the test did not meet the standards of valid measurement. In Wober's study, the reliability and validity were too low for valid measurement (Wober, 1975). Besides, it is rather hard to believe that the highest scoring person in Wober's sample (whose raw score was 27) did not reach the cognitive level of an average 8-year-old British child. Unless one subscribes to the view that these employed men are mentally retarded, these data cannot be taken seriously.

Three of the remaining samples show average IQs below 70. These are Owen's large sample of Black school children South African tested in the 1980s, the 17 (not 26 as

[33] Were one to correct for the Flynn Effect in a way comparable to Lynn (cf. Footnote 32), the average IQ equals 77.

Lynn reports) Black South Africans carefully selected because of their illiteracy by Sonke (2001), and a group of uneducated Ethiopian Jewish children who lived isolated from the western world in Ethiopia, and immigrated to Israel in the 1980s (Kaniel & Fisherman, 1991). Apart from Owen's sample, these samples cannot be considered population samples.

Carefully selected samples are Irvine's (1969b) random selection from the 1962 standardization of several tests among schooled children in Zimbabwe (then Southern Rhodesia), the standardization data from the Northern Rhodesia Mental Survey (MacArthur et al., 1964), Notcutt's (1950) standardization samples of Zulu school children and literate and illiterate Zulu adults in South Africa, and Jedege and Bamgboye's (1981) randomly selected secondary school students in Nigeria. These more carefully sampled groups of test takers all show average IQs of 75 or higher.

There are some large discrepancies between our IQ estimates and those provided by Lynn. In some instances this is due to the Flynn Effect correction employed by Lynn (e.g., Nkaya, Huteau, & Bonnet, 1994). Other discrepancies are due to the use of different norm tables. For samples from before 1966, we used the UK norms of 1938/1948, whereas Lynn calibrated all samples against the UK norms of 1979 for children, and the 1993 US norms for adults. Despite his Flynn Effect correction, Lynn's use of recent norm tables for older samples leads to lower IQ estimates in several samples (Latouche & Dormeau, 1956; Notcutt, 1950; Ombredane, Robaye, & Robaye, 1957; Pons, 1974). In some instances, however, we were unable to replicate Lynn's IQ estimate (Kozulin, 1998; Laroche, 1959; Lynn & Holmshaw, 1990).[34] For instance, Laroche's sample of adolescent boys (average age 12.7) tested in 1955 scored on average 29.5, which corresponds roughly to the 8th percentile (i.e., IQ of 79) for these age groups in the 1979 British standardization. Because of Lynn's Flynn Effect correction (i.e., 24 years), this should be increased to an IQ of 84. However, Lynn writes that the average scores were below the 1st percentile (Lynn & Vanhanen, 2002, p. 202), and assigns an average IQ of 68 to this sample. Our estimate of IQ in this sample is based on the comparison with 1938 norms, which gives a mean score near the 17th percentile (IQ of 86) for these age groups.[35] Further differences may have arisen from the fact that we added two points to SPM scores when persons were tested individually (Ahmed, 1989; Grieve & Viljoen, 2000; Lynn & Holmshaw, 1990; Sonke, 2001), in accordance with the explicit instructions in the test manual (J. C. Raven et al., 1996). For several adult samples, Lynn's estimates lie somewhat higher than ours, possibly because Lynn employed United States norms for adults, where IQ is slightly lower than in Britain.

[34] Lynn's IQ estimate of children aged 9 and a half (Lynn & Holmshaw, 1990) is 65 (he additionally subtracts two points for the Flynn Effect). This is probably based on a (rounded) score of 12. However, the mean of this sample (i.e., 12.7) is closer to 13 and that score corresponds to an IQ of 72 for 9-1/2 year-olds. Our estimate is higher still because the SPM was administered individually in this study.

[35] Note that with the addition of the Flynn Effect correction of 2 points per decade, this value should be lowered by 3 IQ points, resulting in an IQ of 83.

Table 5.1

Sub-Saharan African scores on the Standard Progressive Matrices

Source	Country	Sample description	N	Year	Age	Edu	Rel	M	SD	IQ	IQ Lynn
(Ahmed, 1989)	Sudan	School children from Khartoum	146	±1988	8-12	100	-	18.56	-	77	72
(Crawford Nutt, 1976)	South Africa	Children from high school in Soweto	228	±1975	19	100	.82	45.00	5.6-6.1	83	-
(Grieve & Viljoen, 2000; Sonke, 2001)	South Africa	Impoverished University students in rural Venda	30	1996	19-29	100	-	37.37	6.79	75	77
(Irvine, 1969b)	Zimbabwe	Random selection of children with 8 years of education	200	1962	14-18	100	-	27.8	9.89	81	-
(Jedege & Bamgboye, 1981)	Nigeria	Random selection of secondary school students in Oyo State	755	1977	11-15	100	-	28.49	-	77	-
(Kaniel & Fisherman, 1991)	Ethiopia	Uneducated Ethiopian Jews in Israel	250	±1985	14-15	0	-	27	-	69	69
(Kozulin, 1998)	Ethiopia	Ethiopian Jews immigrated to Israel	46	±1995	14-16	100	-	28.41	8.81-10.50	72	65
(Laroche, 1959)	Congo-Zaire	Boys in schools in Elizabethville	222	1955	10-15	100	.94	29.5	8.9-11.9	86	68
(Latouche & Dormeau, 1956)	Central Afr. Republic	Candidates for centre for accelerated technical learning in Bangui	1144	±1953	17+	100	-	19.54	7.82-9.45	72	64
(Latouche & Dormeau, 1956)	Congo-Braz.	Candidates for centre for accelerated technical learning in Brazzaville	1596	±1953	17+	100	-	23.93	9.15-9.74	78	64
(Latouche & Dormeau, 1956)	Congo-Braz.	Candidates for centre for accelerated technical learning in Pointe-Noire	580	±1953	17+	100	-	23.55	7.90-9.14	78	-
(Lynn & Holmshaw, 1990)	South Africa	Children from socially representative state primary schools	350	1988	9.5	100	-	12.7	4.5	77	63
(MacArthur et al., 1964)	Zambia	Repr. sample of students in class 6	759	1963	15.5	100	-	≅ 27	-	79	77
(MacArthur et al., 1964)	Zambia	Repr. sample of students in Form II	649	1963	17.5	100	-	≅ 34	-	87	-
(MacArthur et al., 1964)	Zambia	Technical college students	195	1963	18+	100	-	≅ 30	-	84	-
(MacArthur et al., 1964)	Zambia	Mine farm youth students	292	1963	16.5	100	-	≅ 26	-	79	-
(Maqsud, 1997)	South Africa	High school students of Batswana Tribe	140	±1995	17-20	100	.83	≅ 39	-	75	-
(Maqsud, 1980b)	Nigeria	Secondary school girls in Kano city	136	±1979	13-15	100	-	38.7	5.33-6.12	85	-
(Maqsud, 1980a)	Nigeria	Boys from two primary schools	120	±1979	11-12	100	-	22.1	4.1	72	-
(Morakinyo, 1985)	Nigeria	Psychiatric out-patients and controls	28	±1983	18+	?	-	≅ 47	-	87	-

(table continues)

Table 5.1 *(continued)*

Source	Country	Sample description	N	Year	Age	Edu	Rel	M	SD	IQ	IQ Lynn
(Nkaya et al., 1994)	Congo-Braz.	Secondary school children	88	±1992	13.25	100	.91	29.6	11.6	75	73
(Notcutt, 1950)	South Africa	Zulus in primary schools near Durban	1008	1948	8-16	100	-	22.49	3.70-10.90	81	75
(Notcutt, 1950)	South Africa	Literate and illiterate Zulu adults	703	1949	17+	44	-	22.15	6.90-11.85	75	64
(Ombredane et al., 1957)	Congo-Zaire	Members of Baluba tribe	320	1954	17-29	74	-	22.14	-	75	64
(Okunrotifa, 1976)	Nigeria	Rural primary school children	50	1974	5.5	100	-	≅ 12	-	87	-
(Okunrotifa, 1976)	Nigeria	Urban primary school children	100	1974	7.0	100	-	≅ 13	-	84	-
(Owen, 1992)	South Africa	Children from schools in PWV and Kwazulu-Natal	1093	1986	16	100	.93	27.65	10.72	69	63
(Pons, 1974)	Zambia	Bemba adult males employed in mining	152	±1961	18+	100	.82	23.18	8.5	77	64
(Pons, 1974)	Zambia	Bemba adult males employed in mining	1011	±1965	18+	100	.88	33.66	9.79	87	-
(Raveau, Elster, & Lecoutre, 1976)	Madagascar	African adults working in France	143	±1975	18-49	100	-	40.92	12.47	79	82
(Raveau et al., 1976)	Various	African adults working in France	588	±1975	18-49	100	-	38.47	12.02	74	-
(Rushton & Skuy, 2000)	South Africa	University students in psychology	173	1998	17-23	100	.91	43.32	8.79	80	83
(Rushton et al., 2002)	South Africa	University students in engineering	198	±2000	17-23	100	.87	50	6.4	92	93
(Skuy et al., 2002)	South Africa	University students in psychology	70	±2000	17-29	100	-	43.20	7.84-10.24	80	81
(Sonke, 2001)	South Africa	Illiterates from rural Venda	17	1995	13-20	50	-	25.7	7.67	69	68
(Verhaegen, 1956)	Congo-Zaire	Unschooled adults from Kasai	67	±1955	18+	0	-	12.89	3.87	NA	64
(Wober, 1969)	Nigeria	Male factory workers	86	1965	18+	?	.59	15.9	4.84	NA	64
(Zaaiman, van der Flier, & Thijs, 2001)	South Africa	Disadvantaged university students	147	1995	18+	100	-	52.3	4.2	97	100

Note: The assignment of IQ values is problematic and these values are only provided in order to compare them to the IQs estimated by Lynn.

From Table 5.1, it is apparent that the samples not considered by Lynn have considerably higher average IQ than the samples that he did consider. In some cases, Lynn chose not to include in his review particular data despite the fact that these additional data were presented in the same sources from which he drew his data (Crawford Nutt, 1976; MacArthur et al., 1964; Raveau et al., 1976). In this respect, Lynn's exclusion of the large representative sample that MacArthur and colleagues collected for Form II students in Zambia is particularly striking.[36]

Note that five of the SPM samples reviewed here contain Black university students from South Africa (Grieve & Viljoen, 2000; Rushton & Skuy, 2000; Rushton et al., 2002; Skuy et al., 2002; Zaaiman et al., 2001). These students (N = 618) score higher on average (IQ: M = 88, median = 92) than the remaining samples. In some studies (Grieve & Viljoen, 2000; Rushton & Skuy, 2000; Skuy et al., 2002), the university samples scored lower than would be expected from academically selected groups. One aspect of these samples is that they were all tested by White researchers, which may have lowered test performance among these Black students (Dambrun & Taylor, 2005; but see Jensen, 1980). Moreover, various studies have shown that African American students may suffer from the performance lowering effects of stereotype threat (Steele & Aronson, 1995; Steele et al., 2002). For instance, in one study (McKay, Doverspike, Bowen Hilton, & Martin, 2002) African American students were administered the Advanced version of Raven's Progressive Matrices under one of two conditions that differed in the presentation of this test. Students who were told that they were doing an IQ test supposedly suffered from stereotype threat (i.e., the fear of conforming to the stereotype of lower IQ among African Americans), which lowered their scores by about 5 IQ points as opposed to African American students who were led to believe they were making a non-intellectual test (i.e., a less threatening condition in which the stereotype is irrelevant). Although we are not familiar with any studies of the effect of stereotype threat on test performance in (South) Africa, given the long history of constitutionalized discrimination of Blacks in South Africa, it would not be surprising if stereotype threat has an effect (Suzuki & Aronson, 2005). According to stereotype threat theory (Steele et al., 2002), this effect should have particularly strong negative effects on test performance of test takers who are academically well motivated and for whom intelligence is an important aspect of their identity, such as university students. Further research into the effects of stereotype threat in (South) Africa is clearly needed.[37]

In contrast to the university students, 734 cases (5.3%) in Table 5.1 had no formal schooling (defined as 3 years of education or less). These 734 uneducated test takers had an N-weighted average IQ of approximately 71, which is considerably below the overall average. Note that the SPM may lack validity in samples with no formal schooling (Dague, 1972), but lower scores among non-schooled test takers may also reflect true levels of

[36] Lynn only used the representative sample of Standard 6 students of this elaborate study. It is unclear why he excluded the other large sized samples in the Northern Rhodesia Mental Ability Survey, most notably the large representative sample of Form II students. The raw median values of all samples are presented next to one another in one table on page 84 of the report, so he could not have missed the other sample medians.

[37] Dambrun and Taylor (2005) claimed that the entire Black-White IQ gap in the US can be accounted for by the effects of stereotype threat, but this conclusion is not warranted (Sackett, Hardison, & Cullen, 2004; Wicherts, 2005b). In addition, the effects of stereotype threat have not been studied on representative samples of African Americans, but mostly on small samples of university students.

lower ability. Because in sub-Saharan Africa the percentage of unschooled young people is around 20%,[38] we may want to correct for this underrepresentation of unschooled persons. A rough stratification for educational level could be achieved by adding 2514 fictional uneducated cases with an IQ of 71 to the total sample. This would lower the average IQ by one point to 77. In sum, average IQ of sub-Saharan samples covered in this review equals 78, or 77 when corrected for the underrepresentation of uneducated subjects. This needs to be raised to 79 or 80 when compared to US norms. This is considerably higher than Lynn's estimate of sub-Saharan African IQ based on the SPM data, which, when weighted by sample size, would result in a mean IQ of 69 and a median of 64.

Psychometric properties. As can be seen in Table 5.1, the reliability of the SPM was computed in several studies in Africa. Reliabilities are generally above 0.80, which is comparable to those found in western samples (J. C. Raven et al., 1996). Only in Wober's (1969) sample, the reliability was unacceptably low.

Convergent validity of the SPM is studied mainly in South-Africa. Grieve and Viljoen (2000) report a correlation of 0.40 between SPM scores and a reasoning test. MacArthur et al. (1964) and Notcutt (1950) correlated SPM scores with various achievement and cognitive ability tests, which resulted in reasonably high correlations. Moreover, SPM scores were found to correlate considerably with the performance of a perceptual learning potential test among South-African students (Skuy et al., 2002), and with the performance on a verbal learning task among healthy and unhealthy adults in Nigeria (Morakinyo, 1985). Likewise, Crawford Nutt (1977) reports significant correlations between the SPM and several reasoning tests.

In Crawford Nutt's (1977) principle axis analyses on these data, the SPM scores did not show the highest factor loading on the dominant axis, indicating that the SPM may not be as highly *g* loaded as it is in western samples. Irvine (1969b) conducted a factor analysis of SPM items and concluded that, unlike in the western samples studied by him, the SPM was not unidimensional in African samples. In a large scale factor analytic study employing data from Zambia and Zimbabwe, Irvine (1969a) found that the SPM was not solely an indicator of *g* in one sample, although it was in another sample.

Predictive validity of the SPM was studied by Zaaiman et al. (2001), who found that the SPM correlated reasonably well with college performance. In addition, Maqsud (1980a) found highly significant correlations between SPM scores and school grades. An interesting aspect of Maqsud's study was that these correlations were generally higher in the modern school than in the more traditional school. However, this could also be due to differences between schools in grading practices or student population.

Some studies failed to support validity of the SPM in Africa. Laroche (1959) found non-significant or only low correlations between SPM scores and school grades. In stark contrast to comparable samples of children from Britain, Japan, and Hong Kong (cf. Lynn, 1991), Lynn and Holmshaw (1990) did not find SPM scores in their sample of Black South African children to correlate significantly with reaction time tasks (at least not with the cognitive aspects of these tasks; Jensen, 1998). Ogunlade (1978, not in Table 5.1) obtained a correlation of only 0.15 between SPM scores and school achievement among 537

[38] Based on UNESCO estimates of gross enrollment ratio in primary education over the period 1970-2003.

secondary school students in Nigeria. Finally, Wober (1966; cf. Wober, 1969) reported a non-significant negative correlation between SPM scores and assessments of job efficiency among 173 Nigerian employees. These negative results signal the need for more validity studies of the SPM in Africa.

To summarize, judged by correlations with criteria and other tests, the SPM has been found to be valid as well as invalid in Africa. Several studies support convergent and predictive validity of the SPM in Africa, particularly among samples with relatively high scores. However, in light of the many claims of unsuitability of the SPM in Africa (Irvine, 1969b; Ogunlade, 1978; Verhaegen, 1956; Wober, 1966, 1975) more research into the construct validity of the SPM is clearly needed. There is also a clear need for more work on the factorial status of the SPM in African samples (Irvine, 1969a). It is unclear (Crawford Nutt, 1977) whether the SPM is as highly *g*-loaded in Africa as it is in the west (Carroll, 1993), and whether the SPM is factorially pure remains to be seen. Additional factor analyses with a larger battery of tests could shed light on this issue.

In none of the studies reported in Table 5.1 was measurement invariance studied with the methods of contemporary item response theory, and we are unfamiliar with any study of differential item functioning in which western SPM scores are compared to scores of African samples. We return to the issue of measurement invariance below. In the absence of information on measurement invariance of the SPM, the degree to which measurement bias may have led to an underestimation of ability in the sub-Saharan African samples remains unknown. Nevertheless, considering the low scores in some samples, such an underestimation is rather likely.

The Flynn Effect in the SPM. The results in Table 5.1 are based on data from diverse samples, of varying age groups, and from different countries. Therefore, any secular trend in these data represents only a tentative indication of African IQ trends. Nevertheless, the adult samples (ages 17 and higher) are fairly comparable with respect to age, because they all include young adults (even the Raveau samples only include a handful of cases above the age of 40). In the study of adult trends, we excluded the university samples, because all of these are quite recent. We studied the Flynn Effect in the current samples by comparing all raw scores to the norms from the older standardization samples (i.e., 1938 for children and 1948 for adults). For the newer samples this resulted in higher IQs than the values in Table 5.1. The results are plotted in Figure 5.1. In this figure, we present separate (*N*-weighted) regression lines of IQ on year of administration for adults (solid line) and children (dashed line), separately.

As can be seen, the steep regression line for the adults suggests the presence a considerable Flynn Effect, while the regression line of the children samples is more flat. Both (*N*-weighted) regression lines deviate significantly from zero ($p < .001$). These regression lines are equivalent to increases of 7 IQ points per decade for adults,[39] and 2 IQ points per decade for children. The rise of adults is comparable to that reported for male adults in the Netherlands from 1952-1982 (Flynn, 1987), while the increase for children is comparable to the increase in Great Britain among children from 1949 to 1982 (Lynn &

[39] Note that a Flynn Effect correction of the average IQs on the basis of this result is not necessary, because we aim to compare these IQs to British norms. In Britain, the gain in SPM scores equals about 2 IQ points (Lynn & Hampson, 1986). As said, this correction lowers average IQ by one point.

Hampson, 1989). A comparison of the adult samples from the different eras does not provide a compelling reason to think that samples are incomparable, so the rise in adult samples appears to be a robust phenomenon. When tested for significance without weighing for sample size, the Flynn Effect for adults remains significant ($p < .001$), whereas in the samples of children, the Flynn Effect is no longer significant ($p > .05$). In sum, there appears to have been a considerable Flynn Effect for African adults on the SPM among the samples considered here. More comparable adult samples are needed for more accurate estimates of the Flynn Effect in Africa. Nonetheless, the secular gain in Africa suggests that the IQ gap between British and African adult test takers has diminished over the years. There is an indication of a smaller rise in the children's samples, but a more definitive indication of a Flynn Effect among African children should await more comparable samples.

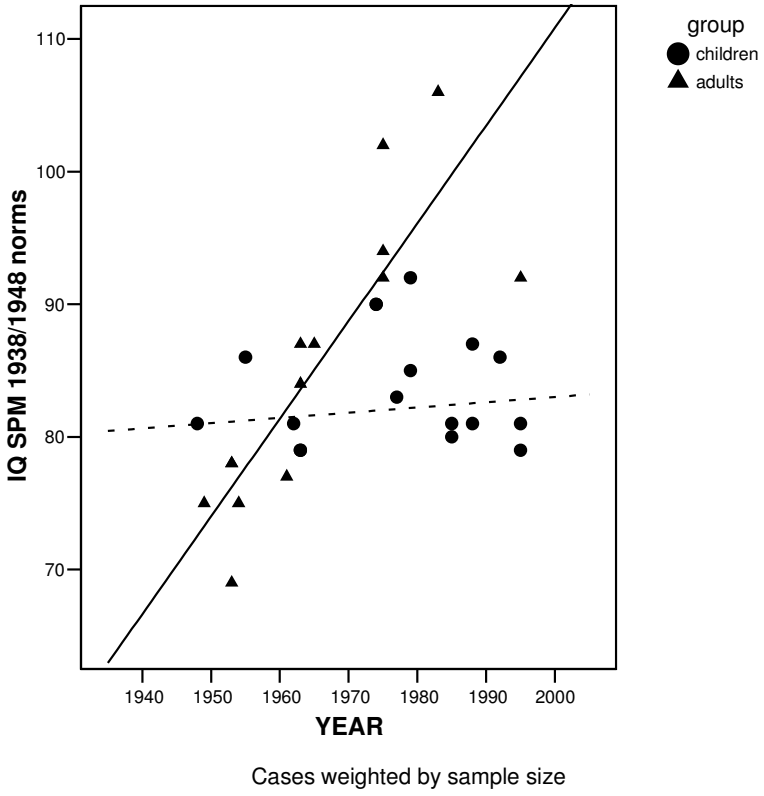

Figure 5.1 *Secular trends in IQ for adult and children samples on the SPM.*

Coloured Progressive Matrices

IQ estimates. Table 5.2 reports on the twelve studies in which the Raven's Coloured Progressive Matrices was administered to sub-Saharan African children (combined N = 4,313). The average IQs vary from 68 to 94. The N-weighted average of the twelve samples equals 78, (median 77, SD = 7.2). If we exclude Fahrmeier's (1975) study, in which the

CPM was administered in a non-standard fashion, we arrive at an average IQ of 79 (median 82, *SD* = 6.9). Therefore, the average IQ of children in sub-Saharan Africa on the CPM appears to be 78 or 79.[40] When compared to an average IQ of 100 for the US, this IQ among sub-Saharan African children equals 80 or 81.

Besides the Fahrmeier's data, the samples that score relatively low are the children from poor rural areas tested by Jinabhai et al. (2004), a sample of Ethiopian orphans (Aboud, Samuel, Hadera, & Addus, 1991), and the representative sample of Ghanaian children (Glewwe & Jacoby, 1992; Heady, 2003). The low IQ for orphans is not surprising (but see Wolff, Tesfai, Egasso, & Aradom, 1995), given the harsh circumstances that such children often encounter (Aboud et al., 1991). Moreover, IQ in rural areas is often lower than in urban areas (e.g., Loehlin, 2000). However, the low average IQ of the representative sample of children in Ghana is peculiar, given that of all sub-Saharan countries Ghana is relatively well-developed (UN Development Programme, 2005). The low scores could be explained by the fact that the tests were administered in children's houses. As the principle investigator put it: "[the test takers] may have been sitting in a chair or even on the ground" while taking the tests (P. Glewwe, personal communication, January, 17, 2006). This may have lowered the scores. Two recent representative standardization samples in Kenya (Costenbader & Ngari, 2001) and South-Africa (Knoetze, Bass, & Steele, 2005) show average IQs around 80. The highest scoring samples are those of private school children, whose fathers' SES is high (Okonji, 1974), and a small sample of Ethiopian Jewish children in Israel (Tzuriel & Kaufman, 1999).

Of the total of 4,165 children in the school-aged range, about 793 children (19 %) did not attend school. This is approximately equivalent to the population estimates of school attendance in current day sub-Saharan Africa (cf. Footnote 38). Moreover, the number of rural children and urban children in the samples in Table 5.2 appear to roughly reflect the population distribution in sub-Saharan Africa. Moreover, of the 11 samples considered, four are considered by the authors to be representative for a particular population. Although definitive statements require completely stratified random population samples, the data in Table 5.2 appear to provide a reasonable estimate of average IQ of African children on the CPM.

In two instances, Lynn's estimate of average IQ is lower than would be expected from his Flynn Effect correction (Costenbader & Ngari, 2001; Jinabhai et al., 2004). It is conceivable that Lynn's estimates are lower because he used the CPM-to-SPM conversion method to estimate IQ. With respect to the Ghanaian data, Lynn used as his source one average CPM score from a larger sample of ages 9-18 given by Glewwe and Jacoby (1992; see also Rushton & Jensen, 2005a). Our IQ estimate of this sample is based on mean scores reported separately for age and gender (from Heady, 2003), and can be regarded more accurate. In addition, we excluded age groups for which there no British CPM norms exist.

[40] With the addition of a Flynn Effect correction of 2 IQ points per decade, the average IQ should be lowered by 2 points.

Table 5.2

Sub-Saharan African scores on the Coloured Progressive Matrices

Source	Country	Sample description	N	Year	Age	Edu	M	SD	IQ	IQ Lynn
(Aboud et al., 1991)	Ethiopia	Children in an orphanage in Jimma	134	±1989	5-11	100	13.56	-	72	-
(Costenbader & Ngari, 2001)	Kenya	Children from representative schools	1222	±1998	6-10	100	15.86	3.51-7.82	82	75
(Daley et al., 2003)	Kenya	Children from rural district of Embu	118	1984	7.5	100	12.82	3.21	75	76
(Daley et al., 2003)	Kenya	Children from rural district of Embu	537	1998	7.5	100	17.31	2.56	90	89
(Fahrmeier, 1975)	Nigeria	Schooled and unschooled children in town in North-Nigeria	334	±1973	6-11	57	11.42	-	68/NA	69
(Heady, 2003)	Ghana	Representative population sample	589	1988	9-11	82	15.80	-	72	-
(Heyneman & Jamison, 1980)	Uganda	Students in 61 representative primary schools	1907	1972	10-18	100	27.07	8.47	NA	73
(Jinabhai et al., 2004)	South Africa	Children from 11 rural primary schools in poor Vulamehlo district	806	±2002	8-10	100	13.9	3.9	72	67
(Knoetze et al., 2005)	South Africa	Xhosa-speaking primary school students in peri-urban Eastern Cape	172	±2002	7.5-11	100	17.21	2.50-5.84	77	-
(Okonji, 1974)	Nigeria	Children in private school in Lagos	73	1972	8-11	100	23.52	4.09-6.10	94	-
(Ombredane, Robaye, & Plumail, 1956)	Congo-Zaire	Children of "very underdeveloped" Asalampasu tribe	151	±1955	6-11	79	14.50		76	-
(Tzuriel & Kaufman, 1999)	Ethiopia	Ethiopian Jews immigrated to Israel	29	±1992	6-7	100	15.60	1.65	94	-
(Wolff et al., 1995)	Eritrea	Orphans and refugee children during war	148	1990	4-7	NA	12.4	3.0-3.6	87	-

Note: The assignment of IQ values is problematic and these values are only provided in order to compare them to the IQs estimated by Lynn.

The average IQ in samples considered by Lynn is lower than the average IQ in than the samples that Lynn did not consider. But then we did not consider several adult and adolescent samples that were administered the CPM. It is certainly the case that the adult samples studied by Berry (1966) and others (Berlioz, 1955; Binnie Dawson, 1984; Kendall, 1976) showed very low CPM averages as compared to western samples (Measso et al., 1993). However, in some cases these averages are too low to be credible. For instance, Berry's sample of adults scored below 14 on average. It is hard to believe that these men were all severely mentally handicapped (if so, Berry would have presumably mentioned this in his paper). Besides, these samples cannot be considered random population samples. In two studies with the CPM (Berry, 1966; Binnie Dawson, 1984), the authors deliberately sampled adults with very little knowledge of western culture. The samples of adults in Tanzania and Kenya (Boissiere et al., 1985), and a large sample of adolescents from Uganda (Heyneman & Jamison, 1980) were more carefully sampled, and showed much higher average CPM scores. Even if we had included these adult and adolescent samples, average IQ based on the CPM scores in sub-Saharan Africa would not change very much. The reason is that most of these low scoring samples are small, whereas the large sample of adolescents in Uganda (N = 1907) showed an average score equivalent to an average IQ above 79.[41]

Psychometric properties of the CPM. Only in the study by Costenbader and Ngari (2001), is the reliability of the CPM reported. Both the internal reliability (0.87) and the test-retest reliability (0.84) in this study are sufficient and comparable to those in western samples. Several studies in Table 5.2 focused on convergent and predictive validity of the CPM. For instance, Tzuriel and Kaufman (1999) found that the CPM correlated reasonably well with two dynamic tests of cognitive ability, Aboud et al. (1991) found the CPM predicted school grades (r of 0.36 with age partialled out), and Heady found that CPM scores correlated significantly with scores on math and reading tests. In a study among 85 adolescents in Kenya (not in Table 5.2), CPM scores correlated reasonably well with vocabulary test scores, but non-significantly with the scores on a practical intelligence test (Sternberg et al., 2001).

In several studies, the validity of the CPM was not supported. For instance, the correlation of the CPM with school grades in Fahrmeier's of sample of school children was only significant in two of the seven classrooms, where these correlations could be computed (although power may have been low due to small sample sizes). Similarly, correlations of CPM scores with other cognitive ability tests were low in two other studies (Jinabhai et al., 2004; Okonji, 1974). In yet another study (not in Table 5.2) among 196 children in Benin, the CPM correlated quite lowly with seven other cognitive tests (van den

[41] Lynn (2006) states that the 1907 primary school students tested with the CPM in Uganda (Heyneman & Jamison, 1980) are 11 years old, but most of these students are around 13. It is important to note that the score reported in Lynn's source (Heyneman & Jamison, 1980) is based on the number correct out of 33 instead of 36 items (Heyneman, 1975), but Lynn's source does not mention this. The first three items were used for instruction, so the average score needs to be raised by 3 points (these items typically have p-values of 1). If we add three points to the score and employ the CPM to SPM conversion (J. C. Raven et al., 1996), we can compare scores to the SPM norm table for the correct age range. This results in a rough estimate of an average IQ of 79. Due to the ceiling effect discussed earlier, this figure is likely to be too low.

Briel et al., 2000). In sum, validity of the CPM in sub-Saharan Africa shows some promise, but has not been established, and needs to be studied further.

Unfortunately, in none of the studies in Table 5.2, measurement invariance was studied. Nevertheless, Ombredane and colleagues did study item characteristics in their samples (Ombredane, 1957; Ombredane et al., 1956) and found that the CPM showed a relatively large number of Guttman errors (e.g., Meijer & Sijtsma, 2001) in their samples. We were unable to locate any rigorous study of DIF in which CPM scores were compared across western and African samples, neither did we find studies into the factorial characteristics of the CPM. In the absence of such studies, it is uncertain to what degree lower CPM scores of sub-Saharan African children as opposed to western children reflect lower levels of general intelligence in the former group. The degree to which measurement bias may have led to an underestimation of ability in the sub-Saharan African samples remains unknown.

Flynn Effect in the CPM. Daley et al. (2003) already documented a Flynn Effect in the CPM among two comparable samples of children from rural Kenya. If we exclude these two Kenyan samples, there is no indication of a Flynn Effect in the remaining samples in Table 5.2. The number of studies is fairly small, and all but three of the samples antedate 1980. More definitive conclusions with respect to a Flynn Effect on the CPM in Africa require more comparable samples.

Draw-a-Man Test

Goodenough-Harris Draw-a-Man test (DAM; Goodenough, 1926; Harris, 1963) is a non-verbal intelligence test for children aged two to thirteen in which children are required to make a drawing of a man. This drawing is rated on 51 (original version; Goodenough, 1926) or 73 (revised version; Harris, 1963) criteria that reflect cognitive development. Scores on the Draw-a-Man test have been shown to correlate reasonably well with scores on cognitive ability tests such as the Stanford-Binet (e.g., J. H. Williams, 1935) and the SPM (Carlson, 1970). It should be noted that the DAM test is not generally considered as good an indicator of general intelligence as tests like the SPM or CPM. We have nevertheless included this test in the current review, because the DAM can be administered easily and at low cost. For that reason, it is used commonly throughout Africa. Also, Lynn used DAM scores to estimate the average IQ in Africa.

IQ estimates. The results of eleven studies of the DAM in sub-Saharan Africa are reported in Table 5.3. There are several samples to which we did not assign an IQ estimate for the simple reason that the administration of the DAM in these samples proved was fraught with difficulties. The first of these is Fahmy's (1964) study among schooled and unschooled children of a primitive tribe in Sudan. He indicates that "[the] children who had no schooling, never used a pencil, and have no experience in how to conceptualize their visual image" (p. 172). Moreover, most of the unschooled children "recruited from under the bush" by Fahmy were naked. It is noteworthy that within Goodenough's scoring scheme of the DAM test, five out of a total of 51 points are awarded for clothing worn by the drawn man. Not surprisingly, Fahmy considered the DAM test unsuitable for these Sudanese children, regardless of school attendance. In a study with the DAM also involving Sudanese children, Badri (1965b) noted that: "Many [children from remote villages] hold

pencils in unusual ways and say they have never before made a drawing on paper" (p.333). Badri therefore reaches a conclusion similar to Fahmy's with respect to the unsuitability of the DAM for these Sudanese children he tested. Despite these obvious problems, Lynn assigned these samples low IQs on the basis of their DAM performance. However, the DAM appears to unsuitable for African children without schooling (cf. Serpell, 1979). The most obvious reasons for this are inexperience with pencil drawing and the unfamiliarity with two-dimensional pictures, which is often encountered among these children.

The nine samples to which we assigned an average IQ (combined N = 4,459) have average IQs varying from 76 to 99. The N-weighted average IQ equals 81 (median 76, SD = 6.3) when compared to the US norms published in 1926 for the older and 1963 for the more recent samples.[42] Combined, the samples with assigned DAM IQs appear to be roughly representative for school-going children. However, more definitive statements on average IQ of African children on the DAM have to await more carefully sampled data. The lowest scoring sample, which is described as fairly representative for the urban school children in South Africa, is also the largest sample. The DAM in this study appeared not to have been administered under ideal circumstances: "Classroom conditions were not ideal from the point of view of scientific test administration" (Hunkin, 1950, p. 54). This may have lowered test performance to an unknown degree.

One aspect that needs attention is the fact that the IQs in the three samples for which the 1926 norms were used (Badri, 1965a; Bardet, Moreigne, & Sénécal, 1960; Hunkin, 1950; Vernon, 1969) are not regular IQs (i.e., those based on the standardized normal distribution with M = 100, SD = 15), but IQs based on the outdated concept of mental age (i.e., the mental age times 100, divided by the chronological age). Raw scores in the Badri and Vernon studies are not given. However, Hunkin and Bardet et al. provide the average raw scores of their samples. Together with the means and SDs for US norm groups for ages 6-10 (from Goodenough, 1926), this enables a computation of the standardized IQ for these age groups in the samples in Hunkin (N = 1067) and Bardet et al. (N = 494), which results in average IQs of 83 and 76, respectively. When only the data of standardized IQs are used (combined N = 2,805), the N-weighted average IQ as compared to the US DAM norms equals 84 (median 83, SD = 5.1).[43] It is again apparent from Table 5.3 that the samples not considered by Lynn showed higher IQs than the samples he did consider. In this table, however, any discrepancy between Lynn's IQ estimates and ours are due to Lynn's corrections for outdated norms.

[42] The norms used are relatively old. A correction for the Flynn Effect according to Lynn's approach (3 IQ points per decade), results in an overall average IQ of 74.
[43] With a Flynn Effect correction, this needs to be lowered with 6 points.

Table 5.3

Sub-Saharan African scores on Goodenough's Draw-A-Man test

Source	Country	Sample description	N	year	age	Edu	IQ	IQ Lynn
(Badri, 1965a)	Sudan	4th grade boys from rural and urban areas	293	±1963	Not given	100	86[1]	74
(Badri, 1965b)	Sudan	Culturally deprived preschool boys	80	±1963	6	NA	NA	64
(Bakare, 1972)	Nigeria	Upper-class and lower-class school children	393	±1970	6-15	100	87	-
(Bardet et al., 1960)	Senegal	School children from Dakar and rural area	750	±1958	6-15	100	76 / 76[1]	-
(Fahmy, 1964)	Sudan	Children from primitive Shilluk tribe	184	1954	7-13	M	NA	52
(Hunkin, 1950)	South-Africa	Children from native schools in Durban	1729	1947	6-13	100	83 / 76[1]	70
(Minde & Kantor, 1976)	Uganda	Children in three primary schools	514	1972	9-14	100	89	-
(Nwanze & Okeowo, 1980)	Nigeria	Children with reading problems	13	±1978	5-10	100	99	-
(Ohuche & Ohuche, 1973)	Sierra Leone	Children in experimental school	202	1968	5-11	100	95	-
(Richter, Griesel, & Wortley, 1989)	South-Africa	Urban school children from townships	415	1988	5-13	100	84	77
(Skuy, Schutte, Fridjhon, & O'Carroll, 2001)	South-Africa	Soweto secondary school children	100	±1998	12-24	100	83	-
(Vernon, 1969)	Uganda	Boys of above average SES	50	±1965	12	100	95[1]	-

Notes. [1] IQs based computation with mental age. Remaining IQs are based on standardized IQs.
The assignment of IQ values is problematic and these values are only provided in order to compare them to the IQs estimated by Lynn.

The 1963 norms of the DAM have been strongly criticized for being inaccurate (Howard Scott, 1981). If we add to the IQs in the recent samples the 10 points to correct for the inaccuracy of the 1963 norms (as suggested by Howard Scott), the average *N*-weighted standardized IQ on the DAM of the samples (N = 2,805) becomes 88 (median 83, SD = 9.3).[44]

In sum, the average IQ on the DAM test can not be accurately determined. It is clear, however, that the average IQ of African samples is well above 80, and not the average IQ of 70 that Lynn reported.

Psychometric properties of the DAM. Two studies in Table 5.3 provide information of reliability of the DAM in sub-Saharan Africa. Test-restest reliability reported by Ohuche and Ohuche (1973) equals 0.82, but those reported by Minde and Kantor were considerably lower (0.63 - 0.66). Nonetheless, these values are in the range of values given in Harris' (1963) manual. Predictive and convergent validity was studies in several samples. Richter and colleagues (Richter et al., 1989) found a strong correlations (multiple r = 0.64) between the DAM and five cognitive ability tests among the younger age group (ages 5 - 7). However, among children aged 8 - 13, the DAM did not correlate significantly with four (other) cognitive ability tests (multiple r = .20, NS).[45] In the same sample, DAM scores correlated significantly with school performance, although common variance was rather small (r = .37, r^2 =.14). Other studies in Africa also documented low correlations between DAM scores and school performance (Bakare, 1972; Minde & Kantor, 1976), particularly for older age groups (Hunkin, 1950; Ohuche & Ohuche, 1973). Predictive validity of the DAM for school performance appears to be reasonable for young children, but insufficient for those above age 8 (but see Nwanze, 1985). These results are in line with studies in the US that showed that DAM scores do not predict academic achievement very well (Howard Scott, 1981).

Of all studies in Table 5.3, only Hunkin (1950) considered item characteristics in a rigorous manner. She documents several items on which African children score lower than US children. For instance, among the African children, the item related to clothing on the drawn man (Item 9a) showed marked lower performance than in the US standardization sample. Whereas Hunkin (1950) concludes that the test is suitable in principle for Urban Black children, she clearly states that US norms should not be used for that population. In fact, several authors (Badri, 1965a; Minde & Kantor, 1976; Munroe & Munroe, 1983; Serpell, 1979), including the test developers themselves (Goodenough & Harris, 1950), have argued that the comparison of DAM scores across cultures is problematic, because of cultural differences in experience with pencil-drawing on paper, and because several aspects of the scoring scheme are clearly culturally loaded. These problems signal a strong need for more insight into differential item functioning of the DAM test. However, we were unable to locate studies into measurement bias of this test using modern methods. In light of the absence of such studies, severe caution should be entertained in the interpretation of these

[44] With a Flynn Effect correction, this estimate of average overall IQ equals 82.

[45] Richter et al. (1989) argue that the DAM test underestimates the IQ of test takers of eight years and older. This appears to be based on the fact that the mean scores of these older test takers differ more from the US mean than the mean scores of younger children. However, they have failed to take into account that the *SD* increases with age. When IQs are computed, the age groups above 7 have slightly higher IQ than the younger age groups. The exclusion of these age groups does not alter the overall *N*-weighted average IQ.

average scores in Table 5.3, and the average IQs we reported above. There is a real possibility that the DAM underestimates latent cognitive ability among African test takers. Nevertheless, the DAM shows some promise as a test of cognitive ability in Africa, particularly for younger test takers, provided that these are familiar with pencils and drawing.

Flynn Effect in the DAM. Richter et al. (1989) reported a secular rise in test scores among Zulu children in South Africa from 1947 to 1988, but this may be due to the suboptimal testing conditions in Hunkin's (1950) study. The number of remaining samples in Table 5.3 is small, and there is no apparent secular rise in these samples. The study of Flynn Effect in the DAM is further complicated because of the use of different scoring schemes in older and newer samples.

Kaufman Assessment Battery for Children

In a series of studies, Boivin, Giordani and co-workers administered the K-ABC to children in the Democratic Republic of Congo (Boivin & Giordani, 1993; Boivin, Giordani, & Bornefeld, 1995; Giordani et al., 1996). Lynn used these data sets to substantiate his claim of low IQ levels among Africans. However, the African data from the K-ABC are not very convincing as far as average IQ of the African population is concerned. The first problem with these data is that the studies were mainly concerned with the effect of intestinal parasites (Boivin & Giordani, 1993) and malaria (Boivin, Giordani, Ndanga, Maky, & et al., 1993) on cognitive development. For that reason, the children in these samples were all from underdeveloped rural areas. In some studies, children were especially selected for their poor health (Boivin & Giordani, 1993). Of course, malaria and intestinal parasite infections are common in tropical Africa, but such selective samples cannot be used to estimate the average IQ of the African population.

To make matters worse, in these samples K-ABC tests were adapted to be administrable to rural children in Africa (Giordani et al., 1996). For that reason, the instructions and items were changed. It is unclear to what extend this has altered the measurement properties of the K-ABC. For all of these children, individual cognitive assessment was an entirely new experience. More importantly, for most of the children, it was their first encounter with color-printed material. Giordani and colleagues (1996) studied the psychometric properties of the K-ABC in their rural African samples. However, they are also severely cautious with respect to the comparability of these African scores to US norms. For instance, in some K-ABC subtests, items included objects that were rather unfamiliar to these test takers, such as telephones. It is therefore likely that at least some items in the K-ABC show DIF (Giordani et al., 1996), and that several subtests are not comparable across Western and African samples. Lynn uses these samples in his overview without regard of the clear warnings by the original authors with respect to the incomparability of these scores to western samples. Lynn also included in his overview the scores from a sample of 184 Kenyan rural children who all suffered from malaria (Holding et al., 2004). In this study, *all* subtests of the K-ABC were altered. Lynn assigned the sample an IQ of 63, but it is entirely unclear to what degree the alterations in the test even allows for the comparison to US norms.

There exist additional data from the K-ABC in Africa, but these data are not considered by Lynn. First, in one study (Skuy, Taylor, O'Carroll, Fridjhon, & Rosenthal, 2000) the K-ABC was administered to 21 Black children from South Africa, and the average IQ was found to be 98. In yet another study, the average IQ on the basis of the K-ABC in a sample of Senegalese children equaled 81 (Boivin, 2002). As was the case with the other tests, the samples not considered by Lynn show higher average IQs than the samples he did consider. In sum, because of the special nature of the samples, the changes in this test, and the likely presence of measurement bias, the data from the K-ABC considered by Lynn cannot reasonably be used to arrive at an estimate of the average IQ of African children.

Wechsler Scales

Lynn included in his overview of African IQ several studies using the Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981) (Avenant, 1988 in Nell, 2000; Nell, 2000) and the Wechsler Intelligence Scale for Children-Revised (WISC-R; Wechsler, 1974) (Fernández-Ballesteros, Juan Espinosa, Colom, & Calero, 1997; Skuy et al., 2001; Zindi, 1994a). Lynn's choice of studies with the Wechsler scales is rather awkward. First, he uses data from Nell, who argued strongly that the use of WAIS-R and WISC-R among South African Blacks can lead to underestimations of ability (Nell, 2000). Nell provides the results of the Avenant study and of one of his own studies to illustrate his point, and Lynn subsequently presented the results obtained in these two samples in support of a low IQ among sub-Saharan Africans. Nell concludes on the basis of these studies that "the Wechsler tests lack validity for these subjects" (p. 27). Lynn has every right to disagree with Nell's assessment of the unsuitability of the Wechsler scales for African test takers. However, in the WAIS-R[46] subtests used by Avenant some items were changed, and it is uncertain whether this has changed the difficulty of items. It is also noteworthy that these samples cannot be considered population samples.

One study often referred to in the literature (Lynn, 2006; Rushton & Jensen, 2005a) is that by Zindi (1994a). This particular study was concerned with the suitability of the US version of the WISC-R for Zimbabwean high school children. Zindi clearly indicated that the WISC-R needed adaptation to remove language difficulties, and he stressed that some instructions and items in the WISC-R may not be appropriate for Zimbabwean children. In a subsequent study, Zindi (1994b) eventually found that some small alterations in the WISC-R greatly enhanced average IQ in Zimbabwean children, a fact neither discussed nor mentioned by those who attach value to the average IQ found by Zindi in his first study.

In yet another study with the WISC-R, Skuy and colleagues indicate that "language has a considerable effect on test performance" (Skuy et al., 2001, p. 1422). In fact, in this sample average IQ is lowered because of the low performance on the vocabulary subtest and other verbal subtests. For most Black Africans, English is not the native language, and it is well known that the Wechlser scales have a strong English language component. In addition, several of the non-verbal (performance) subtests in the Wechsler scales have items displaying typically western objects and situation that may be less familiar to African

[46] Lynn reports that the WISC-R was given, but these subjects completed the WAIS-R.

test-takers. Thus, cultural bias in the Wechsler scales cannot simply be ignored. Unfortunately, in none of the studies with the WAIS-R or WISC-R mentioned by Lynn, reliability, inter-subtest correlations, or validity were reported. Besides, we are not familiar with any factor analyses on western Wechsler scales among Africans. The WISC-R data from Zindi (1994a) and Skuy et al. (2001) were submitted to analyses with the method of correlated vectors (Rushton, 2001; Rushton & Jensen, 2003). However, these analyses did not test whether the factorial structure the WISC-R of both these African samples was comparable to that of western samples. Moreover, the method of correlated vectors is not a suitable method to study measurement invariance (Dolan, 2000; Dolan et al., 2004; Lubke et al., 2001). Moreover, we were unable to locate a single study into the measurement invariance of Wechsler tests between western samples and samples of Africans.

In the last study with the WISC-R in Africa that Lynn reports to substantiate his claim of low African IQ, the IQ of a sample of forty-eight 10-14 year-old children was found to be 63 (which Lynn corrected downwards because of the Flynn Effect). This small sample was used to estimate the average IQ of the entire population of Equatorial Guinea, resulting in an IQ estimate of 59 for this country (Lynn & Vanhanen, 2002). Unfortunately, the use of this particular sample cannot possibly be more inaccurate. The average IQ of the people of Equatorial Guinea is based on a lengthy book chapter (Fernández-Ballesteros et al., 1997). Although this chapter reports research conducted on an illiterate tribe in Equatorial Guinea, the WISC-R was not administered to these African subjects. The forty-eight children who were administered the WISC (not the WISC-R) were not from Equatorial Guinea, and not even from Africa. In fact, the sample in question solely contains Spanish children who attended a Spanish school for handicapped children. Half of these subjects were mentally handicapped; the other half attended the school because of their low IQ. Clearly, Lynn has made a mistake in using this sample to estimate African IQ.[47]

In addition to ignoring explicit statements on the inappropriateness of the WISC-R and WAIS-R by original authors who provided Wechsler data, Lynn also missed other Wechsler data from sub-Saharan Africa. These additional data on Wechsler IQ of Black South-Africans provided higher IQs than those Lynn reported. In one study, 40 educated adults scored an average IQ of 94 on the US WAIS-III (Shuttleworth Edwards et al., 2004). In yet another study, the average WISC-R IQ of 21 Black children *with learning difficulties* was found to be 84 (Skuy et al., 2000).[48] To conclude, data from the Wechsler scales in Africa provided by Lynn does not lend much credibility to his claim that average IQ in Africa is below 70. In addition, there is a need for more research into the appropriateness of the

[47] It is rather disconcerting that Lynn makes a bold statement to the effect that the majority of the people in this West-African country are mentally retarded, yet has not read his source more carefully. The chapter clearly indicates that this experimental study with 48 subjects was conducted in Spain. The mean IQ is mentioned two times, the first time as follows: "A similar design was used in our second experiment with forty-eight subjects, 10 to 14-year-olds, attending a school for handicapped children (63.025 IQ mean)" (Fernández-Ballesteros et al., 1997, p. 253). Indeed, this is the only IQ mean mentioned in the entire chapter, and there are no other samples of size 48 of this age range in the chapter. In a later part of the chapter, the same sample and the same mean IQ are again mentioned. There, the text clearly states that half of the subjects were diagnosed as having brain organic disorders. Moreover, judging by the reference list, the test at hand was the Spanish WISC, and not the US WISC-R.

[48] This is the same sample of 21 Black South African children who completed the K-ABC.

Wechsler scales in Africa. Without such research it is unclear what low Wechsler IQ scores in Africa mean.

Remaining Tests

We now discuss the additional data sets used by Lynn to support his claim that average IQ in Africa lies below 70. One of these studies was concerned with the effects of coaching on test performance (Lloyd & Pidgeon, 1961). In this study, a "fairly representative sample" (p. 147) of South African Zulu children were administered the Non-Verbal Test, a test normed among English children and published in 1951 (Buros, 1959). The Zulu children (N = 275) had an average IQ of 87 on the pretest (i.e., without coaching). Lynn (2006) does not discuss how he arrived at his estimate of an IQ of 74 for this sample, but his estimate is clearly off the mark (in his 1978 review he correctly reported a value of 87).

In another study, Buj (1981) provided the results of the administration of the Culture Fair Intelligence Test to 225 adults in the Ghanaian capital of Accra. This sample, which was stratified for gender, age (6 groups), and Socio-Economic Status (3 levels), was assigned an average IQ of 80 by Lynn. The original source provides an average IQ of 82, but Lynn lowered this IQ estimate by two points because he aims to use an IQ of 100 for Britain as calibration. The 2-point correction was based on the fact that in the same study British adults from London scored an average IQ of 102. Lynn (2006) claimed that the average IQ of 80 for the inhabitants of Accra is "exceptionally high for sub-Saharan Africa" (p. 30). He explains this "high" average IQ by the fact that "the [Ghanaian] sample came from the capital city [and] people in capital cities typically have higher IQs than those in the rest of the country" (p. 30). However, average IQ scores of sub-Saharan Africans on the Culture Fair Test may be considerably higher than the average scores found by Buj in Ghana's capital. Namely, Nenty and Dinero (1981) administered this test to 803 students in seven secondary schools in both urban and rural areas in Nigeria. Interestingly, they found that these Nigerian adolescents scored on a par with a sample of 600 high school students from four schools in Portage County, Ohio. The average IQ on the Culture Fair Test in this large Nigerian sample was 98. In contrast to studies considered thus far, this study actually considered the possibility of measurement bias, which was studied using contemporary item response theory modeling. Some evidence for DIF was found, although the effects were not large and not necessarily in one direction. Lynn did not consider these data in any of his reviews of IQ in Africa.

Besides the DAM test, Fahmy (1964) administered additional tests to his sample of Sudanese children. The average IQ on these three tests was 94, 76.5, and 73.5, respectively. It appears that the average IQ of this sample given by Lynn was considerably lowered because of the unfamiliarity of these children with drawing, resulting in their low performance on the DAM test. Note also that the unweighted average IQ on the four tests should be 74, not the 69 that Lynn (2006) provides (in Lynn & Vanhanen, 2002 an IQ of 73.5 was given). As said, Fahmy considered the DAM test unsuitable for these children, so a fair estimate of IQ in this sample should be the average of the remaining tests (i.e., 81). This could still probably represent a considerable underestimation of these children's

cognitive capacity, because the administration of the remaining tests in Fahmy's study was all but successful.

Vernon administered a battery of IQ tests to fifty Ugandan boys of above average SES. On 16 of the 21 tests, mean IQ was above 80. The mean of the 21 subtests equals 86 (median 86). This becomes 88 when we leave aside the IQ of an English vocabulary subtest on which these boys scored very low (M = 57). Lynn gives an estimate of 80 for this sample, but provides no rationale for his downward correction. Vernon himself computed inter-subtest correlations in this sample and found no indication of a g factor comparable to that in other samples. A later factor analysis over part of Vernon's data by Hakstian and Vandenberg (1979), corroborated that "the cognitive structure among Ugandan subjects may be slightly different from that of other cultures" (p. 87). This is an interesting result, if only because several tests used by Vernon were also used in older studies in Africa.

In one of those old studies, fifty 5-13 year-old children from the Sousou tribe in rural Guinea were administered the Army Beta Test (Nissen, Machover, & Kinder, 1935). These unschooled test takers suffered from "inexperience in manipulating a pencil" (p. 325), which can be considered a serious handicap in taking the Army Beta. Moreover, on some subtests it was clear that most test takers did not understand what was expected from them. For instance, "[t]he subjects appeared utterly bewildered" (Nissen et al., 1935, p. 331) when confronted with the Manikin and Feature Profile subtest of the Army Beta. These difficulties notwithstanding, Lynn assigned this particular sample an IQ of 63. The Army Beta was also administered to 293 Black South-African children by Fick (1929). With respect to representativeness of samples, Fick clearly stated that "sweeping generalizations regarding whole groups should be avoided" (p. 904). He also acknowledged that the test scores may have been lowered due to the fact that "the native does not grow up with pictures and diagrammatic representations of things" (p. 909). In light of these difficulties, and because of the absence of *any* indication of the reliability, validity, or correlational structure of the Army Beta tests in this sample, we do not adhere to Lynn's assignment to this sample of an average IQ of 65. Another old study on the suitability of western intelligence tests among Black South Africans is that by Dent (1937). Dent considered his sample of 80 test takers too small for making any generalizations. With regard to the use of the Koh's Block test (the predecessor of the Block Design test in the Wechsler scales), Dent remarks that "all subjects experienced difficulty with this test" (p. 462). Difficulty with a test may either mean that the subjects did not understand instructions, or that their cognitive ability is low. Lynn apparently subscribes to the second option, and used the scores on this particular test to estimate the average IQ of this sample at 68 (which is a mental age IQ).

The studies reported in the last paragraph, which represent the dark past of IQ testing in Africa, were severely criticized as early as the 1940s (Biesheuvel, 1943), and cannot be taken seriously anno 2006. To begin with, the Army Beta test originates from the first years of intelligence testing and is now completely obsolete (Jensen, 1982 called this test "primitive"). More importantly, administering a paper and pencil test with pictures and diagrammatic representations to persons inexperienced with pencils and unfamiliar with pictures and diagrammatic representations does *not* provide a valid indication of intelligence. The situation is exacerbated by the fact that the pictures used in the Army Beta

are likely to be culturally biased because the pictures display typically American objects and situations. For instance, one item displays a tennis match, and test takers are required to draw the missing tennis net (Lane, 1994). These old papers are surely an interesting read for anyone interested in the invalid use of intelligence tests, and for those interested in the political role of psychology in the pre-apartheid era in South Africa (e.g., Krige, 1997). However, these old studies cannot be taken seriously by modern psychometric standards, certainly not to estimate average IQ of the African population.

Ferron (1965)[49], who states that Fick's "work is obviously biased" (p. 50), reports test results from an unknown IQ test in seven samples of children in Nigeria and Sierra Leone. Ferron considers this test unsuitable for African children. Despite this, Lynn included in his review the average IQs from the two lowest scoring samples, and briefly discussed (but did not include in his review table) a third sample with an average IQ of 80. Unfortunately, Lynn did not explain why he excluded the scores of the four higher scoring samples in Ferron's overview, such as a sample of 100 Sierra Leonean children who scored an average IQ of 93.

In several studies, sub-Saharan African children were administered the Wisconsin Card Sorting Test (WCST; Akande, 2000; Skuy et al., 2001; Sternberg et al., 2002). Note that this test is not meant to be a measure of general intelligence. In addition, none of these samples can be considered representative of a particular population. From one paper (Skuy et al., 2001), Lynn used only WCST data ("IQ of 64"), but did not include additional data from the DAM (IQ of 83) to estimate IQ. It is a commonly accepted that the use of more intelligence scores provide a more accurate estimate of general intelligence. Whereas he did use both the PMA test and the CPM test to estimate IQ of Fahrmeier's sample in his earlier work (Lynn, 1991), he excluded PMA (IQ=78) data in his later reviews (Lynn, 2003, 2006; Lynn & Vanhanen, 2002). Moreover, Lynn used WCST data from a study in Tanzania by Sternberg and colleagues (2002) to argue for a low IQ among Africans. Additional data of the WCST of Black South African children showed considerably higher average scores on this particular test (Akande, 2000), but Lynn did not consider these additional data.

Lynn also mentions data from the JAT in South-African Blacks (Lynn & Owen, 1994). At the level of the subtests, this test is severely biased, as is shown in a study by Dolan and colleagues (Dolan et al., 2004). In none of the samples considered by Lynn, was measurement invariance tested. Although we did not search for additional data on all remaining tests, additional data from the Culture Fair Test and the Wisconsin Card Sorting Test again shows considerably higher average IQ scores than the data Lynn has considered. No sample provided by Lynn in any way adds credibility to his claim that African IQ is below 70. Our review suggests strongly that additional tests show average IQ around or above 80 among African test takers. However, this still cannot be taken to mean that IQ scores are valid and free of measurement bias.

Differential Item Functioning

The question of measurement invariance is central to the question of the meaning of sub-Saharan IQ. Measurement invariance assures that the test measures the same

[49] This study is referred to by Lynn as Farron, 1966.

construct across groups. Measurement invariance is a starting point to understand the nature of group differences in test scores.

In a series of studies, Rushton and coworkers (Rushton, 2002; Rushton & Skuy, 2000; Rushton et al., 2004; Rushton et al., 2002, 2003) studied whether the Raven tests have similar item characteristics for Whites and Blacks in South Africa (cf. Owen, 1992). Rushton claims that these studies establish the construct validity for IQ tests among Africans (Rushton et al., 2004). Unfortunately, in none of these studies was DIF studied across groups. Instead, studies employed two straightforward methods to study biasedness of Raven's items. Central to the methodology employed by Rushton and Owen is the rank-order correlation between item p-values across groups (see also Mpofu & Watkins, 1994). This method is a simple method to study group differences in scale characteristics, which dates back to the 1920s (L. L. Thurstone, 1925). However, this method, and more refined methods based on it (e.g., the Delta-Plot method; Angoff & Ford, 1972) have been criticized extensively in the psychometric literature for not being sensitive to item bias. On the basis of their simulation study of the merits of various methods to detect bias, Shephard and colleagues conclude concerning the Delta-plot method that: "It should not be used for bias detection" (Shepard, Camilli, & Williams, 1985, p. 103). This method is simply incorrect when groups differ markedly in latent ability, and when items differ in discrimination parameter (Angoff, 1982; Ironson, Homan, Willis, & Signer, 1984; Lord, 1977, 1980; Shepard et al., 1985). In the comparison of African samples to western samples, group differences in test scores are generally large. Moreover, item analyses of the SPM and its advanced version (i.e., the Advanced Progressive Matrices or APM) have generally shown that items show considerable differences in discrimination parameter (Abad, Colom, Rebollo, & Escorial, 2004; J. C. Raven et al., 1996). Another method often employed in testing the suitability of Raven's tests in Africa employs the (point) biserial correlations.[50] This method is also shown to be rather suboptimal. "Using the classical point biserial item statistic and taking the discrimination differences between groups as a measure of bias appears to be inadequate" (Ironson & Subkoviak, 1979, p. 222).

Rushton uses yet another method (Rushton, 2002; Rushton & Skuy, 2000; Rushton et al., 2002, 2003), which may be seen as a combination of these two methods. In this third method, Rushton correlates the vector of group differences in item difficulty (i.e., group differences in p-values) with the vector of item-total correlations (i.e., point-biserial or biserial item-total correlations). In fact, this new method is an item-level equivalent of the method of correlated vectors (Jensen, 1998). This method has been shown to be problematic in factor analytic work (Dolan & Hamaker, 2001; Dolan et al., 2004; Lubke et al., 2001). Until this new item-level method is spelled out formally and investigated

[50] Rushton and Jensen (2005a) claim that the "item–total score correlations for Africans, Whites, and East Indians were also similar, indicating that the items measured similar psychometric constructs in all three groups" (p. 243). This statement is false both logically and empirically. The vectors of item-total score correlations in Whites, Blacks and East Indians will not be similar when groups differ in ability (Gulliksen, 1950). More importantly, in none of the studies mentioned, these vectors were similar across groups. For instance, the correlation between the vector of Whites and the vector of Blacks in Rushton's (Rushton et al., 2003) study of the APM among South African students equals 0.105 (pmcc) or 0.099 (rho). This means neither that the items measure similar constructs, nor that they measure something else. In fact, a comparison of item-total correlations does not adequately address the issue of measurement properties.

properly, we cannot tell whether or not it actually works to detect DIF. However, ingenious though it may seem, in the presence of group differences in latent ability and under most common IRT models, this method does not appear to work. First, in the presence of group differences in latent ability, the item-total correlations will *not* be equal across groups (excluding some special conditions under highly restrictive assumptions, which will certainly not hold in the SPM or APM; see, e.g., van der Ven & Ellis, 2000). Second, in all but a few special cases, the vector of these item-total correlations (that will differ across groups) will have a non-linear relation with the vector of group differences in item difficulty.[51] Even if a test is fully measurement invariant across groups, and the IRT model fits perfectly, this correlation between vectors will not equal one. Because we do not know how this correlation works under ideal conditions (i.e., equal item response functions across groups), we have no idea of how it will work in cases in which the test is severely biased across groups. Thus, the merits of this new method are unclear, but it does not show much promise.

The field of psychometrics[52] has provided a host of methods to detect DIF with crystal clear underlying assumptions and with well-established sensitivity to detect DIF (Holland & Wainer, 1993; Millsap & Everson, 1993). Unfortunately, *none* of these have been applied to the issue of African IQ (but see Nenty & Dinero, 1981). It is about time that rigorous methods to detect DIF were applied to shed some light on the meaning of IQ test scores in Africa. One cannot employ outdated or non-established methods that appear insensitive to bias, and reasonably conclude that bias does not exist. Such would be equivalent to claiming that a Petri dish is sterile, because no microorganisms are visible through a magnifying glass. The claim that IQ tests are unbiased with respect to Africans is simply baseless. Clearly, more research is needed to clear up the present obscure meaning of IQ test scores in Africa.

More on Validity

Lynn has estimated the average IQ of countries over the world and set out to validate his estimates of national IQ using data from several internationally comparable surveys of school achievement, in which representative samples of primary and secondary students were given Math and Science tests. In his latest book, he uses a combination of such studies given by Hanushek and Kimko (2000), in which the average IQs of Nigeria (IQ according to Lynn 69), Swaziland (IQ according to Lynn 68), and Mozambique (IQ according to Lynn 64) appear alongside that of 34 other countries[53] (Lynn, 2006). In Figure 5.2, we display the results of his validity study, which is typical of Lynn's validity studies. Lynn reports a correlation of 0.81 between these two variables and claims that this result

[51] Some preliminary computations using a scenario with established SPM item parameters in a 3 parameter logistic model and a large group difference in ability indicated that this relation has the shape of an inverted U. The results of Rushton's method depended greatly on the choice of group from which the item-total correlations were drawn. This is also apparent in the results of these studies (Rushton, 2002; Rushton et al., 2002, 2003).

[52] We are referring to rigorous psychometrics here (e.g., Hambleton & Swaminathan, 1985; Lord & Novick, 1968; Van der Linden & Hambleton, 1996).

[53] The original source also reports data from South-Africa, but Lynn did not include South Africa in this analysis. The data from South Africa would nevertheless also represent a bivariate outlier. Inclusion of South Africa lowers the correlation further to 0.77.

validates the estimates of national IQs. However, a look at the scatterplot suggests otherwise. In fact, this scatterplot shows clearly the presence of three outliers, which are the three data points on the low-left side. In the absence of these three data points, the correlation is 0.864. Incidentally, these three outliers correspond to the three countries from sub-Saharan Africa. These outliers have large negative residuals of over 13 IQ points, indicating that in the regression of IQ on Math and Science scores, the estimated IQs of these African countries is much higher than the IQs reported by Lynn. Lynn argues that the correlation of 0.81 is lowered by measurement error of the educational measures (begging the question of how an *average* score of several thousand test takers in each country would be affected by *random* measurement error). There is a more straightforward explanation for this result, namely that Lynn's estimates of national IQ in Africa are consistently too low. In fact, if anything can be learned from Lynn's validity study, it would be that the average IQ in these countries is around 80 or higher.

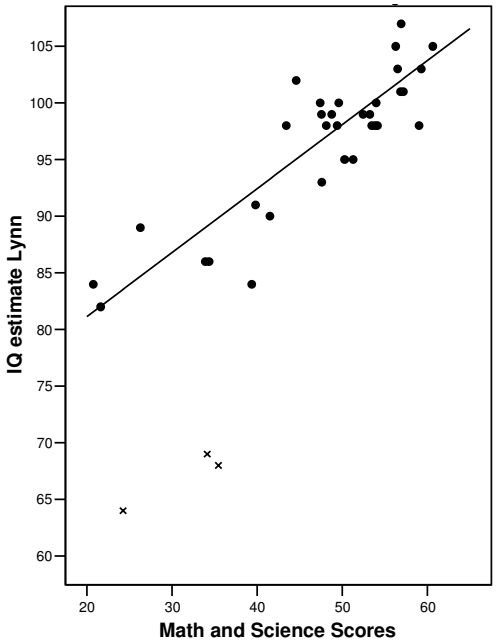

Figure 5.2 *Regression lines of Lynn's national IQ estimates on average Math and Science scores of international studies of student achievement (Hanushek & Kimko, 2000).*

African IQ Conclusion

Based on published data of the CPM and SPM, average IQ in Africa lies somewhere around 78 or 79, when compared to British norms. When compared to US norms, average IQ in Africa equals 80 or 81. There are several large samples in which IQ is considerably higher (Lloyd & Pidgeon, 1961; MacArthur et al., 1964; Nenty & Dinero, 1981). Despite the many measurement difficulties with the DAM, and the inaccuracy of its norms, the IQs on this test are higher than the IQs based on the CPM and SPM. We must again stress the importance of caution in the interpretation of these IQ scores. There have been no published IRT analyses that studied item bias of the SPM, CPM, and the DAM. It

is quite likely that these IQs represent an underestimation of ability, because with these tests even a few biased items can decrease IQ scores considerably.

The bulk of the data on which Lynn based his claim of low average IQ in Africa, is based on the following nine tests: K-ABC, WISC, WAIS, DAM, CPM, SPM, WCST, Culture Fair Test, and the IQ test discussed by Ferron. For all these tests, we came across additional African samples that showed markedly higher average IQs than the samples that were considered by Lynn. In none of the samples that were not included in Lynn's overviews, the average IQ was below 70. The literature missed by Lynn appears not to be missing completely at random (in the sense of Little & Rubin, 1986). In fact, the 31 samples in Tables 5.1-5.3 that Lynn considered in his latest review show significantly lower average IQs than the 27 samples he did not consider: $t(56) = 2.44$, $p < .05$. Lynn presented his review as a "fully comprehensive review" of the literature on African IQ. However, because he missed a sizeable portion of the relevant literature, his estimate of average IQ of Africans is too low.

The most serious omissions in the literature are rigorous tests of measurement invariance. In none of the samples used by Lynn IQ, measurement invariance was tested and found to be tenable, so the degree to which measurement bias has lowered IQ levels in African samples is unclear. The conclusion that the average IQ in sub-Saharan Africa is lower than average IQ in western countries is warranted, but the degree to which these low scores reflect lower general intelligence is unknown. These low scores might not reflect an accurate or valid assessment of general intelligence.

We found a clear indication of a Flynn Effect among adults on the SPM. Besides two studies (Daley et al., 2003; Richter et al., 1989), there is no clear indication of a comparable Flynn Effect among African children. The various samples are not ideal to study the Flynn Effect. The absence of gains among children may be due to the fact that the older samples are primarily of school-going children in times when school attendance in Africa was mainly restricted to higher SES levels. Hence, differences in sampling may be an issue. More comparable samples are required to shed some light on the Flynn Effect among African children.

In what follows, we are going to leave aside the measurement problems with IQ. The reason is that we would like to understand more fully the implications of these low scores if they would eventually prove to be accurate and valid. The main question we are left with is as follows: Suppose that an average IQ of 80 would be valid, would they lend credence to the idea that low African IQs are impervious to environmental variables?

5.6 Nature versus Nature & Nurture

In his most recent book, Lynn (2006) claims, as did Rushton (2000b), that racial differences in general intelligence have evolutionary causes (cf. Jensen, 1998). Rushton's (2000b) evolutionary theory supposes that the races differ in their reproductive strategies, causing racial differences in intelligence. Lynn's theory states that ancestors of Europeans (i.e., Whites) and Asians have developed higher genetic intelligence during their evolutionary struggle to survive in colder and therefore more demanding climates outside of Africa. According to Lynn, people from European and Asian descent are more

intelligent than people from Africa because the latter group did not encounter a similar evolutionary pressure towards high intelligence in the relatively warm climates of Africa. The theories by Lynn and Rushton stand in stark contrast with archeological evidence that clearly shows that people in sub-Saharan Africa have been as advanced as Eurasians (MacEachern, 2006). In a recent study that was criticized by Hunt and Sternberg (2006), Templer and Arikawa (2006) set out to substantiate Lynn's evolutionary theory. To that end, Templer and Arikawa correlated Lynn and Vanhanen's estimates of average IQ in 81 countries all over the world to the average temperature in these countries and to an estimate of national skin color.[54] They found a correlational structure that, according to them, substantiated Lynn's theory. However, any claim to causality using correlational data requires the consideration of confounding variables. There exist a host of alternative explanations for national differences in IQ. To these, we turn next.

Ecological Correlations

In this section we correlate Lynn & Vanhanen's (2002) national IQ estimates with several environmental variables that are suggested to play a role in the Flynn Effect. Various studies have shown that these national IQ estimates correlate with national wealth as expressed in Gross Domestic Product (GDP) per capita (Dickerson, 2006; Jones & Schneider, 2006; Lynn & Vanhanen, 2002; Morse, 2006; Weede & Kampf, 2002; Whetzel & McDaniel, 2006). It is noteworthy that several authors have claimed that the relation between national IQ and GDP is nonlinear (Dickerson, 2006; Morse, 2006; Whetzel & McDaniel, 2006). This nonlinearity may be partly due to the fact that IQ of African countries is underestimated considerably by Lynn and Vanhanen. This results in (impossible) negative predictions of GDP on the basis IQ, which, in turn, results in a non-linear relation between these two variables. Nevertheless, it is clear that mean IQ levels across the world are related to economic development.

In several studies, the correlation between national IQ as estimated by Lynn and Vanhanen and adult literacy rate was found to be around .70 (Barber, 2005; Meisenberg, 2004; Morse, 2006), suggesting that national differences in school attendance are related to national IQ levels. The relation of IQ levels to health variables at the national level is less clear. Barber (2005) found moderate correlations between national IQs and several health variables, but his analysis suffered from missing data. In addition, Whetzel and McDaniel (2006) found a correlation of 0.56 between Lynn and Vanhanen's estimates of national IQ and health expenditure per person. The many variables related to social and economic development are probably not only related with national IQ, but will also show strong intercorrelations. Therefore, it is worthwhile to integrate several variables in one analysis. Because we are mainly concerned with the Flynn Effect, we focus on variables that have been proposed to have caused the Flynn Effect.

[54] The study by Templer and Arikawa is concerned with the evolution of intelligence, yet uses contemporary data on temperature to study this. Moreover, the measurements of skin color and temperature in that study do not stand up to scientific scrutiny. The temperature estimates are based on a weather guide for travelers, which is unsuitable for estimating national temperature. The skin color estimates are based on students' judgments of a 65-year old skin color world map of 10 by 5 inches (Biasutti, 1959). This outdated (Robins, 1991) and inaccurate (Coon, 1966) map was based on a subjective method to measure skin color, which has been in scientific disrepute since the 1950s (Jablonski, 2004).

Method

The results provided below are based on analyses not weighted by size of populations of countries. Although such an N-weighted analysis has a small effect on the some of the correlations, this alternative analysis does not alter our main conclusion. We now discuss our choice of variables to consider, and we provide descriptions of the data employed.

National IQ. We employ Lynn and Vanhanen's estimates of national IQ in 81 countries over the world, excluding Equatorial Guinea (for obvious reasons given above) and Taiwan because of missing data. To enable a comparison to the literature, we used Lynn and Vanhanen's original IQ estimates. We also computed more accurate IQs for African countries based on IQs in Tables 5.1, 5.2, and 5.3, which we corrected for the Flynn Effect in a way similar to that employed by Lynn and Vanhanen (cf. Footnote 32). For instance, in the case of Nigeria we used a weighted average based on eleven studies which results in an IQ of 76 for this country. For countries not included in Tables 5.1-5.3, we added 10 IQ points to Lynn and Vanhanen's estimates, because this is approximately the underestimation of IQ in African countries in Lynn and Vanhanen's list.

Nutrition. Poor nutrition during childhood is generally considered to lower adult IQ (e.g., Sigman & Whaley, 1998). An improvement in nutrition has been suggested as one of the prime reasons for the Flynn Effect (Lynn, 1989, 1990). The data on nutrition were retrieved from the Food and Agriculture Organization (FAO), an agency of the United Nations. We use three nutrition variables that are averages per capita of calories per day, proteins in kg per day, and fat in kg per day. We averaged the numbers over the years 1985-2000, but the use of data from alternative or separate years does not greatly alter the results provided below.

Health. Poor health is generally considered to have a negative effect on IQ (e.g., Mackintosh, 1998), and improvements in health have been proposed as an important contributor to the Flynn Effect (W. M. Williams, 1998). We used three indicators of a countries' health status. These are under five mortality rate, maternal mortality rate, and neonatal mortality rate. The under five mortality rate was estimated by UNICEF for the years 1990-2003. The neonatal and maternal mortality rates are estimates from the WHO for the year 2003. The use of data of alternative years does not have a large effect on the correlations we provide below.

Education. Education has been suggested to be an important factor in the Flynn Effect (Barber, 2005; Ceci, 1991; Husén & Tuijnman, 1991; Tuddenham, 1948). We use data from UNESCO of gross enrollment ratio in primary and secondary education, as well as estimates of teacher-to-student ratio within each country. All educational variables are averages over the period 1970-2003, but the use of data from alternative or separate years does not greatly alter the results provided below.

Computers. The introduction of computers and computer games may have enhanced test-specific skills, contributing to the Flynn Effect (Greenfield, 1998). We use estimates of the number of computers per 1000 inhabitants over the period 1998-2002, provided by the International Telecommunication Union (ITU) and retrieved from the World Bank database.

Family size. It has been suggested that the trend towards smaller families has also been partly responsible for the Flynn Effect (Zajonc & Mullally, 1997). Fertility rate per country was retrieved from the World Development Index. We took the averages over years 1970-2003, but the use of data from separate years does greatly not alter the results provided below.

Urbanization. The transition from a rural to a (sub)urban society has also been suggested as a cause of the Flynn Effect, because of a decrease in inbreeding depression (Mingroni, 2004) and an increase in environmental complexity (Dickens & Flynn, 2001; Schooler, 1998). Urbanization estimates for 2005 were retrieved from World Health Organization tables, but the results provided below are robust to the use of data from alternative years.

Water quality. The lack of improved drinking water and sanitation may have a negative effect on health, which may negatively affect cognitive development through the effects of intestinal parasites (Boivin et al., 1993). Because increases in improved drinking water and sanitation in the developing world are both part of the millennium goals of the UN, the UNICEF has estimated these variables for 2002. We used the data for that year.

Results

The ecological correlations between Lynn & Vanhanen's (2002) IQ estimates (excluding Equatorial Guinea and Taiwan) with the environmental Flynn Effect variables are given in Table 5.4. Because data did not exist for all countries on some variables (particularly water quality variables), missing values (6.3 % of all data points) were imputed by using multiple imputation with the program PRELIS (Jöreskog & Sörbom, 2003). The correlations based on the imputed data are however similar to those computed using pair-wise deletion. The correlations in Table 5.4 are a textbook example of multicolinearity. All environmental variables correlate highly and significantly ($p < 0.001$) with IQ, and (with one exception) significantly ($p < 0.005$) with each other. As a matter of fact, a principal components analysis on these 15 variables results in one highly dominant principle component, as can be seen by the scree plot in Figure 5.3. This first principle component explains 74 % of the variance. This dominant component is nothing more than developmental status of countries. The loadings on this component are given in the last row of Table 5.4. Viewed in this light, IQ is just another indicator of development. On *all* the variables in Table 5.4, sub-Saharan African countries fall on the negative side of the world wide distribution. It is also apparent from Table 5.4 that the use of more accurate estimates of national IQs does not have a large effect on the correlations between IQ and the exploratory variables.[55]

[55] The correlation between our adjusted estimates of IQ and the estimates of GDP for the year 1998 (from Lynn & Vanhanen, 2002) for the 79 countries equals 0.77 ($p < .01$). This is slightly higher than the correlation between GDP and Lynn and Vanhanen's estimates of national IQ ($r = 0.75$). The exponential relation between our estimates of national IQ and GDP does not add much to the linear relation (i.e., r^2 increases by .08) indicating that the nonlinear relation between GDP and national IQ (Dickerson, 2006; Morse, 2006; Whetzel & McDaniel, 2006) is indeed partly due to the underestimation of African IQs by Lynn and Vanhanen. Note also that GDP correlates highly with all the variables in Table 4.

Table 5.4

Correlations between estimates of national IQ with explanatory variables (N=79)

	IQ	Prim. educ. enroll.	Sec. educ. enroll.	Pupil-teach. ratio	PCs per 1000	Fer-tility	Urba-niza-tion	% impr. sanit.	% impr. water	Ch. mort. rate	Neona tmort. rate	Mat. mort. rate	Cal. /day cap.	Prot. g/day cap.	Fat. g/day cap.
IQ	1	.426	.737	-.665	.700	-.817	.630	.658	.607	-.699	-.721	-.626	.688	.730	.731
Prim.educ.enrollment	.518	1	.540	-.376	.124*	-.492	.454	.555	.610	-.667	-.572	-.617	.450	.364	.372
Sec. educ.enrollment	.784	.540	1	-.765	.682	-.846	.659	.801	.757	-.817	-.827	-.759	.757	.791	.827
Pupil-teacher ratio	-.719	-.376	-.765	1	-.522	.764	-.575	-.850	-.741	.731	.718	.677	-.734	-.743	-.767
PCs per 1000 persons	.656	.124*	.682	-.522	1	-.617	.573	.516	.482	-.472	-.559	-.382	.538	.621	.675
Fertility	-.860	-.492	-.846	.764	-.617	1	-.628	-.761	-.729	.853	.868	.759	-.711	-.745	-.734
Urbanization	.666	.454	.659	-.575	.573	-.628	1	.654	.611	-.626	-.625	-.588	.609	.588	.604
% Improved sanitation	.727	.555	.801	-.850	.516	-.761	.654	1	.879	-.827	-.815	-.750	.713	.714	.764
% Improved water	.702	.610	.757	-.741	.482	-.729	.611	.879	1	-.831	-.771	-.719	.747	.715	.699
Child mortality rate	-.811	-.667	-.817	.731	-.472	.853	-.626	-.827	-.831	1	.932	.916	-.694	-.705	-.676
Neonatal mortality rate	-.790	-.572	-.827	.718	-.559	.868	-.625	-.815	-.771	.932	1	.818	-.677	-.700	-.714
Maternal mortality rate	-.763	-.617	-.759	.677	-.382	.759	-.588	-.750	-.719	.916	.818	1	-.694	-.670	-.617
Calories/day per cap.	.728	.450	.757	-.734	.538	-.711	.609	.713	.747	-.694	-.677	-.694	1	.935	.872
Proteins g/day per cap.	.757	.364	.791	-.743	.621	-.745	.588	.714	.715	-.705	-.700	-.670	.935	1	.875
Fat g/day per cap.	.712	.372	.827	-.767	.675	-.734	.604	.764	.699	-.676	-.714	-.617	.872	.875	1
Loading on 1st PC	.884	.560	.926	-.891	.701	-.944	.776	.904	.839	-.929	-.933	-.865	.837	.867	.872

Note: Correlations below diagonal are based on Lynn & Vanhanen's IQ estimates, correlations above diagonal adjusted IQs; All correlations $p < .005$, except $*p > .05$

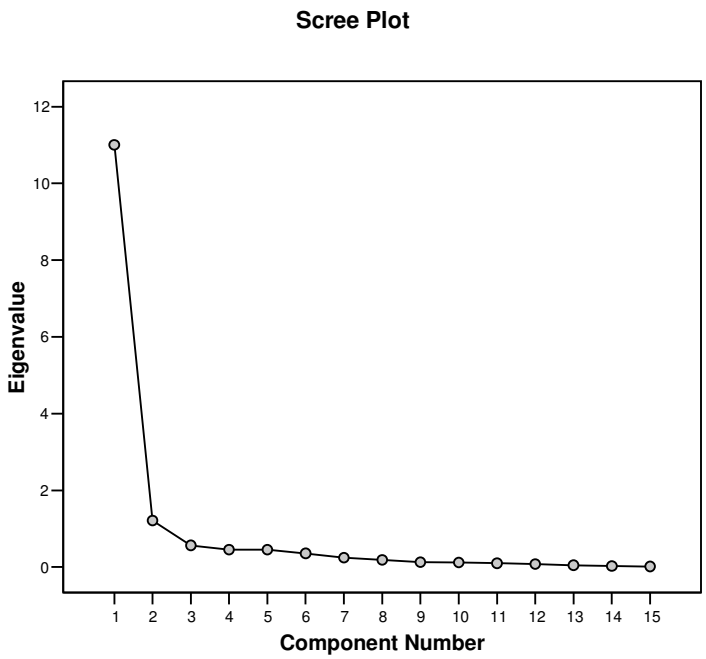

Figure 5.3 *Scree plot of principal components analysis of variables in Table 5.4.*

Conclusion

These correlations indicate that of countries across the globe, environmental variables that have been proposed to have caused the Flynn Effect are also the variables that have yet to show improvements in Africa. Because of the strong relation between these environmental variables and average national IQ, the claim that low levels of IQ in Africa are due to genetic factors (Lynn, 2006; Rushton, 2000b; Templer & Arikawa, 2006) is hard to maintain.

As is the case with any cross-sectional study employing ecological correlations, we can not claim to have established any causal relation of the variables in Table 5.4. It is certainly the case that these data are consistent with a host of environmental explanations of why average IQ in Africa is around 80, as opposed to 100 in developed countries. The comparison of African countries to developed countries is fraught with confounds. In light of all these confounding variables, any claim to causality needs to be made very carefully. Unfortunately, such caution is all but absent in Lynn and Vanhanen's claim that the wealth of nations is caused by intelligence levels of a population. In view of Table 5.4, one could equally claim that insufficient computers, insufficient food, an unhealthy population, poor schooling, etc. are responsible for the fact that some countries are poorly developed economically. Such variables have generally been ignored or dismissed as irrelevant in studies claiming that there are evolutionary reasons related to climate that cause lower IQ in countries in Africa (Lynn, 2006; Templer & Arikawa, 2006).

Because on all variables in Table 5.4, Africa is on the negative end, it is safe to assume that these variables might have a depressing effect on IQ levels. Schooling, health, nutrition, and urbanization are serious confounds in the comparison of IQs across the world. When viewed in this light, national IQ does not appear to be anymore than just another indicator of the development of a country.

5.7 General Discussion

Several conclusions can be drawn on the basis of our review of the literature on African IQ. First, the proposition that Africans have an average IQ of 67 is untenable. It appears to lie somewhere around 80, and it is likely to be even higher. Second, the claim that African IQ scores are comparable to western IQ scores in terms of the construct of general intelligence or *g* has to date not been substantiated by rigorous data analyses. Third, low IQ levels in Africa are not surprising in light of the fact that environmental variables that are believed to suppress IQ levels are omnipresent in Africa. Fourth, the Flynn Effect has occurred in Africa among adults on the SPM. Fifth, as the environmental effects on IQ will continue to improve, they will almost certainly raise IQ levels in Africa in the years to come. Sixth, our results do not sit well with the genetic theory of race differences in intelligence as put forth by Lynn. We will now focus on each of these conclusions more closely.

The Dark Past of African IQ

There has been a long history of IQ testing in Africa. In some instances IQ tests were administered under non-standard circumstances to test takers that were so unfamiliar with IQ tests and the material in it (Dent, 1937; Fick, 1929; Nissen et al., 1935), that their scores cannot and should not be used to claim anything concerning latent cognitive ability (Biesheuvel, 1943). It is about time we leave that dark past of IQ testing in Africa behind us.

It is apparent that average IQ in Africa is lower than average IQ in western countries. However, average IQ in Africa does not appear to be as low as Lynn maintains. The majority of studies on African IQ not taken into account by Lynn showed considerably higher IQs than the studies he reviewed over the years. Lynn's reviews of IQ in sub-Saharan Africa are skewed, and have resulted in an underestimation of average IQ in Africa. Clearly, Lynn has missed a large part of the literature on African IQ. However, in several cases (Crawford Nutt, 1976; Ferron, 1965; Irvine, 1969b; MacArthur et al., 1964; Pons, 1974; Skuy et al., 2001), he must have been familiar with additional data, for the simple reason that he used these sources in his own work. It is unfortunate that Lynn did not discuss his reasons to exclude these additional data. Without knowing his reasons, it would not be fair to jump to conclusions with respect to Lynn's scientific integrity (e.g., Kamin, 1995). For instance, in some cases, tests were administered with additional instruction (Crawford Nutt, 1976; Pons, 1974). We felt it reasonable to include these data because this instruction is highly similar to an instruction as described in the test manual (J. C. Raven et al., 1996), but some have argued that this instruction artificially heightens test performance (cf. Rushton & Skuy, 2000). Nonetheless, it is safe to say that Lynn's

conclusion that average IQ in Africa is around 67 is based on unsound reviews of the literature.

The Obscure Present of African IQ

An IQ score should never be equated uncritically with a particular level of general intelligence because IQ tests are fallible instruments, particularly for test takers less familiar with western culture as reflected in these IQ tests. IQ testing in Africa is a complicated issue (MacArthur et al., 1964; Nell, 2000). Based on our reading of the literature, validity studies of the DAM, SPM, or CPM in Africa provide little support that these tests provide accurate assessments of *g*. More importantly, the degree to which IQ differences between countries in any way reflect national differences in general intelligence is unknown. The reason is that the comparison of test scores across such various groups as westerners and Africans requires standardized testing conditions and measurement invariance across groups. The testing conditions in Africa are not always ideal. More importantly, measurement invariance between western and African samples has not been studied using contemporary methods such as multi-group confirmatory factor analysis (with mean structure) or DIF analyses based on IRT models. Where it has been studied rigorously, results have shown that measurement invariance is rejected (Dolan et al., 2004; Nenty & Dinero, 1981). Therefore, there is a real possibility that IQ averages in sub-Saharan African samples are an underestimate of latent cognitive ability.

The average IQ scores of Africans that we documented in our review are nothing more than average transformed scores on measurement instruments that we call IQ tests. These tests may be well-validated in developed countries, but they are not well-validated in African countries. Further research should shed light on what these test scores may or may not mean. The true meaning of IQ scores differences between western samples and African samples only becomes clear after thorough psychometric modeling. What is required is a study in which testing conditions across groups are controlled and in which it is ascertained that test instructions are crystal clear to all test-takers. This study should involve a battery of tests each of which can be studied for DIF. After that, one can establish that between group mean differences are on the (higher order) latent factor called *g*, by employing multi-group factor analysis with mean structure (Dolan, 2000; Dolan & Hamaker, 2001; Lubke et al., 2003a). Suppose we would establish that tests are fully measurement invariant, and that between-group differences are mainly (or entirely) due to between-group differences in *g*. All we know then is that we have tackled the enormous measurement problem. This opens the door to study of why groups differ in this latent variable we call *g*, and we can study which reasons may lie behind the group difference in *g*. Until that day, we do not know what group differences in IQ scores mean, and evolutionary, genetic, and environmental theories with respect to race differences in intelligence only have a very weak empirical foundation. In addition, evolutionary and genetic theories also should take into account the fact that global differences in national IQ are strongly correlated with a vast number of environmental variables that are known or at least suspected to be responsible for the Flynn Effect.

The Bright Future of African IQ

An average IQ of 80 among Africans may appear to be low, but from a historical perspective this average is not low at all. That is, when we compare the SPM scores of a representative sample of British adults in 1948 to British norms collected in 1992 (J. C. Raven, 1960; J. C. Raven et al., 1996), average adult British IQ in 1948 would be approximately 81. Likewise, compared to the test performance of Dutch 18-year-olds in 1982, a sample containing 79% of all 18-year-old Dutch males in 1952 has an average IQ of 80 (Flynn, 1987). Despite the supposedly "low average IQ" of their populations around 1950, Great Britain and The Netherlands developed fairly well economically, scientifically, and culturally in the last five decades. In fact, the average IQs (based on more recent norms) of samples around 1950 turned out so low *because* these countries developed since 1950! Therefore, the average IQ of Africans would be close to 100 if we would have compared SPM and CPM performance to British norms of around 1950. This is evident in Figure 5.1, where we compared SPM scores of Africans to older norms. In this figure, the average IQ of several African samples is clearly above 100. Note that in terms of societal development, contemporary African countries are more similar to developed countries in 1950 than in 2006.

The rise in average intelligence test scores over the years has been shown to occur in most developing countries over the world (Flynn, 1984, 1987, 1999c; Neisser, 1998). The fate of intelligence test scores in Africa should not be cause for pessimism, because there is much room for improvement of IQ levels in Africa. Whethzel and MacDaniel (2006) indicated that countries (with low average IQ) could improve their IQ levels by encouraging high IQ individuals to procreate and discouraging low IQ individuals from procreation. That appears to be a very slow and highly inefficient way to improve IQ levels in countries like Sierra Leone, where more than 25% of children die before the age of five, because of malnutrition and disease. Improving education, health care, sanitation, and nutrition would seem to be a better idea. Luckily, these are also variables that the UN aims to have improved before the year 2015, as formulated in the so-called Millennium Goals (United Nations, 2005). As we saw, it is safe to say that there has also been a rise in IQ scores in sub-Saharan Africa. There is a lot of empirical support for the claim that malnutrition (Sigman & Whaley, 1998), health (W. M. Williams, 1998), sanitation (Boivin et al., 1993), and schooling (Ceci, 1991) have an effect on IQ. When the Millennium Goals will be accomplished, IQ levels in Africa will surely go up.

What, in terms of the exploratory variables we studied, is the potential of the Flynn Effect in Africa? The average infant mortality rate in current day sub-Saharan Africa is about 84. This is comparable to the infant mortality rate in 1920 in the US. Urbanization in Africa is about 40%, which is comparable to urbanization in the US around 1900. Fertility rate in Africa is comparable to the fertility rate in the US in 1870. The average pupil-to-teacher ratio in primary schools in current-day sub-Saharan Africa roughly equals that in the US before 1910. Thus, in terms of the variables that have been proposed as causes of the Flynn Effect, people in sub-Saharan Africa grow up under circumstances that are comparable to a western civilization before the First World War. Future will tell whether average IQ in sub-Saharan Africa will show gains similar to those found in western countries. Either way, given that the Flynn Effect has stood at about 3 points rise in IQ per

decade in the developed world, the Flynn Effect has a potential of at least 27 IQ points (i.e., 90 years worth of Flynn Effect) in sub-Saharan Africa. Anyone who claims that African IQ is low because of genetic or evolutionary factors, should take this simple fact into account.

Admixture

The results of our study do not sit well with theories that assign a substantial role to genes in racial differences in intelligence. It has been argued that because most African Americans are admixtures of European and African genes, that the IQ of "pure blacks" in Africa should be much lower than the average IQ among African Americans (Lynn, 1991; Rushton & Jensen, 2005a). This appears not to be the case. The average IQ of African Americans has been around 85 for quite a long time (Gottfredson, 2005; Jensen, 1998; Rushton & Jensen, 2005a), although it appears to have changed upwards to approximately 89 in recent years (Dickens & Flynn, 2006). Based on our extensive review of studies, average IQ of the African population lies somewhere in the neighborhood of 80, when compared to a mean IQ of 100 for the US. We are left with a mean difference somewhere from 5 to 9 IQ points between Africans and African Americans. If we take into account the real possibility that African IQ represents an underestimation of ability because of measurement bias and sub-optimal testing conditions, this difference is likely to be smaller. Were one to correct for the large differences in environmental circumstances (e.g., education, nutrition, health) between those two groups, this difference could easily drop to zero. This would falsify the genetic theory of racial differences in intelligence, as put forth by Lynn (2006). In light of our results, the admixture argument in favor of the genetic theory of race differences in intelligence is unconvincing.

Concluding Remarks

Controversial topics such as group differences in IQ should not deter researchers, but should encourage better research (Hunt & Carlson, 2006). Group differences in IQ exist, whether one likes it or not. The fact remains that science has an important role to play in understanding these group differences. One does not learn much by claiming that IQ tests are simply unsuitable for Africans (e.g., Berry, 1974), or that race differences in IQ are not worthy of study (Sternberg, 2005). Some have argued that IQ tests are suitable for Africans (Lynn, 2006; Rushton & Jensen, 2005a), and particular social and political conclusions are drawn on the basis of (incorrect) IQ levels in Africa (e.g., Herrnstein & Murray, 1994; Lynn & Vanhanen, 2002; Rushton & Jensen, 2005a). Scientists do not contribute to knowledge by claiming that certain persons are racist (Kamin, 1995), or that people are being too politically correct to see the truth (Rushton, 1996). Scientists contribute to knowledge by doing what they are good at, namely conducting rigorous, fair, and open research. Besides, the exposure of erroneous claims (e.g., that Africans have an average IQ of 67) is an empirical issue, not a matter of a priori belief. We certainly hope that our study has shed some more light on the complicated issue of IQ scores in sub-Saharan Africa. Regardless of what these scores may eventually turn out to mean.

6

Discussion

6.1 Introduction

"Does IQ have a future? The short answer is: no" (Bartholomew, 2004, p. 33).

The study of cognitive abilities requires the use of measurement models. In this chapter, I will highlight the merits of using such models by focusing on an idealized (measurement) model of intelligence. Practical issues often hinder the implementation of this idealized model. I will argue that the analysis of IQ test scores in the absence of an explicit measurement model cannot add much to our understanding of group differences in cognitive ability. In addition, I will discuss the results of the approaches employed in the studies of this thesis, as well as the results of other approaches, in light of this idealized model.

6.2 Idealized Model of Cognitive Abilities

Figure 6.1 displays a simplified hierarchical model of cognitive abilities, on which there is considerable consensus in the literature (e.g., Carroll, 1993; Jensen, 1998; McGrew, 2005). On the top or apex of this model is the second order factor called *g* or general intelligence. Below *g* is a particular first order factor, which is influenced by g, and by other factors independent of *g*. This first order factor influences the two narrowly defined latent traits, both of which again are also subject to other factors, independent of the first order factor. The two narrow latent traits are each measured by a collection of items composing a scale (i.e., subtest). The item characteristic curves of these items are displayed on the bottom of the figure. Suppose that *g*, the first order factor, and the narrow traits are all linearly related, and suppose that the dichotomous item scores conform to an unidimensional[56] Rasch model. The model is far from complete. As Carroll (1993) has shown, there might be an intermediate level between the first order and second order factor, and there exist several first order factors (e.g., crystallized intelligence, processing speed, long term retrieval, etc.). Despite the incompleteness of this model, Figure 6.1 illustrates how complicated the accepted inter-individual structure of human cognitive abilities actually is. Things are further complicated by the fact that none of the depicted variables in Figure 6.1 are directly observable; they are latent traits. Even the item characteristic curves require estimation by fitting an Item Response Theory (IRT) model on dichotomous item scores, which are the only observed variables.

[56] Note that the hierarchical model might be inconsistent with the unidimensional IRT model. This may be solved by employing multidimensional IRT models.

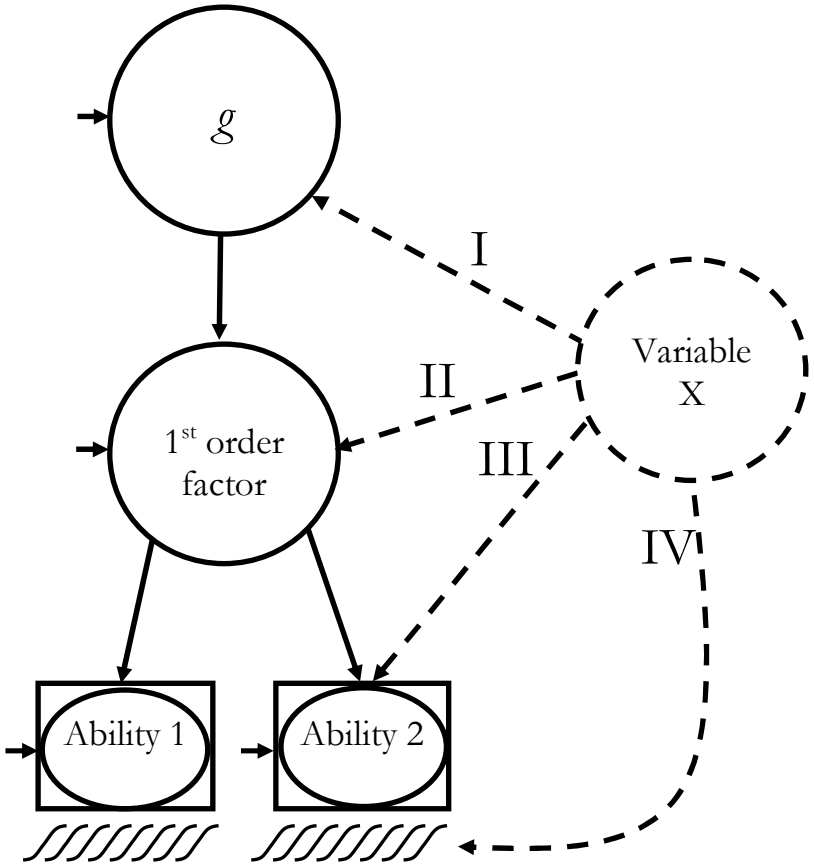

Figure 6.1 *Idealized model of cognitive abilities and different effects of Variable X*

As any model, the model in Figure 6.1 requires empirical verification, which can be accomplished by first fitting the Rasch model on item scores of a sufficiently large sample of test takers. This results in item parameter estimates and ability estimates. Subsequently, these ability estimates (or the sufficient statistics in case of a multivariate normally distribution) can be used as input in a confirmatory factor analysis.[57] With the program M-Plus (Muthen & Muthen, 2003) this analysis can be conducted in one run. Fitting of the model can shed light on the dimensions of inter-individual differences, and on the merits of the measurement model. Fitting the entire model is not always feasible for practical reasons. For instance the IRT model is often not fitted, but the item scores are summed to arrive at scale scores. This was the approach employed in the studies in the current thesis, where we focused on the factorial structure of subtest scores. Note that the summation of item scores is not ideal, but may provide a reasonable approximation, provided that the number of items is sufficiently large.

[57] More 1st order factors and indicators are required to identify the factor model, but this is immaterial to the current discussion.

Now consider the exogenous Variable X, upon which the variables in the model are regressed. X may be a continuous variable, like the amount of intellectual stimulation during childhood, test sophistication, or the additive influence of a large number of genes. X may also represent group membership (e.g., race, cohort, sex), although in this case it might be more appropriate to not speak of a causal effect, but rather of a correlation between X and the variables in the model. As depicted in Figure 6.1, Variable X can affect (or be related to) the variables in the model at four levels: (I) The higher order factor called *g*, (II) the first order factor, (III) a particular narrow trait represented by a subtest, and (IV) the location of individual item characteristic curves. Suppose that the regression on X of the variables at Levels I, II, and III is linear, and that the regression on the dichotomous item is suitably linearized (e.g., probit or logit regression). Thus, a Level IV effect is such that it affects the difficulty parameter of particular items. The difference between these four levels is highly relevant to the understanding of the (causal) relation between X and cognitive abilities. For instance, if X affects item parameters, this amounts to uniform bias (Mellenbergh, 1982) or Differential Item Functioning (DIF) with respect to X in the Rasch model. In that case X is related to the measurement of cognitive ability, but not to any of these abilities themselves. Hence, an effect on Level IV may be considered a measurement artifact. If X affects the narrow trait (i.e., Level III effect), this would mean that the effect of X is limited to the unique ability tapped by a subtest. If X represents group membership, such an effect at Level III implies the presence of an intercept difference across groups (cf. Chapter 2). Such an effect may be seen as a measurement artifact, but it may also be interesting in its own right (cf. Chapter 3 and 4). In addition, it makes quite a difference, both theoretically and practically, whether X affects the first order factor (i.e., Level II effect) or the second order *g* factor (i.e., Level I effect). For instance, if intellectual stimulation during childhood affects the first order factor (e.g., crystallized intelligence), this would imply that intellectual stimulation has an effect limited to this particular first order cognitive ability. On the other hand, if intellectual stimulation during childhood affects *g*, this effect generalizes to other first order factors and narrow abilities, which are affected directly or indirectly by *g*.

6.3 The IQ approach

Figure 6.2 displays an approach to study cognitive ability denoted the IQ approach. In this approach, IQ is used as a proxy for *g*. The IQ approach looks neat and tidy, but looks can be deceiving. In fact, in the IQ approach, all test score information is wiped onto a big pile (i.e., the summation of item scores), and denoted by the catchall term IQ. In reality IQ is a hodgepodge of first order factors, narrow traits, item scores, and *g*.

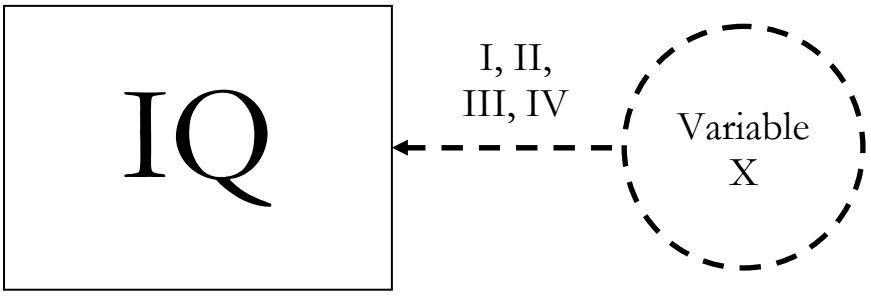

Figure 6.2 *The IQ approach and the effect of Variable X on cognitive ability.*

The use of IQ in the study of cognitive ability is quite common. For instance IQ is used to study intelligence in Africa (Lynn, 2006; cf. Chapter 5), sex differences in intelligence (Jackson & Rushton, 2006), the relation between intelligence and brain size (Thoma et al., 2005), and the Flynn Effect (Dickens & Flynn, 2001). In addition, a variant of the IQ approach is the dominant approach in experimental research, where summed item scores are generally treated as if they were latent variables. The IQ approach has led to much misunderstanding. For instance, the lay public generally sees IQ as synonymous with *g*. Even intelligence researchers sometimes make such mistakes. For example, the following sentence appears to confuse directions of causality. "IQ determines the efficiency of learning and comprehension of all cognitive tasks" (Lynn & Vanhanen, 2002, p. 39).

The IQ approach does not do justice to the complexity of human cognitive abilities, nor is the IQ approach appropriate for the difficult task of measuring these. This concept of IQ ignores the fact that first order factors and subtests invariably measure additional traits besides *g*, and that IQ may not be a good indicator of *g*. *g* may make a large contribution to the variance of IQ scores, but *g* is certainly not the whole story. For instance, a(2003) confirmatory factor analysis (Carroll, 2003) of the US standardization sample of the 29 cognitive ability tests of the Woodcock-Johnson-Revised (WJ-R; Woodcock & Johnson, 1989) showed that of the total test score variance, 33 % could be attributed to *g*, 22 % to nine first order factors, and 45 % to subtest specific factors and measurement error.

The use of the IQ does not contribute very much to our understanding of the nature and causes of cognitive abilities (see also Bartholomew, 2004), or to our understanding of the nature of group differences in intelligence test scores. That is, the (causal) relation between Variable X and IQ can be due to the effect of X on g (i.e., Level I), the effect of X on the first order factors (i.e., Level II), direct effects of X on subtest specific ability (i.e., Level III), or DIF with respect to X (i.e., Level IV). In other words, Level I, II, III, and IV effects are all confounded in the IQ approach. For instance, the relation between intelligence and educational attainment appears to be rather more complicated (Dolan et al., 2006) than expected on the basis of previous research that used IQ. Even when a particular IQ test (e.g., Raven's Progressive Matrices) has a high loading on *g*, this does not mean that the effects of X on other levels are irrelevant. Group

differences in IQ cannot be simply dubbed group differences in *g*, just because IQ is based on IQ test scores (cf. Chapter 5).

6.4 Analytical Approaches

Given the hierarchical model, an appropriate approach to the study of cognitive abilities is based on statistical techniques from item response theory and/or Structural Equation Modeling (SEM) approaches, like Confirmatory Factor Analysis (CFA). These approaches are not always employed in the study of cognitive abilities. There exist several approaches to study intelligence that are intermediate to the idealized modeling approach described above (cf. Figure 6.1) and the IQ approach (cf. Figure 6.2). Most of these intermediate approaches are applied for practical reasons, although in many instances the use of less sophisticated models is not warranted. For instance, consider the widely used Method of Correlated Vectors (MCV; Jensen, 1998). In this method, subtests' factor loadings on *g* are estimated by means of Exploratory Factor Analysis (EFA), Principal Components Analysis (PCA), or principal axis factor analysis (i.e., a variant of PCA). Subsequently, the subtests' *g* loadings are correlated with subtests' correlations with Variable X. In the model underlying MCV, two variables of cognitive ability remain: a causal effect of (or a group difference in) Variable X is either on *g* (i.e., Level I), or on all other variables in the model (i.e., Levels II, III, and IV). In other words, a *g*-or-not-*g* conceptualization underlies the method of correlated vectors. If the correlation that forms the crux of this method is close to one (i.e., a "Jensen Effect"; Rushton, 1998), this is interpreted as an indication that X is related to *g*. Any correlation larger than 0.50 is generally seen as an indication of the rather vague notion that Variable X is mostly related to *g* (e.g., Lynn & Owen, 1994). Likewise, a correlation of zero between *g* loadings and a subtests' correlations with Variable X are interpreted as if *g* is not correlated with X at all (e.g., Rushton, 1999). This, however, is not necessarily the case (Ashton & Lee, 2005). There are several reasons that the method of correlated vectors is suboptimal, the most important being a lack in specificity (Dolan, 2000; Dolan & Hamaker, 2001; Dolan et al., 2004; Lubke et al., 2001; Lubke et al., 2003a). Specifically, with MCV it is not possible to disentangle effects on the different levels in which X can be related to cognitive ability and/or test scores. Whenever feasible, the use of SEM approaches with latent variables (Bollen, 1989) is preferred. With SEM X's relation to the different levels in Figure 6.1 can be studied from a statistically sound perspective. If X represents group membership, Multi-Group Confirmatory Factor Analysis or MGCFA is preferred, for the simple reason that MGCFA models approach the idealized model much more closely. If the factorial structure of a battery of subtests is unclear, Multi-Group Exploratory Factor Analysis (MGEFA) can be employed (Hessen et al., 2006). In most applications of the method of correlated vectors, the presence of sufficient statistics and sufficiently large sample sizes allow for the use of MGCFA or MGEFA. Hence, in such cases, there is no reason to use the method of correlated vectors, as this method is suboptimal.

There are many more approaches that are intermediate to the IQ approach and the idealized model, such as Principal Components Analysis (PCA), multiple linear regression, and equivalent techniques such as analysis of variance (ANOVA). However, the application

of these techniques does not do justice to the fact that cognitive abilities are latent variables that underlie test scores (e.g., PCA). When group differences in intelligence test scores are studied, MANOVA approaches can be used, but these are only related to mean differences. Most importantly, MANOVA, PCA, and regression analyses do not comprise a measurement model. For example, Camarata and Woodcock (Camarata & Woodcock, 2006) recently used ANOVA to study sex differences on a large battery of cognitive ability tests. However, this approach does not identify the exact level on which group differences lie. The use of MGCFA or MGEFA allows for a more parsimonious approach, which may shed light on the nature of sex differences at the level of common factors (Dolan et al., 2006; Van der Sluis et al., 2006).

6.5 MGCFA

In the studies in Chapters 2, 3, and 4, we employed multi-group confirmatory factor analysis, but we did not consider item level data. Therefore, in the approach used in these studies, the effects of X on Level IV and Level III were confounded. Variable X represented ethnic groups and cohorts in Chapters 2 and 4, respectively. In Chapter 3, X represented ethnic groups/sex groups and experimental conditions. In these applications of MGCFA, we were mainly interested in the degree to which group differences in subtest scores could be attributed to group differences at the level of the common factors (i.e., Levels I and II). This also requires that measurement parameters are invariant with respect to X.

When we came across mean effects that could not be accounted for by the common factors in the model, measurement invariance with respect to X was said to be violated in an uniform manner (Mellenbergh, 1982). Because of the confounding of Levels III and IV, the uniform measurement bias we came across in these studies may have been due either to DIF or to group differences in ability unique to subtests. Level I and II effects were also not always distinguishable in the data sets we analyzed in this thesis, for the simple reason that the number of subtests and factors did not always suffice to estimate the higher order factor structure. Nevertheless, in the first study of Chapter 4, the second order factor was modeled in the comparison of two cohorts of test-takers. That particular study showed, like other studies (Dolan, 2000; Dolan et al., 2006; Dolan & Hamaker, 2001; Van der Sluis et al., 2006) that the disentanglement of Level I and Level II effects is not always straightforward empirically. This is due to insufficient power, i.e., the difficulty to detect differences in mean structures at Level I and II. On the other hand, the results of our applications of MGCFA suggested that effects on Level III/IV effects can be readily distinguished from effects on Levels I and II. That is, we came across Level III and IV effects in most of the data sets in this thesis. These effects are called intercept differences.

6.6 Intercept Differences

Chapter 2 focused on the disentanglement of Level I/II from Level III/IV effects when X represents group membership. In our overview of MGCFA studies published in 2005, we found that it is quite common that between-group difference in test scores are not

solely related to the level of the (first order) common factors. This is a requirement for group comparisons at Levels I and II, because the model is a bottom-up model, at least in an empirical sense (i.e., in terms of estimation). In Chapter 2, we argued that within MGCFA, the absence of intercept differences is a necessary condition for measurement invariance. Measurement invariance is central to the understanding of group differences in test scores. Therefore, it is rather surprising that in many studies with MGCFA, the possibility of intercept differences is simply ignored, even in cases where the mean structure was modeled explicitly (e.g., Chirkov et al., 2005; Corwyn & Bradley, 2005; Hagger et al., 2005; McInerney et al., 2005).

As we saw in the illustrative re-analysis of a published study of the suitability of an IQ test for ethnic minority children in The Netherlands (Te Nijenhuis, Tolboom et al., 2004), ignoring Level III/IV effects may have serious practical consequences. It is noteworthy that Te Nijenhuis and colleagues also studied DIF of several subtests of this IQ test (i.e., they did study Level IV effects). Interestingly, the subtest that showed the largest intercept difference in our factor analyses (cf. Figure 2.5), did *not* show item level bias in their DIF analyses. This indicates that Level III effects may indeed be present in the absence of Level IV effects. The effect of ethnicity on this subtest, which measured knowledge of Dutch vocabulary, is a clear example of a Level III effect. One of the two other subtests (i.e., Learning Names) that showed an intercept difference did display DIF (the third biased subtest was not suitable for DIF analyses), which suggests that the intercept difference on this subtest was probably due to a Level IV effect. DIF in this subtest may have been due to the fact that it contained Dutch names from various fairy tales, with which the ethnic minority children may have been less familiar. Yet another subtest of this IQ test showed considerable DIF with respect to ethnicity, but did not display an intercept difference. Combined, these results indicate that Level IV and Level III effects are distinct, and that Level IV effects may or may not show up as intercept differences at the subtest level. Therefore, both analyses at the item level and analyses on the level of subtests are required to fully establish measurement invariance across groups.

The RAKIT test appears to be biased with respect to ethnic minority children in The Netherlands. Other research has shown that the GAT-B in the Netherlands was also biased with respect to minorities at Levels III/IV (Dolan et al., 2004), and the WAIS-III in Spain and The Netherlands was biased with respect to females at this level (Dolan et al., 2006; Van der Sluis et al., 2006). In IQ test development, the use of MGCFA (if applied at all; see, e.g., Wechsler, 2000) is mostly restricted to testing group differences in factor loadings, but these tests do not provide reassurance whether tests are measurement invariant across groups. Considering the apparent omnipresence of intercept differences, and the likelihood of DIF, there is a strong need for more research on measurement invariance of IQ tests across demographic groups. The claim that "the issue of test bias is scientifically dead" (Hunter & Schmidt, 2000, p. 151) seems to be divorced from reality.

Ideally, studies of group differences should include data on a large battery of tests with a clear theoretically based underlying factorial structure (e.g., WJ-III test battery; Woodcock, McGrew, & Mather, 2001), item level data, and covariates that could help explain the possible group differences on different levels. Several important unresolved issues in the study of intercept difference are (1) the disentanglement of Level III and Level

IV effects, (2) power to detect intercept differences when more than one indicator of a factor are affected by a biasing variable, and (3) effects of nonlinearity.

6.7 Stereotype Threat

The effects of stereotype threat on test performance are generally seen as measurement artifacts (Steele, 1997). Stereotype threat theory (Steele et al., 2002) states that the performance lowering effects of stereotype threat are mainly restricted to items and subtests that are sufficiently difficult to be stereotype threatening. That is, on an easy task one does not run the risk to conform to the stereotype of low performance. Furthermore, the performance on easy tasks is not likely to be strongly affected by decreases in working memory capacity, which is considered an important mediator of stereotype threat effects (Schmader & Johns, 2003).[58] Within the model in Figure 6.1, the effects of stereotype threat may be seen as Level III and Level IV effects on the most cognitively demanding subtests and items, respectively.

In the studies into the effect of stereotype threat on test performance in Chapter 3, we employed basic one-factor models with subtest scores as indicators. The theory of stereotype threat allowed for quite specific predictions of the effect of stereotype threat (i.e., Variable X in Figure 6.1) on test performance. We predicted and found that the experimentally induced effects of stereotype threat were most pronounced on the most cognitively demanding subtests. Most, but not all, effects of stereotype threat we found were linear and resulted in intercept differences.

In the first study of Chapter 3, stereotype threat had a non-linear effect on test performance. In many (albeit not all; Lubke et al., 2003b) circumstances, such non-linear effects can also be detected readily by means of MGCFA. Note, that the factor models employed in Chapter 3 were quite small. In such models it is not always possible to pinpoint exactly the subtests that display uniform or non-uniform bias. Suppose that in a one factor model with three indicators, the first subtest shows misfit after a particular between-group restriction is implemented. This effect could be due either to bias in this first subtest, or to biasing effects on the other two subtests. Ultimately, theoretical arguments guide the identification of biased subtests, not solely indicators of model misfit. Ideally, one would incorporate in the model covariates that could explain the bias.

The modeling approach in Chapter 3 showed the usefulness of MGCFA in experimental settings. The experimental paradigm in psychology focuses strongly on mean effects, while covariance effects are often ignored (but see Baron & Kenny, 1986). In addition, in experimental psychology manifest test scores are generally viewed as latent variables, and individual differences are usually not modeled (i.e., they act as error terms in ANOVA; Cronbach, 1957). As we showed in Appendix C of Chapter 3, the use of analysis of covariance (ANCOVA) to accommodate individual differences in experiments does not always sit well with predictions derived from theories that relate to individual differences (e.g., stereotype threat theory). The use of MGCFA in experiments allows for the use of a

[58] It would be interesting to study the effects of stereotype threat on working memory capacity from a modelling perspective. Note that when the effect of stereotype threat is related to a common factor representing working memory, this effect represents a Level II effect.

measurement model of both the mean and covariance structure. With this approach measurement artifacts on Level III/IV can be disentangled from effects on the level of common factors (i.e., Levels I and II). In addition, the use of MGCFA allows for a test of measurement invariance across design cells. Therefore, the use of measurement models in experimental settings could greatly enhance the construct validity of experiments (Shadish, Cook, & Campbell, 2002). Not only would the use of such measurement models allow for the detection and correction of many methodological artifacts (e.g., demand characteristics in self-report questionnaires), it would also shed more light on the exact nature of the latent dependent variables and of the nature of causal effects on these variables.

The studies in Chapter 3 illustrated the usefulness of rigorous modeling in experimental settings and in our understanding of stereotype threat. Future work on stereotype threat could look at item level effects (e.g., Stricker & Bejar, 2004). Also, it would be interesting to employ more elaborate factor models, and to include covariates that can shed light on mediating and moderating variables. Stereotype threat theory states that not all test-takers are equally susceptible to the effects of stereotype threat, but this theory (like many psychological theories) is not very explicit in whether stereotype threat susceptibility is a latent class or a latent trait. If stereotype threat susceptibility turns out to be a latent class, an analytical approach to study stereotype threat effects on test performance would be to use factor mixture analyses (Lubke & Muthen, 2005), which could be used in both experimental and non-experimental settings.

6.8 The Flynn Effect

Chapter 4 was concerned with the Flynn Effect. The large gain in IQ test scores is quite remarkable given its size and consistency over time and over populations. However, the apparent consistency of the effect over the developed world is mostly a function of the use of IQ to document the effect. The fact that the summed scores of a battery of tests (i.e., IQ) increase over the years can be due to different causes raising scores on different levels of the idealized model. It is quite conceivable that a large portion of the gain is caused by Level III effects on different narrow abilities. Only with rigorous modeling, can we hope to understand the nature of this phenomenon.

The continued use of IQ in the study of the Flynn Effect is remarkable, because early on it was noted that the gains were dependent on the type of subtest (Flynn, 1987). Differential increases have raised the question whether the gains can be related to an increase in *g* (Colom & García-López, 2003; Colom et al., 2001; Flynn, 1999a, 1999b, 2000a; Jensen, 1998; Must et al., 2003; Rushton, 1999, 2000a). This discussion revolved mainly around the method of correlated vectors, which does a poor job in disentangling the effects on the different levels of the idealized model.

The results from the studies in Chapter 4 shed some light on the level at which the Flynn Effect appears to be operating. It became clear that Level III/IV effects (both positive and negative) were present in the comparison across cohorts, although the gain was also related to Levels I and II.

Proposed causes for the Flynn Effect differ in the level of effects within the idealized model. Gains in test sophistication (Brand, 1987) and improvements in test

specific skills (Greenfield, 1998) may be seen as Level III and Level IV effects. Such causes are consistent with the results of Studies 1, 2, and 4 in Chapter 4. Urbanization (Barber, 2005), greater environmental complexity (Schooler, 1998), improvements in health care (W. M. Williams, 1998), a trend towards smaller families (Zajonc & Mullally, 1997) are most likely Level I and Level II effects. Increases in educational attainment (Husén & Tuijnman, 1991; Tuddenham, 1948), betterment of educational practice (Blair et al., 2005), are likely to be effects on Level II (e.g., crystallized ability) and Level III (e.g., math ability). The working of gene by environment correlation in the increasing presence of more intelligent others (Dickens & Flynn, 2001), the genetic effect of heterosis (Mingroni, 2004), and improvements in nutrition (Lynn, 1989, 1990), are Level I effects. If the absence of measurement invariance across cohorts proves to be robust, it follows that variables related to Levels I and II cannot be the sole causes of the Flynn Effect.

Measurement invariance in the studies in Chapter 4 was rejected mostly due to the presence of intercept differences across groups. If we consider the arguments put forth in Chapter 2 and in Appendix B of Chapter 3, these intercept differences would imply that at least part of the Flynn Effect is related to uniform effects on Levels III and IV. The uniformity means that whichever variable has caused the gain at this level, it does not interact (strongly) with latent ability and it does not correlate strongly with latent ability. What kind of variable could this be? Likely variables are variables that large portions of the population encounter, such as the introduction of the television, computer games, and toys (Greenfield, 1998).

Figure 6.3 displays two examples of children's toys, which strongly resemble, and might even have been copied from, particular subtests in the Wechsler Adult Intelligence Scale (WAIS) and the Wechsler Intelligence Scale for Children (WISC). The toy on the left resembles the Block Design test, the toy on the right is almost identical to the Object Assembly test. Note that such toys are disseminated widely since the 1960s. In fact, most primary schools in the Netherlands have toys like these. Such "educational" toys may have provided excellent test coaching that may have contributed to the Flynn Effect on the WAIS and WISC. Note that in the US both these subtests have shown consistent and relatively strong gains from 1947 to 2002 (Flynn, 2006).[59]

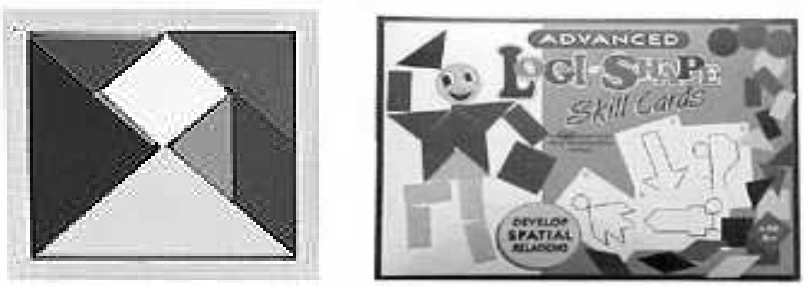

Figure 6.3 *Two smart types of toys.*

[59] Note that these two subtests showed only moderate gains in the Dutch data we analyzed in Chapter 4. In addition, these subtests did not show intercept differences in these analyses.

The Flynn Effect may in large part be due to increases in the specific ability tapped by such subtests, which would constitute a Level III effect. Level IV effects are equally likely. For instance, the Vocabulary subtest of the US WAIS contained an item asking for the meaning of the word "terminate". It is quite conceivable that since Schwarzenegger's 1984 film *The Terminator*, this item has become considerably easier.[60] A study of DIF could shed light on this issue. Note that the changes over time in item parameters is well-known in educational measurement, where the effect is known as item parameter drift (Chan, Drasgow, & Sawin, 1999). With one notable exception (Flieller, 1988), DIF analyses have not been used to study the nature of the Flynn Effect. The results of Flieller's study clearly showed DIF over time.

Future work on the Flynn Effect should focus on data at both the scale and item level. In addition, modeling of covariates that could explain the gain, may shed light on the causes of the Flynn Effect. Unfortunately, existing raw data sets, particularly those including item scores, are often difficult to acquire (Wicherts, Borsboom, Kats, & Molenaar, 2006). Nevertheless, measurement models are the key to our understanding of this fascinating phenomenon. Although there is an indication that the Flynn Effect appears to have stopped in some western countries (Teasdale & Owen, 2005; Wicherts, 2005a), the Flynn Effect will in all likelihood continue in the developing world, such as in Africa.

6.9 IQ in Africa

In Chapter 5, we were guilty of the use of IQ scores for the simple reason that item level data and multivariate test scores were not available. In light of the lack of results from rigorous psychometric modeling, it is unclear what IQ scores in African samples mean psychometrically. The mean IQ difference between western samples and African samples in scores on an intelligence test such as Raven's Progressive Matrices may be on Level I, II, III, and/or IV. Given the likelihood of measurement bias (see, e.g., Dolan et al., 2004), caution needs to be entertained in the interpretation of IQ scores in Africa. Rushton and Lynn (Lynn, 2006; Rushton et al., 2004) maintain that the difference in IQ scores between western samples and African samples lie on *g* (i.e., Level I). However, this is mere speculation. The analyses employed by Rushton and coworkers (Rushton, 2002; Rushton & Skuy, 2000; Rushton et al., 2004; Rushton et al., 2002, 2003) do not establish the level at which these groups differ. A rigorous study of DIF with well-established methods would be a good starting point in the study of the psychometric meaning of IQ test scores in Africa. To gain insight into the factorial nature of these test scores, analyses using MGCFA or MGEFA are also needed.

There exist a host of variables that could account for the relatively low performance of Africans on the Raven's tests. Considering the many cultural differences between Africans and westerners, several of these variables are likely to be on Levels II, III, and IV. For instance, relatively low test sophistication (Irvine, 1966) and misunderstanding

[60]Popular film titles could also have a negative effect on WAIS performance. The WAIS-III Information subtest contains an item asking for information on The Kremlin. In a sample of 416 Psychology freshmen from the University of Amsterdam, no less than 15 students indicated that this is a small, cute, furry creature that turns into a vicious monster at midnight. This error is clearly caused by Spielberg's 1984 film *The Gremlins*.

of test directions (MacArthur et al., 1964) among African test takers could both result in performance lowering effects on Level III.[61] Individual items could display DIF (i.e., a Level IV effect) because these items contain figures that are less familiar to Africans (Bakare, 1972). It is also conceivable that the Raven's test measure spatial abilities in addition to *g* (Irvine, 1969b), and that spatial abilities among Africans are less well developed by televisions and computer games (Greenfield, 1998).

The genetic or evolutionary theories of race differences in intelligence test scores invariably relate to Level I effects. However, group differences in *g* are not likely to fully explain the relatively low IQ scores of Africans. Besides, as we saw, there are many variables related to the Flynn Effect that have not shown gains in Africa comparable to those that occurred in the developed world. A large part of these environmental variables are related to Level I, and could also help explain the reasons for lower IQ among Africans.

In conclusion, the genetic or evolutionary explanations of low average IQ in Africa have a weak empirical basis, because these theories relate to a group difference in *g* which has not been established, and because they are based on correlational evidence in the presence of many highly relevant confounding variables.

6.10 The Nature of Latent Traits

In basically all theories on cognitive ability, cognitive abilities, including *g*, are conceptualized as normally distributed latent variables. Although latent cognitive variables are interesting and valuable in their own right, more research is required to shed light on the exact psychological nature of these variables. The presence of latent dimensions of inter-individual differences in cognitive ability may or may not be a reflection of latent cognitive processes (Borsboom, Mellenbergh, & van Heerden, 2003). For instance, van der Maas and colleagues (2006), recently proposed a dynamic model of cognitive abilities that could account for the positive manifold (i.e., the phenomenon that cognitive ability test scores universally intercorrelate positively) in the absence of a single cognitive quantitative biological process or capacity. Their model provides an interesting view on the nature of cognitive abilities, and could explain many phenomena such as the Flynn Effect. Dynamic models like that presented by van der Maas et al. illustrate the usefulness of formal models in the study of cognitive abilities.

The cognitive processes underlying inter-individual differences in intelligence test performance can also be studied from an experimental perspective. Unfortunately, experimental cognitive psychology and the study of individual differences in cognitive ability appear to be as mutually isolated as they were fifty years ago (Cronbach, 1957). Nonetheless, work on individual differences in working memory capacity (e.g., Engle, 2002; Engle, Tuholski, Laughlin, & Conway, 1999) does appear to show some promise. The continued use of ANOVAs by experimental psychologists, and the continued use of basic correlational techniques by differential psychologists will not help in bringing the two disciplines of psychology together. Bridging the gap between the experimental approach

[61] The effect of such misunderstanding on item performance could also depend on the nature of particular items. In that case, misunderstanding may result in a Level IV effect.

and the individual differences approach to study human cognition rests ultimately on the use of rigorous statistical models (Embretson & Schmidt McCollam, 2000; Lohman, 2000), which should be well grounded in psychological theory.

6.11 Conclusion

Measuring latent variables by means of IQ tests is not an easy task, but the field of psychometrics has provided many tools to study the relation between test scores and the latent traits that are supposed to underlie those test scores. The aim of this thesis was to use one such psychometric tool (i.e., MGCFA) to gain a better understanding of group differences in intelligence test scores. Measurement models should be an integral part of theorizing in all psychological theories that are related to latent traits. However, theories are not always explicit concerning the level at which the effect of exploratory variables lie. If our ultimate aim is to understand human cognitive abilities, and their determinants, the approaches based on IQ do not take us very far. The more approaches are based on explicit statistical and psychometric models, the closer we get in understanding cognitive abilities, their antecedents, and group differences in intelligence test performance. Cognitive abilities are complex phenomena that we will never fully understand by using approaches based on IQ, or by using simple heuristics such as the *g*-or-non-*g* conceptualization underlying the method of correlated vectors. In the study of cognitive abilities, simplistic analytical approaches are best abandoned.

7

Appendix:
A cautionary note on the use of information fit indices in covariance structure modeling with means

Information fit indices such as AIC, CAIC, BIC and ECVI can be valuable in assessing the relative fit of structural equation models that differ with respect to restrictiveness. In cases where models without mean restrictions (i.e., saturated mean structure) are compared to models with restricted (i.e., modeled) means, one should take account of the presence of means, even if the model is saturated with respect to the means. The failure to do this can result in an incorrect rank order of models in terms of the information indices. We demonstrate this point by an analysis of measurement invariance in a multi-group confirmatory factor model.

7.1 Introduction

Often in structural equation modeling, a sequence of increasingly restrictive models is fitted. When both means and covariances are modeled, the situation may arise in which one first fits a series of models to the observed covariance matrix, and one subsequently adds the model for the means. Such a stepwise approach has the advantage that it provides information concerning the drop in fit when structured means are added. This is especially important when the means and the covariance structure are modeled with a common subset of parameters, i.e., when strong hypotheses are tested concerning the common causation of individual and mean differences (e.g., Mandys, Dolan, & Molenaar, 1994; Meredith, 1993). The aim of the present note is to point out that in the calculation of information criteria and the expected cross validation index (ECVI) in this context one should take account of the presence of means, *even if* the model is saturated with respect to the means. The failure to do this can result in an incorrect rank order of models by AIC (Akaike, 1974), BIC (Schwarz, 1978), CAIC (Bozdogan, 1987), and ECVI (Browne & Cudeck, 1989, 1993). Specifically the rank order is incorrect when going from a model in which the model for the means is saturated to a model in which the means are constrained. We identify this problem below and demonstrate it in an illustrative analysis.

7.2 Assessment of Relative Fit Using AIC, BIC, CAIC, and ECVI

In assessing the fit of structural equation models, it is advisable to consider several fit measures, in addition to the χ^2 index (Bollen & Long, 1993). Information criteria such as AIC, CAIC, and BIC form a useful class of indices, as they penalize for the number of parameters, and thus take into consideration the parsimony of models. Although the information statistics have rather different origins, varying from the concept of entropy (AIC) to Bayesian statistics (BIC), they all have a similar structure (see Table 7.1), in that they involve the same information. Lower information index values indicate better fit. We note that the ECVI (Browne & Cudeck, 1989, 1993) is linearly related to the AIC, and thus yields the same rank order of competing models as the AIC.

Information statistics are valuable in analyses, where models without restrictions on the mean[62] are compared to models with such restrictions. However when means are unrestricted, one may be inclined to discard the means. Clearly means need not actually be included in a model, in which the means are not structured. Moreover in certain cases (exploratory factor analyses), it is difficult to actually include the means. However, comparing models that do restrict means to models that do not, these implicit mean parameters have to be considered in the computation of the information indices. If these parameters are overlooked, the information indices are underestimated. This in turn may result in the unjustified rejection of restrictions on the means. The underestimation caused by ignoring the parameters for the means differs for each information criterion, and depends on the number of manifest variables and the number of cases. Table 7.1 contains expression for this underestimation for each information criterion. We illustrate our point by testing for factorial invariance in two groups of children.

Table 7.1
Fit indices and underestimation due to ignoring saturated means

Fit Index	Formula	Underestimation due to ignoring saturated means
AIC	$= \chi^2 + 2t$	$2*p$
CAIC	$= \chi^2 + (1 + \ln N)t$	$(1 + \ln N)*p$
BIC	$= \chi^2 + (\ln N)t$	$(\ln N)p$
ECVI	$= (\chi^2/n) + 2(t/n)$	$2*(p/n)$

Note: t=number of parameters; p=number of manifest means; N=number of cases; n=N-number of groups.

7.3 Illustration: Factorial Invariance

The psychometric theory concerning the definition and meaning of measurement invariance within the context of the common factor model (i.e., factorial invariance) is well developed (Meredith, 1993). This theory gives rise to multi-group confirmatory factor models, in which covariance and mean structures are restricted over groups. Here we

[62] I.e., a saturated mean structure in which a parameter is estimated for each observed mean.

compare two groups. Let μ_i and Σ_i denote the implied mean vector and covariance matrix in group i. These are modeled as follows:

$$\mu_i = \tau_i + \Lambda_i a_i, \quad (1)$$

$$\Sigma_i = \Lambda_i \Psi_i \Lambda_i' + \Theta_i, \quad (2)$$

where the (p x q) matrix Λ_i contains factor loadings, and the p-dimensional vector τ_i contains measurement intercepts. The (p x p)-diagonal matrix Θ_i contains unique/error-variances, and Ψ_i is the (q x q)-covariance-matrix of the q common factors. Finally, a_i is a q-dimensional vector of factor means. For reasons of identification (see Sörbom, 1974) this vector is fixed to zero in an arbitrary group, so that latent differences in means are modeled. Factorial invariance can be tested by fitting a series of increasingly restricted models. These are presented in Table 7.2.

Table 7.2

Summary of models in case of two Groups 1 and 2

No.	Description	$\Sigma_1=$	$\Sigma_2=$	$\mu_1=$	$\mu_2=$
0	Exploratory	$\Lambda^*_1\Lambda^{*\prime}_1+\Theta_1$	$\Lambda^*_2\Lambda^{*\prime}_2+\Theta_2$	τ_1	τ_2
1	Configural invariance	$\Lambda_1\Psi_1\Lambda_1'+\Theta_1$	$\Lambda_2\Psi_2\Lambda_2'+\Theta_2$	τ_1	τ_2
2	Metric invariance	$\Lambda\Psi_1\Lambda'+\Theta_1$	$\Lambda\Psi_2\Lambda'+\Theta_2$	τ_1	τ_2
3	Equal error/unique variances	$\Lambda\Psi_1\Lambda'+\Theta$	$\Lambda\Psi_2\Lambda'+\Theta$	τ_1	τ_2
4a	Strict factorial invariance	$\Lambda\Psi_1\Lambda'+\Theta$	$\Lambda\Psi_2\Lambda'+\Theta$	τ	$\tau+\Lambda a_2$
4b	Strong factorial invariance	$\Lambda\Psi_1\Lambda'+\Theta_1$	$\Lambda\Psi_2\Lambda'+\Theta_2$	τ	$\tau+\Lambda a_2$

Note: Λ^*_1 denotes that all elements are estimated. Except for Step 4b (nested under 2) each model is nested under the previous one.

In addition to an exploratory factor analysis, we fit three models without mean restrictions, namely configural invariance (equal pattern of factor loadings), metric invariance (equal factor loadings; Horn, McArdle, & Mason, 1983), and a model with group-invariant error/unique variances. Furthermore, we fit two models with structured means, denoted strong factorial invariance and strict factorial invariance (Meredith, 1993). Meredith (1993) has shown that, within the factor model, strict factorial invariance is required to demonstrate measurement invariance (i.e., unbiasedness) with respect to groups. To illustrate our point we fit these models and calculate the indices with and without taking the means into account.

The models are fitted on a subset of data published in Naglieri and Jensen (1987), which comprise the K-ABC and WISC-R scores of 86 Black and 86 White children. We first carried out an exploratory factor analysis (EFA) on selected 16 subscales (see Dolan & Hamaker, 2001, for similar analyses of the complete dataset). This resulted in a simple structure with three common factors relating to verbal abilities (V), spatial abilities (S) and memory (M). The scales are: Information (loading on the factor V), Similarities (V), Vocabulary (V), Comprehension (V), Picture Completion (S), Picture Arrangement (S), Block Design (S), Object Assembly (S), and Digit Span (M) from the WISC-R, and Faces and Places (V), Riddles (V), Reading/Understanding (V), Triangles (S), Hand Movement (M), Number Recall (M), and Word Order (M) from the K-ABC. In subsequent confirmatory analyses, we use this simple structure. We fix one factor loading per factor at

1 for scaling purposes. Furthermore, we assume multivariate normality and estimate parameters by Maximum Likelihood (ML).

Table 7.3
Fit indices of models with or without means

					means excluded in 0-3				means included in 0-3			
model	DF	χ^2	p	RMSEA	ECVI	AIC	CAIC	BIC	ECVI	AIC	CAIC	BIC
0	150	177.5	.062	0.032	2.39	407	913	791	2.77	471	1110	956
1	202	233.4	.064	0.020	2.05	349	639	569	2.43	413	836	734
2	215	250.5	.049	0.028	2.02	344	580	523	2.40	408	777	688
3	231	294.1	.003	0.044	2.06	350	520	479	2.44	414	717	644
4a	244	311.3	.002	0.041	2.35	399	648	588	2.35	399	648	588
4b	228	267.4	.038	0.026	2.31	393	708	632	2.31	393	708	632

Note: The χ^2 reported here is the minimum fit χ^2, whereas the slightly different Normal Theory Weighted Least Squares χ^2 is used here (like it is in LISREL) for computation of the information indices.

The fit indices of the models are presented in Table 7.3. For comparison we also report the χ^2's and RMSEA fit indices, which are unaffected by the presence or absence of means in saturated mean models (i.e., models 0-3). We first consider the χ^2 indices. Given the nesting of the models, we employ χ^2 differences as a significance test for each restriction (Jöreskog, 1971). This would lead us to conclude that the equality of unique/error variances over groups is not tenable, but that the other between-group restrictions do not lead to a significant increase ($p < .05$) in χ^2. Based on the χ^2, we therefore conclude that strong factorial invariance holds. Note that the RMSEA does not really help in selecting models. Given the rule of thumb that RMSEA< 0.05 represents a reasonable approximation (Browne & Cudeck, 1993), all models are judged to be acceptable. In view of the equivocality of RMSEA, and given the recommendation to consider a variety of indices (Bollen & Long, 1993), we now turn to the information criteria.

Here we first look at the case in which means are *not* incorporated in the model, i.e., models 0-3. Based on both the ECVI and the AIC, we would conclude the equality over groups of error/unique variances is not tenable and, more importantly, that intercepts cannot be equated across groups. The latter also applies to BIC and CAIC, although these two indices indicate that error/unique variances are invariant across groups. Thus, when the saturated mean structure is ignored, ECVI, AIC, CAIC and BIC lead to the incorrect conclusion that both strong and strict factorial invariance should be rejected. Only when the parameters for the means are taken into account (even though they are unconstrained), do we draw the correct conclusion. Here strong factorial invariance does hold, whereas strict factorial invariance would be rejected (e.g., compare the ECVI and AIC in Model 4a and Model 4b).[63]

[63] However, note that the BIC and CAIC suggest that strict factorial invariance is tenable.

7.4 Conclusion

The use of information criteria such as AIC, BIC, and CAIC, and the ECVI is valuable in the comparison of structural equation models that differ with respect to restrictiveness. However, when mean structure is analyzed in addition to the covariance structure this mean structure should be incorporated in the models at all stages of model fitting, even when the mean structure is saturated (unrestricted). Failure to do so may result in an incorrect rank order of models, and incorrect conclusions. Happily the correct value of the criteria can be obtained by including the means in the input and model specification.[64] In situations where this may not be possible (e.g. exploratory factor analysis) the correct value can be calculated readily by hand (see Table 7.1). Although we have focused on factorial invariance in our illustration, this conclusion applies to other models including structured means such as the latent growth curve model or (quasi-)simplex models with structured means (e.g., Mandys et al., 1994). Finally we note other fit indices (e.g., the various comparative fit indices, such as the non-normed fit index) and related information (standardized residuals, modification indices) are invariant whether saturated means are or are not included in the model.

[64] In Lisrel, the ty vector can be used to this end.

References

Abad, F. J., Colom, R., Rebollo, I., & Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: evidence for bias. *Personality and Individual Differences, 36*, 1459-1470.

Abdel-Khalek, A. M., & Raven, J. (2006). Normative data from the standardization of Raven's Standard Progressive Matrices in Kuwait in an international context. *Social Behavior and Personality, 34*, 169-180.

Aboud, F., Samuel, M., Hadera, A., & Addus, A. (1991). Intellectual, social, and nutritional status of children in an Ethiopian orphanage. *Social science and medicine, 33*, 1275-1280.

AERA, APA, & NCME. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Ahmed, R. A. (1989). The development of number, space, quantity, and reasoning concepts in Sudanese schoolchildren. In L. L. Adler (Ed.), *Cross cultural research in human development: Life span perspectives* (pp. 17-26). New York, NY: Praeger Publishers.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716-729.

Akande, A. (2000). Order effects on neuropsychological test performance of normal, learning disabled and low functioning children: A cross-cultural study. *Early Child Development and Care, 165*, 145-161.

American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: American Psychiatric Association.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96-116). Baltimore, MD: The Johns Hopkins University Press.

Angoff, W. H., & Ford, S. F. (1972). Item-race interaction on a test of scholastic aptitude. *Journal of Educational Measurement, 10*, 95-105.

Aronson, J., Lustina, M. J., Good, C., Keough, K., Steele, C. M., & Brown, J. (1999). When White men can't do math: Necessary and sufficient factors in stereotype threat. *Journal of Experimental Social Psychology, 35*, 29-46.

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

Avenant, T. J. (1988). *The establishment of an individual intelligence scale for adult South Africans. Report on an exploratory study conducted with the WAIS-R on a sample of Blacks.* (No. P-91). Pretoria, South Africa: Human Sciences Research Council.

Badri, M. B. (1965a). Influence of modernization on Goodenough quotients of Sudanese children. *Perceptual and Motor Skills, 20*, 931-932.

Badri, M. B. (1965b). The use of finger drawing in measuring the Goodenough quotient of culturally deprived Sudanese children. *Journal of Psychology, 59*, 333-334.

Bakare, C. G. M. (1972). Social-class differences in the performance of Nigerian children on the Draw-a-Man test. In L. J. Cronbach & P. J. Drenth (Eds.), *Mental tests and cultural adaptation* (pp. 355-363). The Hague, The Netherlands: Mouton.

Barber, N. (2005). Educational and ecological correlates of IQ: A cross-national investigation. *Intelligence, 33*, 273-284.

Bardet, C., Moreigne, F., & Sénécal, J. (1960). Application de test de Goodenough à des écoliers africains de 7 à 14 ans [Application of the Goodenough test to African school children ages 7 to 14]. *Enfance*, 199-208.

Barnett, S. M., & Williams, W. M. (2004). National intelligence and the emperor's new clothes. *Contemporary Psychology, 49*, 389-396.

Baron, R. M., & Kenny, D. A. (1986). The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173-1182.

Bartholomew, D. J. (2004). *Measuring intelligence. Facts and fallacies*. Cambridge, UK: Cambridge University Press.

Beilock, S. L., & Carr, T. H. (2005). When high-powered people fail: Working memory and "choking under pressure" in math. *Psychological Science, 16*, 101-105.

Ben Zeev, T., Fein, S., & Inzlicht, M. (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology, 41*, 174-181.

Bentler, P. M. (1990). Comparative fix indexes in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M. (1995). EQS structural equations program manual. Encino, CA: Multivariate Software.

Berlioz, L. (1955). Étude des Progressive Matrices faite sur les Africains de Douala [Study of the Progressive Matrices among Africans of Douala]. *Bulletin du C.E.R.P., 4*, 33-44.

Berry, J. W. (1966). Temne and Eskimo perceptual skills. *International Journal of Psychology, 1*, 207-229.

Berry, J. W. (1974). Radical cultural relativism and the concept of intelligence. In J. W. Berry & P. R. Dasen (Eds.), *Culture and cognition: Readings in cross-cultural psychology* (pp. 225-229). London: Methuen & Co Ltd.

Berry, J. W. (1976). *Human ecology and cognitive style. Comparative studies in cultural and psychological adaptation*. New York, NY: Sage Publications, Inc.

Berry, J. W. (1983). This week's citation classic. *Current Contents, 24*, 22.

Biasutti, R. (1959). *Le razze e i popoli della terra [The races and the peoples of the earth]*. Torino, Italy: Unione Tipografico Torinese.

Bielinski, J., & Davison, M. L. (1998). Gender differences by item difficulty interactions in multiple-choice mathematics items. *American Educational Research Journal, 35*, 455-476.

Bielinski, J., & Davison, M. L. (2001). A sex difference by item difficulty interaction in multiple-choice mathematics items administered to national probability samples. *Journal of Educational Measurement, 38*, 51-77.

Biesheuvel, S. (1943). *African intelligence*. Johannesburg, South Africa: South African Institute of Race Relations.

Binnie Dawson, J. L. (1984). Bio-social and endocrine bases of spatial ability. *Psychologia: An International Journal of Psychology in the Orient, 27*, 129-151.

Blair, C., Gamson, D., Thorne, S., & Baker, D. (2005). Rising mean IQ: Cognitive demand of mathematics education for young children, population exposure to formal schooling, and the neurobiology of the prefrontal cortex. *Intelligence, 33*, 93-106.

Blascovich, J., Spencer, S. J., Quinn, D., & Steele, C. M. (2001). African Americans and high blood pressure: The role of stereotype threat. *Psychological Science, 12*, 225-229.

Bleichrodt, N., Drenth, P. J. D., Zaal, J. N., & Resing, W. C. M. (1984). *Revisie Amsterdamse kinder intelligentie test [Revised Amsterdam Child Intelligence Test]*. Lisse, The Netherlands: Swets & Zeitlinger.

Boissiere, M., Knight, J. B., & Sabot, R. H. (1985). Earnings, schooling, ability and cognitive skills. *American Economic Review, 75*, 1016-1030.

Boivin, M. J. (2002). Effects of early cerebral malaria on cognitive ability in Senegalese children. *Journal of Developmental and Behavioral Pediatrics, 23*, 353-364.

Boivin, M. J., & Giordani, B. (1993). Improvements in cognitive performance for schoolchildren in Zaire, Africa, following an iron supplement and treatment for intestinal parasites. *Journal of Pediatric Psychology, 18*, 249-264.

Boivin, M. J., Giordani, B., & Bornefeld, B. (1995). Use of the Tactual Performance Test for cognitive ability testing with African children. *Neuropsychology, 9*, 409-417.

Boivin, M. J., Giordani, B., Ndanga, K., Maky, M. M., & et al. (1993). Effects of treatment for intestinal parasites and malaria on the cognitive abilities of schoolchildren in Zaire, Africa. *Health Psychology, 12*, 220-226.

Bollen, K. A. (1989). *Structural equations with latent variables*. Oxford, England: John Wiley and Sons.

Bollen, K. A., & Long, J. S. (1993). Introduction. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 1-9). Newbury Park, CA: Sage Publications.

Borsboom, D. (2006a). The attack of the psychometricians. *Psychometrika, 71*, 425-440.

Borsboom, D. (2006b). When does measurement invariance matter? *Medical Care, 44*, S176-S181.

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review, 110*, 203-219.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52*, 345-370.

Brand, C. R. (1987). Bryter still and bryter? *Nature, 328*, 110.

Brand, C. R. (1990). A "gross" underestimate of a "massive" IQ rise? A rejoinder to Flynn. *Irish Journal of Psychology, 11*, 52-56.

Brand, C. R., Freshwater, S., & Dockrell, W. B. (1989). Has there been a massive rise in IQ levels in the West? Evidence from Scottish children. *Irish Journal of Psychology, 10*, 388-393.

Brown, R. P., & Pinel, E. C. (2003). Stigma on my mind: Individual differences in the experience of stereotype threat. *Journal of Experimental Social Psychology, 39*, 626-633.

Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research, 24*, 445-455.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136–162). Newbury Park, CA: Sage Publications.

Buchanan, J. R. (1889). *Manual of psychometry: The dawn of a new civilization* (3rd ed.). Boston, MA: Press of Cupples, Wilson.

Buj, V. (1981). Average IQ values in various European countries. *Personality and Individual Differences, 2*, 168-169.

Buros, O. K. (1959). *The fifth mental measurements yearbook*. Highland Park, NJ: Gryphon.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456-466.

Camarata, S., & Woodcock, R. W. (2006). Sex differences in processing speed: Developmental effects in males and females. *Intelligence, 34*, 231-252.

Campbell, J. P. (1996). Group differences and personnel decisions: Validity, fairness, and affirmative action. *Journal of Vocational Behaviour, 49*, 122-158.

Carlson, J. S. (1970). A note on the relationship between the Draw-a-Man test, the Progressive Matrices, and conservation. *Journal of Psychology, 74*, 231-235.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97*, 404-431.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. New York, NY: Cambridge University Press.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5-21). Oxford, UK: Elsevier Science Ltd.

Catron, D. W., & Thompson, C. C. (1979). Test-retest gains in WAIS scores after four retest intervals. *Journal of Clinical Psychology, 35*, 352-357.

Cattell, R. B. (1950). The fate of national intelligence; test of a thirteen-year prediction. *Eugenics Review, 42*, 136-148.

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology, 27*, 703-722.

Chan, K. Y., Drasgow, F., & Sawin, L. L. (1999). What is the shelf life of a test? The effect of time on the psychometrics of a cognitive ability test battery. *Journal of Applied Psychology, 84*, 610-619.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*, 233-255.

Chirkov, V. I., Ryan, R. M., & Willness, C. (2005). Cultural context and psychological needs in Canada and Brazil: Testing a self-determination approach to the internalization of cultural practices, identity, and well-being. *Journal of Cross Cultural Psychology, 36*, 423-443.

Cleary, T. A. (1968). Test bias: Predictions of grades of negro and white students in integrated colleges. *Journal of Educational Measurement, 5*, 115-124.

Coatsworth, J. D., Sharp, E. H., Palen, L. A., Darling, N., Cumsille, P., & Marta, E. (2005). Exploring adolescent self-defining leisure activities and identity experiences across three countries. *International Journal of Behavioral Development, 29*, 361-370.

Colom, R., Andres-Pueyo, A., & Juan-Espinosa, M. (1998). Generational IQ gains: Spanish data. *Personality and Individual Differences, 25*, 927-935.

Colom, R., & García-López, O. (2003). Secular gains in fluid intelligence: Evidence from the culture-fair intelligence test. *Journal of Biosocial Science, 35*, 33-39.

Colom, R., Juan Espinosa, M., & Garcia, L. F. (2001). The secular increase in test scores is a "Jensen effect." *Personality and Individual Differences, 30*, 553-559.

Coon, C. S. (1966). *The living races of man*. London: Jonahthan Cape.

Corwyn, R. F., & Bradley, R. H. (2005). The cross-gender equivalence of strains and gains from occupying multiple roles among dual-earner couples. *Parenting: Science and Practice, 5*, 1-26.

Costenbader, V., & Ngari, S. M. (2001). A Kenya standardization of the Raven's Coloured Progressive Matrices. *School Psychology International, 22*, 258-268.

Crawford Nutt, D. H. (1976). Are Black scores on Raven's Standard Progressive Matrices an artifact of method of test presentation? *Psychologia Africana, 16*, 201-206.

Crawford Nutt, D. H. (1977). The effect of educational level on the test scores of people in South Africa. *Psychologia Africana, 17*, 49-59.

Crockett, L. J., Randall, B. A., Shen, Y. L., Russell, S. T., & Driscoll, A. K. (2005). Measurement equivalence of the Center for Epidemiological Studies Depression scale for Latino and Anglo adolescents: A national study. *Journal of Consulting and Clinical Psychology, 73*, 47-58.

Croizet, J. C., Després, G., Gauzins, M. E., Huguet, P., Leyens, J. P., & Méot, A. (2004). Stereotype threat undermines intellectual performance by triggering a disruptive mental load. *Personality and Social Psychology Bulletin, 30*, 721-731.

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

Cullen, M. J., Hardison, C. M., & Sackett, P. R. (2004). Using SAT-grade and ability-job performance relationships to test predictions derived from stereotype threat theory. *Journal of Applied Psychology, 89*, 220-230.

Dagevos, J., Gijsberts, M., & van Praag, C. (2003). *Rapportage minderheden 2003. [Minority report 2003]*. The Hague, The Netherlands: Sociaal en Cultureel Planbureau.

Dague, P. (1972). Development, application and interpretation of tests for use in French-speaking black Africa and Madagascar. In L. J. Cronbach & P. J. Drenth (Eds.), *Mental tests and cultural adaptation* (pp. 63-74). The Hague, The Netherlands: Mouton.

Daley, T. C., Whaley, S. E., Sigman, M. D., Espinosa, M. P., & Neumann, C. (2003). IQ on the rise: The Flynn effect in rural Kenyan children. *Psychological Science, 14*, 215-219.

Dambrun, M., & Taylor, D. M. (2005). Race, sex, and social class differences in cognitive ability: towards a contextual rather than genetic explanation. *Current Research in Social Psychology,, 10*, 189-202.

Davenport, E. C. (1990). Significance testing of congruence coefficients: A good idea? *Educational and Psychological Measurement, 50*, 289-296.

de Frias, C. M., & Dixon, R. A. (2005). Confirmatory factor structure and measurement invariance of the Memory Compensation Questionnaire. *Psychological Assessment, 17*, 168-178.

Dent, G. R. (1937). The educability of the Bantu. In E. G. Malherbe (Ed.), *Educational adaptations in a changing society*. Capetown, South Africa: Juta & Co., Ltd.

DeShon, R. P. (2004). Measures are not invariant across groups without error variance homogeneity. *Psychology Science, 46*, 137-149.

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

Dickens, W. T., & Flynn, J. R. (2006). Black Americans reduce the racial IQ gap: Evidence from standardization samples. *Psychological Science, 17*, 913-920.

Dickerson, R. E. (2006). Exponential correlation of IQ and the wealth of nations. *Intelligence, 34*, 291-295.

Dillon, R. F., Pohlmann, J. T., & Lohman, D. F. (1981). A factor analysis of Raven's Advanced Progressive Matrices freed of difficulty factors. *Educational and Psychological Measurement, 41*, 1295-1302.

Dolan, C. V. (2000). Investigating Spearman's hypothesis by means of multi-group confirmatory factor analysis. *Multivariate Behavioral Research, 35*, 21-50.

Dolan, C. V., Colom, R., Abad, F. J., Wicherts, J. M., Hessen, D. J., & Van der Sluis, S. (2006). Multi-group covariance and mean structure modeling of the relationship between WAIS-III common factors and gender and educational attainment in Spain. *Intelligence, 34*, 193-210.

Dolan, C. V., & Hamaker, E. L. (2001). Investigating Black-White differences in psychometric IQ: Multi-group confirmatory factor analyses of the WISC-R and K-ABC and a critique of the method of correlated vectors. In F. Columbus (Ed.), *Advances in psychology research* (Vol. 6, pp. 31-59). Huntington, NY: Nova Science Publishers, Inc.

Dolan, C. V., & Lubke, G. H. (2001). Viewing Spearman's hypothesis from the perspective of multigroup PCA: A comment on Schoenemann's criticism. *Intelligence, 29*, 231-245.

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

Donaldson, G. W. (2003). General linear contrasts on latent variable means: Structural equation hypothesis tests for multivariate clinical trails. *Statistics in Medicine, 22*, 2893-2917.

Drasgow, F., & Kanfer, R. (1985). Equivalence of psychological measurement in heterogeneous populations. *Journal of Applied Psychology, 70*, 662-680.

Du, L., & Tang, T. L.-P. (2005). Measurement invariance across gender and major: The love of money among university students in People's Republic of China. *Journal of Business Ethics, 59*, 281-293.

Elshout, J. J. (1976). *Karakteristieke moeilijkheden in het denken. [Characteristic difficulties in thinking]*. Unpublished doctoral dissertation, University of Amsterdam, Amsterdam, The Netherlands.

Emanuelsson, I., Reuterberg, S. E., & Svensson, A. (1993). Changing differences in intelligence? Comparisons between groups of 13-year-olds tested from 1960 to 1990. *Scandinavian Journal of Educational Research, 37*, 259-277.

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

Embretson, S. E., & Schmidt McCollam, K. M. (2000). Psychometric approaches to understanding and measuring intelligence. In R. J. Sternberg (Ed.), *Handbook of Intelligence* (pp. 423-444). Cambridge, UK: Cambridge University Press.

Engle, R. W. (2002). Working memory capacity as executive attention. *Current Directions in Psychological Science, 11*, 19-23.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology: General, 128*, 309-331.

Ervik, A. O. (2003). Book review: IQ and the wealth of nations. *The Economic Journal, 113*, 406-408.

Evers, A., & Lucassen, W. (1992). *Handleiding DAT'83 [DAT '83 Manual]*. Lisse, The Netherlands: Swets & Zeitlinger.

Fahmy, M. (1964). Initial exploring of the intelligence of Shilluk children: Studies in the Southern Sudan. *Vita Humana, 7*, 164-177.

Fahrmeier, E. D. (1975). The effect of school attendance on intellectual development in northern Nigeria. *Child Development, 46*, 281-285.

Fernández-Ballesteros, R., Juan Espinosa, M., Colom, R., & Calero, M. D. (1997). Contextual and personal sources of individual differences in intelligence: Empirical results. In J. S. Carlson, J. Kingma & W. Tomic (Eds.), *Advances in cognition and educational practice: Reflections on the concept of intelligence.* (Vol. 4, pp. 221-274). London, England: JAI Press Inc.

Ferron, O. (1965). The test performance of "coloured" children. *Educational Research, 8*, 42-57.

Fick, M. L. (1929). Intelligence test results of poor white, native (Zulu), coloured, and Indian school children and the educational and social implications. *South African Journal of Science, 16*, 904-920.

Flieller, A. (1988). Application du modèle de Rasch à un problème de comparaison de générations. [Application of the Rasch model to a problem of intergenerational comparison]. *Bulletin de Psychologie, 42*, 86-91.

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

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

Flynn, J. R. (1990). Massive IQ gains on the Scottish WISC - Evidence against Brand et al's hypothesis. *Irish Journal of Psychology, 11*, 41-51.

Flynn, J. R. (1998a). IQ gains over time: Toward finding the causes. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 25-66). Washington, DC: American Psychological Association.

Flynn, J. R. (1998b). Israeli military IQ tests: Gender differences small; IQ gains large. *Journal of Biosocial Science, 30*, 541-553.

Flynn, J. R. (1998c). WAIS-III and WISC-III IQ gains in the United States from 1972 to 1995: How to compensate for obsolete norms. *Perceptual and Motor Skills, 86*, 1231-1239.

Flynn, J. R. (1999a). Evidence against Rushton: The genetic loading of WISC-R subtests and the causes of between-group IQ differences. *Personality and Individual Differences, 26*, 373-379.

Flynn, J. R. (1999b). Reply to Rushton: A gang of gs overpowers factor analysis. *Personality and Individual Differences, 26*, 391-393.

Flynn, J. R. (1999c). Searching for justice - The discovery of IQ gains over time. *American Psychologist, 54*, 5-20.

Flynn, J. R. (2000a). IQ gains and fluid g. *American Psychologist, 55*, 543-543.

Flynn, J. R. (2000b). IQ gains, WISC subtests and fluid g: g theory and the relevance of Spearman's hypothesis to race. In G. R. Bock, J. A. Goode & K. Webb (Eds.), *The nature of intelligence: Novartis Foundation Symposium 233* (pp. 202 - 227). Chichester, U.K.: Wiley.

Flynn, J. R. (2006). The Flynn Effect: Rethinking intelligence and what affects it. In C. Flores-Mendoza & R. Colom (Eds.), *Introduction to the Psychology of Individual Differences*. Porto Alegre, Brazil: ArtMed.

Ghorpade, J., Hattrup, K., & Lackritz, J. R. (1999). The use of personality measures in cross-cultural research: A test of three personality scales across two countries. *Journal of Applied Psychology, 84*, 670-679.

Giordani, B., Boivin, M. J., Opel, B., Dia Nseyila, D. N., & Lauer, R. E. (1996). Use of the K-ABC with children in Zaire, Africa: An evaluation of the sequential-simultaneous processing distinction within an intercultural context. *International Journal of Disability, Development and Education, 43*, 5-24.

Glewwe, P., & Jacoby, H. (1992). *Estimating the determinants of cognitive achievement in low income countries*. Washington, DC: World Bank.

Goodenough, F. L. (1926). *Measurement of intelligence by drawings*. Chicago, IL: World Book Company.

Goodenough, F. L., & Harris, D. B. (1950). Studies in the psychology of children's drawings II. 1928-1949. *Psychological Bulletin, 47*, 360-433.

Gottfredson, L. S. (2005). Implications of cognitive differences for schooling within diverse societies. In C. L. Frisby & C. R. Reynolds (Eds.), *Comprehensive handbook of multicultural school psychology* (pp. 517-554). New York, NY: John Wiley & Sons, Inc.

Greenfield, P. M. (1997). You can't take it with you: Why ability assessments don't cross cultures. *American Psychologist, 52*, 1115-1124.

Greenfield, P. M. (1998). The cultural evolution of IQ. In U. Neisser (Ed.), *The rising curve: Long term gains in IQ and related measures* (pp. 81-123). Washington, DC: American Psychological Association.

Grieve, K. W., & Viljoen, S. (2000). An exploratory study of the use of the Austin Maze in South Africa. *South African Journal of Psychology, 30*, 14-18.

Gulliksen, H. (1950). *Theory of mental tests*. New York, NY: John Wiley & Sons, Inc.

Hagger, M. S., Chatzisarantis, N. L. D., Barkoukis, V., Wang, C. K. J., & Baranowski, J. (2005). Perceived Autonomy Support in Physical Education and Leisure-Time Physical Activity: A Cross-Cultural Evaluation of the Trans-Contextual Model. *Journal of Educational Psychology, 97*, 376-390.

Hakstian, A. R., & Vandenberg, S. G. (1979). The cross-cultural generalizability of a higher-order cognitive structure model. *Intelligence, 3*, 73-103.

Hambleton, R., & Swaminathan, H. (1985). *Item response theory : Principles and applications*. Dordrecht, The Netherlands: Kluwer-Nijhoff.

Hanushek, E. A., & Kimko, D. D. (2000). Schooling, labor-force quality, and the growth of nations. *American Economic Review, 90*, 1184-1208.

Harris, D. B. (1963). *Children's drawings as measures of intellectual maturity*. New York, NY: Harcourt, Brace & World, Inc.

Hartmann, P., Kruuse, N. H. S., & Nyborg, H. (2007). Testing the cross-racial generality of Spearman's hypothesis in two samples. *Intelligence, 35*, 47-57.

Heady, C. (2003). The effect of child labor on learning achievement. *World Development, 31*, 385-398.

Hedges, L. V., & Nowell, A. (1995). Sex differences in mental test scores, variability, and numbers of high-scoring individuals. *Science, 269*, 41-45.

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. A. (1994). *The bell curve: Intelligence and class structure in American life*. New York, NY: Free Press.

Hessen, D. J., Dolan, C. V., & Wicherts, J. M. (2006). The multi-group common factor model with minimal uniqueness constraints and the power to detect uniform bias. *Applied Psychological Measurement, 30*, 233-246.

Heyneman, S. P. (1975). *Influences on academic achievement in Uganda. A "Coleman Report" from a non-industrial society.* Unpublished doctoral dissertation, University of Chicago, Chicago, IL.

Heyneman, S. P., & Jamison, D. T. (1980). Student learning in Uganda. *Comparative Educational Review, 24*, 207-220.

Holding, P. A., Taylor, H. G., Kazungu, S. D., Mkala, T., Gona, J., Mwamuye, B., et al. (2004). Assessing cognitive outcomes in a rural African population: Development of a neuropsychological battery in Kilifi District, Kenya. *Journal of the International Neuropsychological Society, 10*, 246-260.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning.* Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research, 18*, 117-144.

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *The Southern Psychologist, 1*, 179-188.

Howard, R. W. (1999). Preliminary real-world evidence that average human intelligence really is rising. *Intelligence, 27*, 235-250.

Howard, R. W. (2001). Searching the real world for signs of rising population intelligence. *Personality and Individual Differences, 30*, 1039-1058.

Howard Scott, L. (1981). Measuring intelligence with the Goodenough-Harris Drawing test. *Psychological Bulletin, 89*, 483-505.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Hunkin, V. (1950). Validation of the Goodenough Draw-a-man Test for African children. *Journal for Social Research, 1*, 52-63.

Hunt, E. B., & Carlson, J. S. (2006). Considerations relating to the study of racial/ethnic differences in intelligence. *paper submitted for publication*.

Hunt, E. B., & Sternberg, R. J. (2006). Sorry, wrong numbers: An analysis of a study of a correlation between skin color and IQ. *Intelligence, 34*, 131-137.

Hunter, J. E., & Schmidt, F. L. (2000). Racial and gender bias in ability and achievement tests: Resolving the apparent paradox. *Psychology, Public Policy, and Law, 6*, 151-158.

Husén, T., & Tuijnman, A. (1991). The contribution of formal schooling to the increase in intellectual capital. *Educational Researcher, 20*, 17-25.

Inzlicht, M., & Ben Zeev, T. (2003). Do high-achieving female students underperform in private? The implications of threatening environments on intellectual processing. *Journal of Educational Psychology, 95*, 796-805.

Ironson, G. H., Homan, S., Willis, R., & Signer, B. (1984). The validity of item bias techniques with math word problems. *Applied Psychological Measurement, 8*, 391-396.

Ironson, G. H., & Subkoviak, M., J. (1979). A comparison of several methods of assessing item bias. *Journal of Educational Measurement, 16*, 209-225.

Irvine, S. H. (1966). Towards a rationale for testing attainments and abilities in Africa. *British Journal of Educational Psychology, 36*, 24-32.

Irvine, S. H. (1969a). Factor analysis of African abilities and attainments: Constructs across cultures. *Psychological Bulletin, 71*, 20-32.

Irvine, S. H. (1969b). Figural tests of reasoning in Africa. Studies in the use of Raven's Progressive Matrices across cultures. *International Journal of Psychology, 4*, 217-228.

Jablonski, N. G. (2004). The evolution of human skin and skin color. *Annual Review of Anthropology, 33*, 585-623.

Jackson, D. N., & Rushton, J. P. (2006). Males have greater g: Sex differences in general mental ability from 100,000 17- to 18-year-olds on the Scholastic Assessment Test. *Intelligence, 34*, 479-486.

Jedege, R. O., & Bamgboye, E. A. (1981). Self-concepts of young Nigerian adolescents. *Psychological Reports, 49*, 451-454.

Jensen, A. R. (1980). *Bias in mental testing.* London: Methuen & Co., Ltd.

Jensen, A. R. (1982). The debunking of scientific fossils and straw persons. *Contemporary Education Review, 1*, 121-135.

Jensen, A. R. (1996). Secular trends in IQ: Additional hypothesis. In D. K. Detterman (Ed.), *The environment. Current topics in human intelligence, Vol. 5.* (pp. 147 - 150). Westport, CT: Ablex Publishing.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT, US: Praeger Publishers/Greenwood Publishing Group, Inc.

Jinabhai, C. C., Taylor, M., Rangongo, M. F., Mkhize, N. J., Anderson, S., Pillay, B. J., et al. (2004). Investigating the mental abilities of rural Zulu primary school children in South Africa. *Ethnicity and Health, 9*, 17-36.

Jones, G., & Schneider, W. J. (2006). Intelligence, human capital, and economic growth: A Bayesian Averaging of Classical Estimates (BACE) approach. *Journal of Economic Growth, 11*, 71–93.

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

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Newbury Park, CA: Sage.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: User's reference guide.* Chicago, IL: Scientific Software International.

Jöreskog, K. G., & Sörbom, D. (2003). LISREL 8.5. Lincolnwood, IL: Scientific Software International.

Kamin, L. J. (1995). Lies, damned lies, and statistics. In R. Jacoby & N. Glauberman (Eds.), *The Bell Curve debate* (pp. 81-105). New York, NY: Random House, Inc.

Kanazawa, S. (2004). General intelligence as a domain-specific adaptation. *Psychological Review, 111*, 512-523.

Kaniel, S., & Fisherman, S. (1991). Level of performance and distribution of errors in the progressive matrices test: A comparison of Ethiopian and Israeli adolescents. *International Journal of Psychology, 26*, 25-33.

Kaufman, A. S., & Kaufman, N. L. (1983). *Kaufman Assessment Battery for Children: Administration and scoring manual.* Circle Pines, MN: American Guidance Service.

Keller, J. C. (2002). Blatant stereotype threat and women's math performance: Self-handicapping as a strategic means to cope with obtrusive negative performance expectations. *Sex Roles, 47*, 193-198.

Kendall, I. M. (1976). The predictive validity of a possible alternative to the Classification Test Battery. *Psychologia Africana, 16*, 131-146.

Kendall, I. M., Verster, M. A., & Von Mollendorf, J. W. (1988). Test performance of Blacks in Southern Africa. In S. H. Irvine & J. W. Berry (Eds.), *Human abilities in cultural context.* (pp. 299-339). New York, NY: Cambridge University Press.

Kirkcaldy, B., Furnham, A., & Siefen, G. (2004). The relationship between health efficacy, educational attainment and well-being among 30 nations. *European Psychologist, 9*, 107-119.

Kleinpenning, G., & Hagendoorn, L. (1991). Contextual aspects of ethnic stereotypes and interethnic evaluations. *European Journal of Social Psychology, 21*, 331-348.

Klingelhofer, E. L. (1967). Performance of Tanzanian secondary school pupils on the Raven Standard Progressive Matrices Test. *Journal of Social Psychology, 72*, 205-215.

Knoetze, J., Bass, N., & Steele, G. (2005). The Raven's Coloured Progressive Matrices: Pilot norms for Xhosa-speaking primary school learners in pert-urban Eastern Cape. *South African Journal of Psychology, 35*, 175-194.

Kozulin, A. (1998). Profiles of immigrant students' cognitive performance on Raven's Progressive Matrices. *Perceptual and Motor Skills, 87*, 1311-1314.

Krige, S. (1997). Segregation, science and commissions of enquiry: The contestation over native education policy in South Africa. *Journal of Southern African Studies, 23*, 491-506.

Lane, C. (1994). Tainted sources. *The New York Review of Books, December 1, 1994*.

Laroche, J. L. (1959). Effets de répétition du Matrix 38 sur les résultats d'enfants Katangais [Results from retesting Katanga children with the Raven's Matrix 38]. *Bulletin du C.E.R.P., 8*, 85-99.

Latouche, G. L. M., & Dormeau, G. (1956). *La formation professionnelle rapide en Afrique Équatoriale Française [Fast vocational training in French Equatorial Africa]*. Brazzaville, Congo: Centre d' Étude des Problèmes du Travail.

Levin, M. (1997). *Why race matters. Race differences and what they mean.* Westport, CT: Praeger.

Little, R. J. A., & Rubin, D. B. (1986). *Statistical analysis with missing data.* New York, NY: John Wiley & Sons, Inc.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research, 32*, 53-76.

Liu, C., Borg, I., & Spector, P. E. (2004). Measurement equivalence of the German Job Satisfaction Survey used in a multinational organization: implications of Schwartz's culture model. *Journal of Applied Psychology, 89*, 1070-1082.

Lloyd, F., & Pidgeon, D. A. (1961). An investigation into the effects of coaching on non-verbal test material with European, Indian and African children. *British Journal of Psychology, 31*, 145-151.

Loehlin, J. C. (2000). Group differences in intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 176-193). Cambridge, UK: Cambridge University Press.

Lohman, D. F. (2000). Complex information processing and intelligence. In R. J. Sternberg (Ed.), *Handbook of Intelligence* (pp. 285-340). Cambridge, UK: Cambridge University Press.

Lord, F. M. (1977). A study of item bias, using item characteristic curve theory. In Y. H. Poortinga (Ed.), *Basic problems in cross-cultural psychology* (pp. 19-29). Amsterdam, The Netherlands: Swets and Zeitlinger B.V.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley Publishing Company.

Lubke, G. H., & Dolan, C. V. (2003). Can unequal residual variances across groups mask differences in residual means in the common factor model? *Structural Equation Modeling, 10*, 175-192.

Lubke, G. H., Dolan, C. V., & Kelderman, H. (2001). Investigating group differences on cognitive tests using Spearman's hypothesis: An evaluation of Jensen's method. *Multivariate Behavioral Research, 36*, 299-324.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003a). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence, 31*, 543-566.

Lubke, G. H., Dolan, C. V., Kelderman, H., & Mellenbergh, G. J. (2003b). Weak measurement invariance with respect to unmeasured variables: An implication of strict factorial invariance. *British Journal of Mathematical and Statistical Psychology, 56*, 231-248.

Lubke, G. H., & Muthen, B. (2005). Investigating population heterogeneity with factor mixture models. *Psychological Methods, 10*, 21-39.

Lynn, R. (1978). Ethnic and racial differences in intelligence: International comparisons. In R. T. Osborne, C. E. Noble & N. Weyl (Eds.), *Human variation. The biopsychology of age, race, and sex*. New York, NY: Academic Press.

Lynn, R. (1982). IQ in Japan and the United States shows a growing disparity. *Nature, 297*, 222-223.

Lynn, R. (1989). A nutrition theory of the secular increases in intelligence - Positive correlations between height, head size and IQ. *British Journal of Educational Psychology, 59*, 372-377.

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

Lynn, R. (1991). Race differences in intelligence: A global perspective. *Mankind Quarterly, 31*, 255-296.

Lynn, R. (1997). Geographical variation in intelligence. In H. Nyborg (Ed.), *The scientific study of human nature: Tribute to Hans J. Eysenck at eighty* (pp. 259-281). Oxford, UK: Elsevier Science Ltd.

Lynn, R. (2003). The geography of intelligence. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 127-146). Oxford, UK: Elsevier Science Ltd.

Lynn, R. (2006). *Race differences in intelligence: An evolutionary analysis*. Augusta, GA: Washington Summit Publishers.

Lynn, R., & Hampson, S. (1986). The rise of national intelligence: Evidence from Britain, Japan and the U.S.A. *Personality and Individual Differences, 7*, 23-32.

Lynn, R., & Hampson, S. (1989). Secular increases in reasoning and mathematical abilities in Britain, 1972-84. *School Psychology International, 10*, 301-304.

Lynn, R., & Holmshaw, M. (1990). Black-White differences in reaction times and intelligence. *Social Behavior and Personality, 18*, 299-308.

Lynn, R., & Irwing, P. (2002). Sex differences in general knowledge, semantic memory and reasoning ability. *British Journal of Psychology, 93*, 545-556.

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

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

MacArthur, R. S., Irvine, S. H., & Brimble, A. R. (1964). *The Northern Rhodesia mental ability survey*. Lusaka, Zambia: Rhodes Livingstone Institute.

MacEachern, S. (2006). Africanist archaeology and ancient IQ: racial science and cultural evolution in the twenty-first century. *World Archaeology, 38*, 72-92.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods, 7*, 83-104.

Mackintosh, N. J. (1998). *IQ and human intelligence*. New York, NY: Oxford University Press, Inc.

Mackintosh, N. J., & Bennett, E. S. (2005). What do Raven's Matrices measure? An analysis in terms of sex differences. *Intelligence, 33*, 663-674.

Maduagwu, S. N. (2003). *Development of education in Nigeria - Past, present and future*. Paper presented at the Institute for Educational Research and Service, International Christian University, Tokyo, Japan.

Mandys, F., Dolan, C. V., & Molenaar, P. C. M. (1994). Two aspects of the simplex model: Goodness of fit to linear growth curve structures and the analysis of mean trends. *Journal of Educational and Behavioral Statistics, 19*, 201-215.

Maqsud, M. (1980a). Personality and academic attainment of primary school children. *Psychological Reports, 46*, 1271-1275.

Maqsud, M. (1980b). Relationship between sociometric status and moral judgment in secondary school girls. *West African Journal of Educational and Vocational Measurement, 5*, 13-17.

Maqsud, M. (1997). Effects of metacognitive skills and nonverbal ability on academic achievement of high school pupils. *Educational Psychology, 17*, 387-397.

Martorell, R. (1998). Nutrition and the worldwide rise in IQ scores. In U. Neisser (Ed.), *The rising curve: Long term gains in IQ and related measures* (pp. 183-206). Washington, DC: American Psychological Association.

Matarazzo, R. G., Wiens, A. N., Matarazzo, J. D., & Manaugh, T. S. (1973). Test-retest reliability of the WAIS in a normal population. *Journal of Clinical Psychology, 29*, 194-197.

McFarland, L. A., Lev Arey, D. M., & Ziegert, J. C. (2003). An examination of stereotype threat in a motivational context. *Human Performance, 16*, 181-205.

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

McInerney, D. M., Dowson, M., & Yeung, A. S. (2005). Facilitating Conditions for School Motivation: Construct Validity and Applicability. *Educational and Psychological Measurement, 65*, 1046-1066.

McKay, P. F., Doverspike, D., Bowen Hilton, D., & Martin, Q. D. (2002). Stereotype threat effects on the Raven Advanced Progressive Matrices scores of African-Americans. *Journal of Applied Social Psychology, 32*, 767-787.

Meade, A. W., & Lautenschlager, G. J. (2004). A monte-carlo study of confirmatory factor analytic tests of measurement equivalence/invariance. *Structural Equation Modeling, 11*, 60-72.

Measso, G., Zappala, G., Cavarzeran, F., Crook, T. H., Romani, L., Pirozzolo, F. J., et al. (1993). Raven's Coloured Progressive Matrices: A normative study of a random sample of healthy adults. *Acta Neurologica Scandinavica, 88*, 70-74.

Meijer, R. R., & Sijtsma, K. (2001). Evaluating person fit. *Applied Psychological Measurement, 25*, 107-135.

Meisenberg, G. (2004). Talent, character, and the dimensions of national culture. *Mankind Quarterly, 45*, 123-168.

Mellenbergh, G. J. (1982). Contingency table models for assessing item bias. *Journal of Educational Statistics, 7*, 105-118.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research, 13*, 127-143.

Mellenbergh, G. J. (1994). A unidimensional latent trait model for continuous item responses. *Multivariate Behavioral Research, 29*, 223-236.

Mellenbergh, G. J., & Kok, F. G. (1991). Finding the biasing trait(s). In P. L. Dann & S. H. Irvine (Eds.), *Advances in computer based human assessment. Theory and decision library: Series D: System theory, knowledge engineering and problem solving* (Vol. 7, pp. 291-306). New York, NY: Kluwer Academic/Plenum Publishers.

Meredith, W. (1964). Notes on factorial invariance. *Psychometrika, 29*, 177-185.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525-543.

Meredith, W., & Horn, J. (2001). The role of factorial invariance in modeling growth and change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 203-240). Washington, DC: American Psychological Association.

Miller, E. M. (1992). On the correlation of myopia and intelligence. *Genetic, Social & General Psychology Monographs, 118*, 363-383.

Miller, E. M. (1995). Environmental variability selects for large families only in special circumstances: Another objection to differential K theory. *Personality and Individual Differences, 19*, 903-918.

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research, 30*, 577-605.

Millsap, R. E. (1997a). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*, 248-260.

Millsap, R. E. (1997b). The investigation of Spearman's hypothesis and the failure to understand factor analysis. *Cahiers de Psychologie Cognitive, 16*, 750-757.

Millsap, R. E. (1998). Group differences in regression intercepts: Implications for factorial invariance. *Multivariate Behavioral Research, 33*, 403-424.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

Millsap, R. E., & Kwok, O. M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods, 9*, 93-115.

Minde, K., & Kantor, S. (1976). Instructing Ugandan primary schoolchildren in the execution of an "intelligence" test: A controlled evaluation. *Journal of Cross Cultural Psychology, 7*, 209-222.

Mingroni, M. A. (2004). The secular rise in IQ: Giving heterosis a closer look. *Intelligence, 32*, 65-83.

Morakinyo, O. (1985). The brain-fag syndrome in Nigeria: Cognitive deficits in an illness associated with study. *British Journal of Psychiatry, 146*, 209-210.

Morse, S. (2006). Making development simple. The genetic deterministic hypothesis for economic development. *Ecological Economics, 56*, 79-88.

Mpofu, E. (2004). Being intelligent with Zimbabweans: A historical and contemporary view. In R. J. Sternberg (Ed.), *International handbook of intelligence* (pp. 364-390). New York, NY: Cambridge University Press.

Mpofu, E., & Watkins, D. (1994). The similarities subtest of the British Ability Scales: Construct and content bias for a sample of Zimbabwe school children. *Educational and Psychological Measurement, 54*, 728-733.

Mung'ala Odera, V., Snow, R. W., & Newton, C. R. J. C. (2004). The burden of the neurocognitive impairment associated with Plasmodium Falciparum malaria in sub-Saharan Africa. *American Journal of Tropical Medicine and Hygiene, 71*, 64-70.

Munroe, R. L., & Munroe, R. H. (1983). Drawings and values in three East African societies. *Journal of Social Psychology, 119*, 135-136.

Must, O., Must, A., & Raudik, V. (2003). The secular rise in IQs: In Estonia the Flynn effect is not a Jensen effect. *Intelligence, 167*, 1-11.

Muthen, B., & Muthen, L. K. (2003). *MPlus version 2.14*. Los Angeles, CA: Muthén & Muthén.

Muthén, B. O. (1989). Factor structure in groups selected on observed scores. *British Journal of Mathematical and Statistical Psychology, 42*, 81-90.

Naglieri, J. A., & Jensen, A. R. (1987). Comparison and black-white differences on the WISC-R and the K-ABC: Spearman's Hypothesis. *Intelligence, 11*, 21-43.

Neisser, U. (Ed.). (1998). *The rising curve: Long-term gains in IQ and related measures*. Washington, DC: American Psychological Association.

Nell, V. (2000). *Cross-cultural neuropsychological assessment: Theory and practice*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Nenty, H. J., & Dinero, T. E. (1981). A cross-cultural analysis of the fairness of the Cattell Culture Fair Intelligence Test using the Rasch model. *Applied Psychological Measurement, 5*, 355-368.

Nguyen, H. H. D., O'Neal, A., & Ryan, A. M. (2003). Relating test-taking attitudes and skills and stereotype threat effects to the racial gap in cognitive ability test performance. *Human Performance, 16*, 261-293.

Nissen, H. W., Machover, S., & Kinder, E. F. (1935). A study of performance tests given to a group of native African negro children. *British Journal of Psychology, 25*, 308-355.

Nkaya, H. N., Huteau, M., & Bonnet, J. P. (1994). Retest effect on cognitive performance on the Raven-38 Matrices in France and in the Congo. *Perceptual and Motor Skills, 78*, 503-510.

Notcutt, B. (1950). The measurement of Zulu intelligence. *Journal for Social Research, 1*, 195-206.

Nwanze, H. O. (1985). Relations between spelling and performance in Nigerian elementary school children. *Journal of Social Psychology, 125*, 45-52.

Nwanze, H. O., & Okeowo, P. A. (1980). The usefulness of a developmental profile in predicting reading retardation in Nigerian children. *West African Journal of Educational and Vocational Measurement, 5*, 50-59.

O'Brien, L. T., & Crandall, C. S. (2003). Stereotype threat and arousal: Effects on women's math performance. *Personality and Social Psychology Bulletin, 29*, 782-789.

Ogunlade, J. O. (1978). The predictive validity of the Raven Progressive Matrices with some Nigerian children. *Educational and Psychological Measurement, 38*, 465-467.

Ohuche, N. M., & Ohuche, R. O. (1973). The Draw-A-Man Test as a predictor of academic achievement. *West African Journal of Educational and Vocational Measurement, 1*, 20-27.

Okonji, M. O. (1974). Predicting reading proficiency in some Nigerian primary school children. *West-African Journal of Educational and Vocational Measurement, 2*, 17-23.

Okunrotifa, P. O. (1976). A comparison of the entry behaviours of Nigerian rural and urban children in geography. *West-African Journal of Educational and Vocational Measurement, 3*, 1-6.

Ombredane, A. (1957). Étude du comportement intellectual des noirs congolais [A study of intellectual behavior in Congo Negroes]. *Psychologie Française, 1*, 19.

Ombredane, A., Robaye, F., & Plumail, H. (1956). Résultats d'une application répétée du matrix-couleur à une population de Noirs Congolais [Results of a repeated administration of the color-matrix test to Congo Blacks]. *Bulletin du C.E.R.P., 5*, 129-147.

Ombredane, A., Robaye, F., & Robaye, E. (1957). *Application expérimentale du test d'intelligence Matrix 38 a 485 noirs Baluba. [Experimental administration of the Matrix 38 intelligence test among 485 Baluba Blacks]*. Brussels, Belgium: Académie Royale des Sciences Coloniales.

Oort, F. J. (1992). Using restricted factor analysis to detect item bias. *Methodika, 6*, 150-166.

Oosterveld, P. (1996). *Questionnaire design methods.* Unpublished Doctoral dissertation, University of Amsterdam, Amsterdam, The Netherlands.

Osborne, J. W. (2001). Testing stereotype threat: Does anxiety explain race and sex differences in achievement? *Contemporary Educational Psychology, 26*, 291-310.

Owen, K. (1992). The suitability of Raven's Standard Progressive Matrices for various groups in South Africa. *Personality and Individual Differences, 13*, 149-159.

Ployhart, R. E., & Oswald, F. L. (2004). Applications of mean and covariance structure analysis: Integrating correlational and experimental approaches. *Organizational Research Methods, 7*, 27-65.

Ployhart, R. E., Ziegert, J. C., & McFarland, L. A. (2003). Understanding racial differences on cognitive ability tests in selection contexts: An integration of stereotype threat and applicant reactions research. *Human Performance, 16*, 231-259.

Pons, A. L. (1974). *Administration of tests outside cultures of their origin.* Paper presented at the 26th Annual congress of the South African Psychological Association, Johannesburg, South Africa.

Poortinga, Y. H., & van der Flier, H. (1988). The meaning of item bias in ability tests. In S. H. Irvine & J. W. Berry (Eds.), *Human abilities in cultural context* (pp. 166-183). Cambridge, UK: Cambridge University Press.

Quinn, D. M., & Spencer, S. J. (2001). The interference of stereotype threat with women's generation of mathematical problem-solving strategies. *Journal of Social Issues, 57*, 55-71.

Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology, 87*, 517-529.

Randi, J. (1982). *Flim-Flam! Psychics, ESP, unicorns, and other delusions.* Amherst, NY: Promotheus Books.

Raveau, F. H. M., Elster, E., & Lecoutre, J. P. (1976). Migration et acculturation différentielle. [Migration and differential acculturation]. *Revue de Psychologie Appliquée, 25*, 145-165.

Raven, J. (2000). The Raven's Progressive Matrices: Change and stability over culture and time. *Cognitive Psychology, 41*, 1-48.

Raven, J. C. (1956). *Guide to using the Coloured Progressive Matrices.* London: H.K. Lewis & Co. Ltd.

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

Raven, J. C., Court, J. H., & Raven, J. (1990). *Manual for Raven's Coloured Progressive Matrices*. Oxford, UK: Oxford Psychologists Press.

Raven, J. C., Court, J. H., & Raven, J. (1996). *Manual for Raven's Progressive Matrices and Vocabulary Scales*. Oxford, UK: Oxford Psychologists Press.

Reeve, C. L., & Hakel, M. D. (2002). Asking the right questions about g. *Human Performance, 15*, 47-74.

Reilly, R. R. (1973). A note on minority group test bias studies. *Psychological Bulletin, 80*, 130-132.

Richter, L. M., Griesel, R. D., & Wortley, M. E. (1989). The Draw-a-Man test: A 50-year perspective on drawings done by black South African children. *South African Journal of Psychology, 19*, 1-5.

Rietveld, M. J. H., van Baal, G. C. M., Dolan, C. V., & Boomsma, D. I. (2000). Genetic factor analyses of specific cognitive abilities in 5-year-old Dutch children. *Behavior Genetics, 30*, 29-40.

Rindermann, H. (2006). Was messen internationale Schulleistungsstudien? Schulleistungen, Schülerfähigkeiten, kognitive Fähigkeiten, Wissen oder allgemeine Intelligenz? [What do international student assessment studies measure? School performance, student abilities, cognitive abilities, knowledge or general intelligence?]. *Psychologische Runschau, 57*, 69-86.

Robins, A. H. (1991). *Biological Perspectives on Human Pigmentation*. Cambridge: Cambridge University Press.

Rodgers, J. L. (1998). A critique of the Flynn Effect: Massive IQ gains, methodological artifacts, or both? *Intelligence, 26*, 337-356.

Roll, W. G. (2003). Poltergeists, electromagnetism and consciousness. *Journal of Scientific Exploration, 17*, 75-86.

Rushton, J. P. (1996). Political correctness and the study of racial differences. *Journal of Social Distress and the Homeless, 5*, 213-229.

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

Rushton, J. P. (1999). Secular gains in IQ not related to the g factor and inbreeding depression - unlike Black-White differences: A reply to Flynn. *Personality and Individual Differences, 26*, 381-389.

Rushton, J. P. (2000a). Flynn effects not genetic and unrelated to race differences. *American Psychologist, 55*, 542-543.

Rushton, J. P. (2000b). *Race, evolution, and behavior. A life history perspective*. Port Huron, MI: Charles Darwin Research Institute.

Rushton, J. P. (2001). Black-White differences on the g-factor in South Africa: A "Jensen Effect" on the Wechsler Intelligence Scale for Children-Revised. *Personality and Individual Differences, 31*, 1227-1232.

Rushton, J. P. (2002). Jensen effects and African/Coloured/Indian/White differences on Raven's Standard Progressive Matrices in South Africa. *Personality and Individual Differences, 33*, 1279-1284.

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

Rushton, J. P., & Jensen, A. R. (2005b). Wanted: More race realism, less moralistic fallacy. *Psychology, Public Policy, and Law, 11*, 328-336.

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

Rushton, J. P., 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.

Rushton, J. P., Skuy, M., & Fridjhon, P. (2002). Jensen effects among African, Indian and White engineering students in South Africa on Raven's standard progressive matrices. *Intelligence, 30*, 409-423.

Rushton, J. P., Skuy, M., & Fridjhon, P. (2003). Performance on Raven's Advanced Progressive Matrices by African, East Indian, and White engineering students in South Africa. *Intelligence, 31*, 123-137.

Sackett, P. R. (2003). Stereotype threat in applied selection settings: A commentary. *Human Performance, 16*, 295-309.

Sackett, P. R., Hardison, C. M., & Cullen, M. J. (2004). On interpreting stereotype threat as accounting for African American-White differences on cognitive tests. *American Psychologist, 59*, 7-13.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review, 103*, 403-428.

Sarich, V., & Miele, F. (2004). *Race: The reality of human differences*. Boulder, CO: Westview Press.

Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 181-204). Newbury Park, CA: Sage Publications.

Satzger, W., Dragon, E., & Engel, R. R. (1996). The equivalence of the German version of the Wechsler adult intelligence scale-revised (HAWIE-R) and the original German version (HAWIE). *Diagnostica, 42*, 119-138.

Schallberger, U. (1987). HAWIK und HAWIK-R: Ein empirischer Vergleich. [HAWIK and HAWIK-R: An empirical comparison]. *Diagnostica, 33*, 1-13.

Schmader, T., & Johns, M. (2003). Converging evidence that stereotype threat reduces working memory capacity. *Journal of Personality and Social Psychology, 85*, 440-452.

Scholderer, J., Grunert, K. G., & Brunso, K. (2005). A procedure for eliminating additive bias from cross-cultural survey data. *Journal of Business Research, 58*, 72-78.

Schooler, C. (1998). Environmental complexity and the Flynn effect. In U. Neisser (Ed.), *The rising curve: Long term gains in IQ and related measures* (pp. 67-79). Washington, DC: American Psychological Association.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6*, 461-464.

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*, 38-56.

Sekaquaptewa, D., & Thompson, M. (2003). Solo status, stereotype threat, and performance expectancies: Their effects on women's performance. *Journal of Experimental Social Psychology, 39*, 68-74.

Serpell, R. (1979). How specific are perceptual skills? A cross-cultural study of pattern reproduction. *British Journal of Psychology, 70*, 365-380.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for causal inference.* Boston, MA: Houghton Mifflin Company.

Shepard, L. A., Camilli, G., & Williams, D. M. (1985). Validity of approximation techniques for detecting item bias. *Journal of Educational Measurement, 22*, 77-105.

Shuttleworth Edwards, A. B., Kemp, R. D., Rust, A. L., Muirhead, J. G. L., Hartman, N. P., & Radloff, S. E. (2004). Cross-cultural effects on IQ test performance: A review and preliminary normative indications on WAIS-III test performance. *Journal of Clinical and Experimental Neuropsychology, 26*, 903-920.

Sigman, M., Neumann, C., Jansen, A. A., & Bwibo, N. (1989). Cognitive abilities of Kenyan children in relation to nutrition, family characteristics, and education. *Child Development, 60*, 1463-1474.

Sigman, M., & Whaley, S. E. (1998). The role of nutrition in the development of intelligence. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 155-182). Washington, DC: American Psychological Association.

Skuy, M., Gewer, A., Osrin, Y., Khunou, D., Fridjhon, P., & Rushton, J. P. (2002). Effects of mediated learning experience on Raven's matrices scores of African and non-African university students in South Africa. *Intelligence, 30*, 221-232.

Skuy, M., Schutte, E., Fridjhon, P., & O'Carroll, S. (2001). Suitability of published neuropsychological test norms for urban African secondary school students in South Africa. *Personality and Individual Differences, 30*, 1413-1425.

Skuy, M., Taylor, M., O'Carroll, S., Fridjhon, P., & Rosenthal, L. (2000). Performance of Black and White South African children on the Wechsler Intelligence Scale for Children-Revised and the Kaufman Assessment Battery. *Psychological Reports, 86*, 727-737.

Smith, J. L. (2004). Understanding the process of stereotype threat: A review of mediational variables and new performance goal directions. *Educational Psychology Review, 16*, 177-206.

Smith, J. L., & White, P. H. (2001). Development of the domain identification measure: A tool for investigating stereotype threat effects. *Educational and Psychological Measurement, 61*, 1040-1057.

Smith, J. L., & White, P. H. (2002). An examination of implicitly activated, explicitly activated, and nullified stereotypes on mathematical performance: It's not just a woman's issue. *Sex Roles, 47*, 179-191.

Sonke, C. J. (2001). *Cross-cultural differences on simple cognitive tasks. A psychophysiological investigation.* Unpublished Doctoral Thesis, Tilburg University, Tilburg, The Netherlands.

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

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

Spencer, S. J., Steele, C. M., & Quinn, D. M. (1999). Stereotype threat and women's math performance. *Journal of Experimental Social Psychology, 35*, 4-28.

Spitz, H. H. (1989). Variations in Wechsler interscale IQ disparities at different levels of IQ. *Intelligence, 13*, 157-167.

Stapleton, L. M., & Leite, W. L. (2005). A review of syllabi for a sample of structural equation modeling courses. *Structural Equation Modeling, 12*, 642–664.

Statistics Netherlands. (2003). *Beroepsbevolking naar onderwijsniveau 1993. [Working population by educational level 1993].* Retrieved May 14th, 2003, from http://statline.cbs.nl

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

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*, 797-811.

Steele, C. M., & Davies, P. G. (2003). Stereotype threat and employment testing: A commentary. *Human Performance, 16*, 311-326.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 34, pp. 379-440). San Diego, CA, US: Academic Press, Inc.

Steenkamp, J. B. E. M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research, 25*, 78-90.

Sternberg, R. J. (2004). Culture and intelligence. *American Psychologist, 59*, 325-338.
Sternberg, R. J. (2005). There are no public-policy implications: A reply to Rushton and Jensen (2005). *Psychology, Public Policy, and Law, 11*, 295-301.
Sternberg, R. J., Grigorenko, E. L., Ngorosho, D., Tantufuye, E., Mbise, A., Nokes, C., et al. (2002). Assessing intellectual potential in rural Tanzanian school children. *Intelligence, 30*, 141-162.
Sternberg, R. J., Nokes, C., Geissler, P. W., Prince, R., Okatcha, F., Bundy, D. A., et al. (2001). The relationship between academic and practical intelligence: A case study in Kenya. *Intelligence, 29*, 401-418.
Stevens, J. (1996). *Applied multivariate statistics for the social sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.
Stinissen, J. (1977). *De constructie van de nederlandstalige WAIS*. Leuven, Belgium: Katholieke Universiteit Leuven.
Stinissen, J., Willems, P. J., Coetsier, P., & Hulsman, W. L. L. (1970). *Handleiding bij de nederlandstalige bewerking van de Wechsler Adult Intelligence Scale (WAIS). [Dutch WAIS Manual]*. Lisse, The Netherlands: Swets & Zeitlinger.
Stricker, L. W., & Bejar, I. I. (2004). Test difficulty and stereotype threat on the GRE general test. *Journal of Applied Social Psychology, 34*, 563-597.
Stricker, L. W., & Ward, W. C. (2004). Stereotype threat, inquiring about test takers' ethnicity and sex, and standardized test performance. *Journal of Applied Social Psychology, 34*, 665-693.
Suzuki, L., & Aronson, J. (2005). The cultural malleability of intelligence and its impact on the racial/ethnic hierarchy. *Psychology, Public Policy, and Law, 11*, 320-327.
Swets & Zeitlinger. (2003). *Aanvullend normonderzoek WAIS-III. [Additional norming study WAIS-III]*. Retrieved May, 7th, 2003, from http://www.swetstest.nl/info/WAIS-III/
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., Evers, A., & Mur, J. P. (2000). Validity of the Differential Aptitude Test for the assessment of immigrant children. *Educational Psychology, 20*, 99-115.
Te Nijenhuis, J., Tolboom, E., Resing, W., & Bleichrodt, N. (2004). Does cultural background influence the intellectual performance of children from immigrant groups? The RAKIT intelligence test for immigrant children. *European Journal of Psychological Assessment, 20*, 10-26.
Te Nijenhuis, J., & van der Flier, H. (1997). Comparability of GATB scores for immigrants and majority group members: Some Dutch findings. *Journal of Applied Psychology, 82*, 675-687.
Te Nijenhuis, J., Voskuijl, O. F., & Schijve, N. B. (2001). Practice and coaching on IQ tests: quite a lot of g. *International Journal of Selection and Assessment, 9*, 302-308.
Teasdale, T. W., & Owen, D. R. (1987). National secular trends in intelligence and education: A twenty-year cross-sectional study. *Nature, 325*, 119-121.
Teasdale, T. W., & Owen, D. R. (1989). Continuing secular increases in intelligence and a stable prevalence of high intelligence levels. *Intelligence, 13*, 255-262.
Teasdale, T. W., & Owen, D. R. (2000). Forty-year secular trends in cognitive abilities. *Intelligence, 28*, 115-120.
Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn Effect in reverse. *Personality and Individual Differences, 39*, 837-843.
Tellegen, P. J. (2002). De kwaliteit van de normen van de WAIS-III. [Quality of the WAIS-III norms.]. *De Psycholoog, 37*, 463-465.
Templer, D. I., & Arikawa, H. (2006). Temperature, skin color, per capita income, and IQ: An international perspective. *Intelligence, 34*, 121-139.
Thoma, R. J., Yeo, R. A., Gangestad, S. W., Halgren, E., Sanchez, N. M., & Lewine, J. D. (2005). Cortical volume and developmental instability are independent predictors of general intellectual ability. *Intelligence, 33*, 27-38.
Thurstone, L. L. (1925). A method of scaling educational and psychological tests. *Journal of Educational Psychology, 16*, 263-278.
Thurstone, T. G. (1958). *Manual for the SRA Primary Mental Abilities Test 11-17*. Chicago, IL: Science Research Associates.
Thurstone, T. G. (1962). *Primary mental abilities for grades 9-12*. Chicago, IL: Science Research Associates.
Tuddenham, R. D. (1948). Soldier intelligence in World Wars I and II. *American Psychologist, 3*, 54-56.
Tzuriel, D., & Kaufman, R. (1999). Mediated learning and cognitive modifiability. Dynamic assessment of young Ethiopian immigrant children to Israel. *Journal of Cross Cultural Psychology, 30*, 359-380.
UN Development Programme. (2005). *Human development report*. New York, NY: UN Development Programme.
United Nations. (2005). *The millennium development goals report 2005*. New York, NY: United Nations.
Van de Vijver, F., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: Sage.
van den Briel, T., West, C. E., Bleichrodt, N., van de Vijver, F. J. R., Ategbo, E. A., & Hautvast, G. A. J. (2000). Improved iodine status is associated with improved mental performance of schoolchildren in Benin. *American Journal of Clinical Nutrition, 72*, 1179-1185.
Van der Linden, W. J., & Hambleton, R. (Eds.). (1996). *Handbook of modern item response theory*. New York, NY: Springer Verlag.

van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: the positive manifold of intelligence by mutualism. *Psychological Review, 113*, 842-861.

Van der Sluis, S., Posthuma, D., Dolan, C. V., de Geus, E. J. C., Colom, R., & Boomsma, D. I. (2006). Sex differences in the Dutch WAIS-III. *Intelligence, 34*, 273-289.

van der Ven, A. H. G. S., & Ellis, J. L. (2000). A Rasch analysis of Raven's Standard Progressive Matrices. *Personality and Individual Differences, 29*, 45-64.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4-69.

Verhaegen, P. (1956). Utilité actuelle des tests pour l'étude psychologique des autochtones congolais [Current usability of tests for the psychological examination of Congolese natives]. *Revue de Psychologie Appliquée, 6*, 139-151.

Verkuyten, M., & Kinket, B. (1999). The relative importance of ethnicity: Ethnic categorization among older children. *International Journal of Psychology, 34*, 107-118.

Verkuyten, M., & Thijs, J. (2004). Psychological disidentification with the academic domain among ethnic minority adolescents in The Netherlands. *British Journal of Educational Psychology, 74*, 109-125.

Vernon, P. E. (1969). *Intelligence and cultural environment*. London, UK: Methuen & Co. Ltd.

Voracek, M. (2004). National intelligence and suicide rate: an ecological study of 85 countries. *Personality and Individual Differences, 37*, 543-553.

Vorst, H. C. M., & Zand Scholten, A. (2000). *Psychometrische analyse van metingen op het cognitieve, structurele en affectieve domein afgenomen in Testweek 31. [Psychometric analysis of measures in the cognitive, structural, and affective domain, administred during Testweek 31]* (Internal report). Amsterdam, The Netherlands: University of Amsterdam, Psychological Methods Department.

Walton, G. M., & Cohen, G. L. (2003). Stereotype lift. *Journal of Experimental Social Psychology, 39*, 456-467.

Wechsler, D. (1955). *WAIS Manual*. New York, NY: The Psychological Corporation.

Wechsler, D. (1974). *WISC-R Manual*. New York, NY: The Psychological Corporation.

Wechsler, D. (1981). *WAIS-R Manual*. New York, NY: The Psychological Corporation.

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale-Third Edition*. San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2000). *WAIS-III. Nederlandstalige bewerking. Technische handleiding. [Techical manual Dutch WAIS-III]*. Lisse, The Netherlands: Swets Test Publishers.

Wechsler, D. (2004). *Wechsler Intelligence Scale for Children - Fourth Edition*. San Antonio, TX, US: Harcourt Assessment.

Weede, E., & Kampf, S. (2002). The impact of intelligence and institutional improvements on economic growth. *Kyklos, 55*, 361-380.

Wheeler, S. C., & Petty, R. E. (2001). The effects of stereotype activation on behavior: A review of possible mechanisms. *Psychological Bulletin, 127*, 797-826.

Whetzel, D. L., & McDaniel, M. A. (2006). Prediction of national wealth. *Intelligence, 34*, 449-458.

Wicherts, J. M. (2005a). *Flynn Effect in the Woodcock-Johnson cognitive ability tests 1987-1999*. Paper presented at the 6th Annual Conference of the International Society for Intelligence Research (ISIR), Albuquerque, NM.

Wicherts, J. M. (2005b). Stereotype threat research and the assumptions underlying analysis of covariance. *American Psychologist, 60*, 267-269.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*, 726-728.

Wicherts, J. M., & Dolan, C. V. (2004). A cautionary note on the use of information fit indexes in covariance structure modeling with means. *Structural Equation Modeling, 11*, 45-50.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89*, 696-716.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (submitted). Measurement invariance and group differences in intercepts in confirmatory factor analysis. *paper submitted for publication*.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., et al. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence, 32*, 509-537.

Wicherts, J. M., Van Asten, E. J., Balcombe, S. J., Boom, S. M., Van den Heuvel, B. B., & Sylva, H. (2003). *Stereotype threat and intelligence test scores of minority and majority high school students* (Internal Report): University of Amsterdam.

Widaman, K. F., & Reise, S. P. (1997). Exploring the measurement invariance of psychological instruments: Applications in the substance use domain. In K. J. Bryant, M. Windle & S. G. West (Eds.), *The science of prevention: Methodological advances from alcohol and substance abuse research* (pp. 281-324). Washington, DC: American Psychological Association.

Widaman, K. F., & Thompson, J. S. (2003). On specifying the null model for incremental fit indices in structural equation modeling. *Psychological Methods, 8*, 16-37.

Wikipedia. (2006). *Psychometry*. Retrieved June 18, 2006, from http://en.wikipedia.org/wiki/Psychometry
Wildt, A. R., & Ahtola, O. (1978). *Analysis of covariance*. Thousand Oaks, CA: Sage Publications.
Williams, J. H. (1935). Validity and reliability of the Goodenough intelligence test. *School and Society, 41*, 653-656.
Williams, W. M. (1998). Are we raising smarter children today? School- and home-related influences on IQ. In U. Neisser (Ed.), *The rising curve: Long term gains in IQ and related measures* (pp. 125-154). Washington, DC: American Psychological Association.
Wober, M. (1966). Sensotypes. *Journal of Social Psychology, 70*, 181-189.
Wober, M. (1969). Meaning and stability of Raven's Matrices test among Africans. *International Journal of Psychology, 4*, 229-235.
Wober, M. (1974). Towards an understanding of the Kiganda concept of intelligence. In J. W. Berry & P. R. Dasen (Eds.), *Culture and cognition: Readings in cross-cultural psychology* (pp. 261-280). London: Methuen & Co Ltd.
Wober, M. (1975). *Psychology in Africa*. London: International African Institute.
Woehr, D. J., Sheehan, M. K., & Bennett, W., Jr. (2005). Assessing measurement equivalence across rating sources: A multitrait-multirater approach. *Journal of Applied Psychology, 90*, 592-600.
Wolff, P. H., Tesfai, B., Egasso, H., & Aradom, T. (1995). The orphans of Eritrea: A comparison study. *Journal of Child Psychology and Psychiatry, 36*, 633-644.
Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock–Johnson Psycho-Educational Battery–Revised*. Allen, TX: Developmental Learning Materials.
Woodcock, R. W., McGrew, K. S., & Mather, N. (2001). *Woodcock-Johnson-III*. Itasca, IL: Riverside Publishing.
Yao, G., & Wu, C. h. (2005). Factorial invariance of the WHOQOL-BREF among disease groups. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care and Rehabilitation, 14*, 1881-1888.
Yzerbyt, V. Y., Muller, D., & Judd, C. M. (2004). Adjusting researchers' approach to adjustment: On the use of covariates when testing interactions. *Journal of Experimental Social Psychology, 40*, 424-431.
Zaaiman, H., van der Flier, H., & Thijs, G. (2001). Dynamic testing in selection for an educational program: Assessing South African performance on the Raven Progressive Matrices. *International Journal of Selection and Assessment, 9*, 258-269.
Zajonc, R. B., & Mullally, P. R. (1997). Birth order: Reconciling conflicting effects. *American Psychologist, 52*, 685-699.
Zand Scholten, A. (2003). *Profielkeuzeadvies via internet. [Profile counseling via the internet]*. Unpublished Master's Thesis, University of Amsterdam, Amsterdam, The Netherlands.
Zindi, F. (1994a). Differences in psychometric performance. *The Psychologist, 7*, 549-552.
Zindi, F. (1994b). Towards the standardization of the WISC--R for early childhood assessment in Zimbabwe. *IFE Psychologia: An International Journal, 2*, 19-32.

Samenvatting

Groepsverschillen in prestaties op intelligentie tests

Groepsverschillen in de scores op intelligentie tests behoren tot de meest controversiële onderwerpen van de psychologie. Dit proefschrift gaat over dergelijke groepsverschillen en benadert deze vooral vanuit de context van het lineaire confirmatieve factor model. Dit psychometrische model is uitermate goed geschikt om groepsverschillen in multivariate intelligentie test scores mee te onderzoeken, omdat ermee kan worden nagegaan of met tests in verschillende groepen dezelfde latente variabelen (of factoren) worden gemeten. Meer specifiek kan ermee worden onderzocht of groepsverschillen in waargenomen test scores kunnen worden toegewezen aan groepsverschillen op de onderliggende latente variabelen die dergelijke tests beogen te meten (Meredith, 1993). Een dergelijke situatie wordt ook wel meetinvariantie ten opzichte van groepen genoemd (Mellenbergh, 1989).

Verschillende Soorten Psychometrie

In het eerste hoofdstuk wordt opgemerkt dat er in de loop der tijd een schisma lijkt te zijn ontstaan tussen de meer technische psychometrie die gericht is op het modelleren van test scores aan de ene kant, en een psychometrie die gericht is op het begrijpen van cognitieve capaciteiten (of intelligentie) aan de andere kant. Daar waar vroeger veel onderzoekers een interesse aan de dag legden voor beide deelaspecten van intelligentie test scores, lijken de technische psychometrische ontwikkelingen en de inhoudelijke ontwikkelingen op het gebied van intelligentieonderzoek tegenwoordig steeds meer uit elkaar te zijn gelopen. Zo worden moderne analytische technieken niet ten volle benut om licht te werpen op de variabelen die gemeten worden aan de hand van cognitieve tests. Het doel van dit proefschrift is om relatief moderne psychometrische technieken toe te passen op groepsverschillen in intelligentie test prestaties, zoals die gevonden worden tussen bijvoorbeeld etnische groepen (Hoofdstuk 2), groepen waarover wel of geen negatieve stereotypen bestaan m.b.t. test prestaties (Hoofdstuk 3) en verschillende cohorten (Hoofdstuk 4). Hoofdstuk 5 is gewijd aan de interpretatie van IQ test scores van personen uit Afrika. In Hoofdstuk 6 wordt betoogd dat gezien de structuur van individuele verschillen in intelligentie, het gebruik van eenvoudige analysetechnieken nauwelijks bijdraagt aan het begrip van effecten van bepaalde variabelen op, of groepsverschillen in, intelligentie test scores.

Meetinvariantie en Groepsverschillen in Meetintercepten

Hoofdstuk 2 heeft betrekking op meetinvariantie van tests en in het bijzonder op eerlijkheid van IQ tests ten opzichte van bestaande groepen. Mellenbergh (1989) stelde een eenvoudige en algemeen geldende definitie op van meetinvariantie ten opzichte van groepen. Volgens zijn definitie zijn testscores meetinvariant ten opzichte van groepen wanneer geldt dat, gegeven een bepaalde waarde op de latente trek, de verwachte waarde op de test onafhankelijk is van groepslidmaatschap. Een schending van meetinvariantie (ook

wel meetonzuiverheid genoemd) betekent dat groepsverschillen in test scores niet eenvoudigweg kunnen worden geïnterpreteerd als groepsverschillen in de latente trek(ken) die de test beoogt te meten.

In Hoofdstuk 2 wordt binnen het kader van het lineaire factormodel inzichtelijk gemaakt dat er geen sprake kan zijn van meetinvariantie ten opzichte van groepen als er bepaalde toetsen op de gemiddeldestructuur van test scores niet zijn uitgevoerd. Binnen dit model wordt dit aspect gemodelleerd aan de hand van meetintercepten. Voor een zinvolle groepsvergelijking dienen deze meetintercepten gelijk te zijn over groepen. Uit een overzicht van de recente literatuur blijkt dat in veel onderzoek naar meetinvariantie aan de hand van het confirmatieve factor model dit centrale aspect van meetinvariantie niet nadrukkelijk is getoetst. In Hoofdstuk 2 wordt beargumenteerd dat zonder deze toets niet geconcludeerd kan worden dat een test meetinvariant is ten opzichte van groepen.

De consequenties van het negeren van meetintercepten in de vergelijking van groepen wordt geïllustreerd aan de hand van een heranalyse van een gepubliceerde studie naar de bruikbaarheid bij allochtone kinderen van een veel gebruikte Nederlandse intelligentietest, te weten de Revisie Amsterdamse Kinder Intelligentie Test (RAKIT). Hoewel er in de oorspronkelijke studie door Te Nijenhuis, Tolboom, Resing en Bleichrodt (2004) gebruik gemaakt is van verschillende methoden om meetinvariantie te toetsen, is er door deze auteurs geen toets uitgevoerd op de gelijkheid van meetintercepten in het factormodel. Daar waar deze auteurs op grond van hun eigen analyses concluderen dat de RAKIT in sterke mate meetinvariant is voor allochtone kinderen, laat de heranalyse in Hoofdstuk 2 zien dat de RAKIT niet meetinvariant is ten opzichte van etnische groepen en dat de RAKIT de latente vaardigheden bij van oorsprong Marokkaanse en Turkse kinderen met ten minste 7 IQ punten onderschat. Dit impliceert dat de RAKIT alleen met grote voorzichtigheid kan worden gebruikt bij het meten van intelligentie bij allochtone kinderen. Dit resultaat laat tevens zien dat er meer behoefte is naar het bepalen van eerlijkheid van veel gebruikte intelligentietests.

Stereotype Bedreiging en Groepsverschillen in Test Scores

Hoofdstuk 3 gaat over de effecten van stereotype bedreiging op test prestaties. Stereotype bedreiging is de angst om onbedoeld te voldoen aan een negatieve stereotype die betrekking heeft op de prestaties van de eigen groep (Steele & Aronson, 1995). Zo kunnen vrouwen angst hebben om laag te scoren op een wiskundetest, omdat er een stereotype bestaat dat vrouwen minder goed zijn in wiskunde. Uit veel experimenteel laboratoriumonderzoek is gebleken dat wanneer personen uit gestigmatiseerde groepen op meer of minder subtiele wijze worden herinnerd aan hun lidmaatschap van die groep, dit een negatief effect kan hebben op hun test prestaties. Omdat een dergelijk effect sterke maatschappelijke gevolgen hebben voor leden van die groepen, is het van belang om na te gaan in hoeverre dit effect optreedt in echte testsituaties. In dergelijke testsituaties is het echter vaak onethisch of onmogelijk om de effecten van stereotype bedreiging te onderzoeken.

In Hoofdstuk 3 wordt beargumenteerd dat het effect van stereotype bedreiging op test prestaties kan worden gezien als een meetartefact dat leidt tot een schending van meetinvariantie ten opzichte van groepen die wel of geen last hebben van het relevante

negatieve stereotype. In drie experimenten is met behulp van multigroep confirmatieve factor analyse nagegaan of de effecten van stereotype bedreiging op test scores inderdaad dit psychometrische effect hebben. In het eerste experiment is gekeken naar de prestaties van allochtone en autochtone middelbare scholieren op een kleine intelligentietest. Hierbij werden scholieren aselect verdeeld over testsituaties die verschilden in de mate waarin deze stereotype bedreiging opwekken voor allochtone leerlingen. Hoewel het gemiddeldeneffect in dit experiment afwezig was, wees een toets op meetinvariantie uit dat deze manipulatie een duidelijke schending van meetinvariantie teweeg heeft gebracht. In het tweede en derde experiment werd gekeken naar de effecten van experimenteel opgewekte stereotype bedreiging op de prestaties van vrouwelijke studenten op wiskundetests. Ook hier werd gevonden dat wanneer stereotype bedreiging een (verlagend) effect heeft op test prestaties, dit leidt tot een schending van meetinvariantie. Deze resultaten wijzen erop dat de effecten van stereotype bedreiging in principe, en ongeacht de testsituatie, detecteerbaar zijn aan de hand van toetsen op meetinvariantie. Dit maakt het mogelijk om deze effecten ook in "echte" testsituaties te onderzoeken. De modelmatige aanpak in de analyse van experimentele resultaten in dit hoofdstuk illustreert bovendien het grote voordeel van dergelijke analyses boven analyses van enkel en alleen gemiddeldeneffecten zoals die in de experimentele psychologie gebruikelijk zijn.

Aard van het Flynn Effect

Het Flynn Effect is de term voor de stijging van gemiddelde intelligentie test scores over de jaren heen. Zo liet Flynn (1987) zien dat Nederlandse mannen bij de dienstkeuring in de tweede helft van de twintigste eeuw steeds hoger zijn gaan scoren op een als cultuurvrij bekend staande niet-verbale intelligentietest. Van 1952 tot 1982 stegen de gemiddelde scores van de rekruten met maar liefst 20 IQ punten. Dergelijke forse trends in populatiegemiddelde IQ test scores zijn inmiddels in veel westerse landen en in enkele niet-westerse landen gedocumenteerd. Dit heeft de vraag opgeworpen wat de aard is van deze toename in intelligentie test scores. Wordt deze veroorzaakt door een toename in de latente trek algemene intelligentie of is er sprake van meetartefacten, bijvoorbeeld omdat personen steeds handiger zijn geworden in het maken van IQ tests?

In Hoofdstuk 4 is onderzocht hoe deze stijging binnen het confirmatieve factor model moet worden geïnterpreteerd. Hiertoe is een vijftal vergelijkingen uitgevoerd van cohorten die dezelfde IQ test batterij in verschillende periodes hebben gemaakt. Zo werden de intelligentietest scores van een steekproef Nederlandse volwassenen uit het eind van de jaren zestig vergeleken met de scores op dezelfde test van een steekproef Nederlandse volwassenen uit het eind van de jaren negentig. Uit multigroep confirmatieve factoranalyses blijkt dat bij alle vijf de vergelijkingen van cohorten de gebruikte intelligentietests niet meetinvariant zijn over de tijd. Dit impliceert dat de stijging in IQ test scores niet alleen maar kan worden toegeschreven aan een stijging van de latente variabelen die ten grondslag liggen aan deze test scores. Met andere woorden, het Flynn Effect lijkt deels te kunnen worden toegeschreven aan meetartefacten. Toch blijkt een deel van de toename te kunnen worden toegeschreven aan toenames in latente trekken. Echter, verklaringen voor het Flynn Effect kunnen niet louter betrekking hebben op effecten op het niveau van (brede) latente cognitieve vaardigheden.

IQ Scores in Afrika

Op de grond van een aantal uitgebreide literatuuroverzichten heeft Richard Lynn (2006) geconcludeerd dat het gemiddelde IQ van zwarte Afrikanen onder de 70 ligt. In Hoofdstuk 5 wordt er op kritische wijze gekeken naar de gegevens waarop Lynn deze bewering heeft gebaseerd. Er moet worden opgemerkt dat scores van Afrikanen op westerse IQ tests niet zomaar kunnen worden geïnterpreteerd in termen van de latente trek algemene intelligentie of *g*, zoals Lynn en anderen hebben gedaan. Voor een dergelijke interpretatie van de relatief lage IQ scores van Afrikaanse steekproeven moet aan een groot aantal methodologische en psychometrische eisen worden voldaan. Zo moet er zekerheid bestaan over dat alle getesten weten wat er van hen verwacht wordt en moeten tests worden afgenomen volgens strikte regels zoals geformuleerd in testhandleidingen. Het is vrij goed denkbaar dat deze ideale testsituaties niet altijd gelden bij afnamen van westerse IQ tests onder Afrikaanse personen en dat daardoor hun latente cognitieve vaardigheden door deze IQ tests worden onderschat.

Uit de resultaten van de literatuurstudie komt naar voren dat van de meest gebruikte IQ tests niet is komen vast te staan of deze in Afrikaanse steekproeven een goede en valide weergave geven van de latente trek algemene intelligentie. Ook is vooralsnog onduidelijk in hoeverre er bij de vergelijking van test prestaties tussen Afrikanen en westerlingen sprake is van meetinvariantie. Niettegenstaande komt uit de literatuurstudie naar voren dat de scores op deze IQ tests in Afrika aanzienlijk hoger liggen dan Lynn doet voorkomen, vooral omdat Lynn een aanzienlijke portie van de literatuur over het hoofd heeft gezien of simpelweg niet in zijn overzicht heeft opgenomen. In vergelijking tot Amerikaanse normen scoren Afrikaanse steekproeven op een tweetal abstracte intelligentietest gemiddeld rond een IQ van 80. Gezien de reële mogelijkheid van psychometrische problemen en de relatief slechte omstandigheden waaronder veel Afrikanen opgroeien is dit lage gemiddelde niet verwonderlijk. Gemiddelde scores op dergelijke IQ tests hebben in de meeste westerse landen een aanzienlijke stijging laten zien die geacht wordt te zijn veroorzaakt door zaken als verbeteringen in gezondheidszorg en voeding, verbeteringen in onderwijsniveau, urbanisatie, trend naar kleinere gezinnen en technologische ontwikkelingen. Uit de in Hoofdstuk 5 gerapporteerde correlaties op het niveau van landen blijkt dat vrijwel alle ontwikkelingen die in de westerse wereld verantwoordelijk worden gehouden voor het Flynn Effect, in Afrika nog niet of nauwelijks hebben plaatsgevonden. Dit suggereert dat de relatief lage IQ scores van Afrikanen alles behalve steun bieden aan genetische theorieën over rassenverschillen in intelligentie test scores zoals geformuleerd door Lynn (2006) en Rushton (2000b).

Discussie

In Hoofdstuk 6 wordt een geïdealiseerd model gepresenteerd van de structuur van individuele verschillen in cognitieve vaardigheden (zie Figuur 6.1). Voorts wordt beargumenteerd dat gezien deze structuur, groepsverschillen in cognitieve vaardigheden, of effecten van een bepaalde Variabele X op deze vaardigheden (bijv. cognitieve stimulatie tijdens de jeugd), op vier verschillende niveaus kunnen plaatsgrijpen. Op Niveau I is er sprake van een effect van (groep of variabele) X op de hogere orde factor g. Op Niveau II

is er sprake van een effect direct op eerste orde factor(en), zoals of Ruimtelijke Vaardigheid of Verwerkingssnelheid. Op Niveau III is er sprake van een effect direct op de subtest in de test batterij oftewel op de subtest-specifieke vaardigheid. Op Niveau IV is er sprake van een direct effect op itemscores, wat kan worden gezien als een schending van meetinvariantie van items.

Er wordt in Hoofdstuk 6 betoogd dat het gebruik van gesommeerde IQ scores in onderzoek naar groepsverschillen in, of naar effecten van een Variabele X op, cognitieve vaardigheden ons niet veel wijzer maakt, omdat aan de hand van gesommeerde IQ scores geen onderscheid kan worden gemaakt tussen de effecten op Niveaus I, II, III en IV. Daarentegen levert het gebruik van multigroep confirmatieve factor analyse in combinatie met item respons modellen deze informatie wel op. De resultaten van de onderzoeken in Hoofdstukken 2, 3, 4 en 5 worden geïnterpreteerd vanuit dit geïdealiseerde model. Groepsverschillen in meetintercepten die onderwerp waren van Hoofdstuk 2 kunnen worden gezien als effecten op Niveau III of IV. Nader onderzoek met de meting van deze eventuele additionele variabelen (zoals test-specifieke vaardigheden) of additionele analyses aan de hand van Item Respons Theorie (IRT) modellen kan licht werpen op de precieze aard van deze effecten.

De effecten van stereotype bedreiging (Hoofdstuk 3) worden in de regel gezien als meetartefacten en vallen onder effecten op Niveau III wanneer sprake is van subtest gerelateerde effecten en onder Niveau IV wanneer sprake is van effecten die specifiek zijn voor bepaalde items.

De verschillende variabelen die in de literatuur zijn geopperd ter verklaring van het Flynn Effect kunnen eveneens worden gezien in termen van de verschillende niveaus. De studies in het vierde hoofdstuk wezen erop dat het Flynn Effect deels kan worden toegewezen aan effecten op het derde en vierde niveau. Er is daarom meer onderzoek nodig om deze effecten in kaart te brengen.

De relatief lage scores van zwarte Afrikanen op westerse IQ tests kunnen door effecten op alle niveaus van de hiërarchie zijn veroorzaakt. Omdat er een gebrek is aan goede grondige psychometrische analyses in deze context, kan er niet zondermeer geconcludeerd worden dat deze lage scores een reflectie zijn van lage gemiddelde g, zoals Lynn wel heeft gedaan. Meer onderzoek in deze context is nodig om de aard van lage IQ scores van Afrikanen juist te kunnen interpreteren.

Gezien de structuur van individuele verschillen in cognitieve capaciteiten moet de voorkeur worden gegeven aan het gebruik van grondige psychometrische modellen die een weergave zijn van de theorieën die worden onderzocht.

Acknowledgements/dankwoord

De productie van de hoofdstukken in dit proefschrift kan als volgt worden beschreven:

Lezen, nadenken, praten, iets opzoeken, databestanden verwerken, analyseren, iets checken, nog wat analyseren, naar de film, praten, nadenken, lezen, in korte tijd het belangrijkste idee optekenen, iemand iets mailen, discussiëren, iets opzoeken, lezen, terug mailen, praten, (uit) eten met je vriendin, weer wat analyseren, nog iets lezen, overleggen, nadenken op de fiets, nog een keer analyseren, tabellen maken, tekst schrijven, lachen met collega's, tekst schrijven, snel iets opzoeken in de bieb, nog meer tekst schrijven, borrelen aan de overkant, veel te laat wakker worden, teveel koffie drinken, te onrustig zijn om iets te schrijven, referenties checken, figuren maken, tekst schrijven, praten, tekst schrijven, weer eens afspreken met goede oude vriend, nog meer tekst schrijven, tekst nog een beetje bijschaven, tekst eindelijk mailen, tekst (snel) terugkrijgen, overleggen, tekst aanpassen, tekst en opmaak checken, artikel opsturen, borrel drinken aan de overkant en echt weer eens je moeder bellen.

Kortom, dit proefschrift had ik niet alleen kunnen schrijven.

Ik dank Conor Dolan voor zijn deskundige commentaar, zijn stimulerende inzichten, zijn technische hulp, zijn snelle correcties van mijn Engels, zijn relativerende humor en de vrijheid die hij me gaf. Een betere begeleider had ik me niet kunnen wensen.

Ik dank Han van der Maas voor het vervullen van de rol van promotor en voor het leveren van commentaar aan het eind van het project en ik dank Peter Molenaar voor de inspiratie in den beginne.

Mijn leermeester Harrie Vorst kan ik niet voldoende bedanken. De komende jaren blijf ik gewoon proberen hem een keer te trakteren op een Duvel in De Roeter.

Paranimf Michiel Waardenburg dank ik voor de goede vriendschap en paranimf AZS voor de deskundige gezelligheid.

Voor de fantastische werksfeer op kamer A529 ben ik dank verschuldigd aan AZS (hij staat uit!), Sophie van der Sluis ("s"), Ellen Hamaker (huuuh!), Michiel Hol (ik heb wel formules) en Willemijn Roorda (zie Hoofdstuk 5!).

Dave Hessen dank ik voor de prettige samenwerking en Sharon Klinkenberg voor zijn grafische hulp.

Voor de prettige en stimulerende werkomgeving bij PML wil ik verder dank betuigen aan collega's Denny Borsboom, Don Mellenbergh, Jan-Willem Romeyn, Lourens Waldorp, Pieter Koele, Katharina Kouwenhoven, Peter van Rijn, Jan Hoogeboom, Marijke Engels, Steven David, aan oud-collega's Maarten Speekenbrink, Johan Hoogstraten, Jaap van Heerden en Wulfert van den Brink, evenals de nieuwe collega's Marthe Straatemeier, Eric-Jan Wagenmakers, Daan Zult en natuurlijk Raoul Grasman.

Ineke van Osch is de beste secretaresse ter wereld.

Dorret Boomsma, Arne Evers, Henk Kelderman, Don Mellenbergh en Daniel Wigboldus dank ik hartelijk voor het zitting nemen in de promotiecommissie.

I would like to thank Jerry Carlson and Earl Hunt for initiating my work on Chapter 5. I am also indebted to Laurie O'Brien and Hannah-Hahn Nguyen for sharing their data for the studies reported in Chapter 3.

Paul Oosteveld, Arne Evers, Jules Stinissen, Dorret Boomsma, Caroline van Baal, Mark Span, Jan Hoogeboom, Harrie Vorst, Thijs van der Vossen, Daan Kramer, Michiel Waardenburg, Suzanne Kats, Kim de Crom, Annemarie Eigenhuis en de dames van OP4425 dank ik voor hun hulp bij de dataverzameling. Suzanne dank ik voor veel meer dan dat.

Eelco, bedankt voor de altijd interessante discussies.

Mijn vader geeft mij acht jaar na zijn dood nog bijna dagelijks het gevoel dat het met mij wel goed komt. Mijn moeder dank ik voor de steun, de trots en de interesse in wat ik doe.

Dit proefschrift draag ik op aan mijn ouders.

Jelte M. Wicherts
Amsterdam, 12 januari 2007