Available online at www.sciencedirect.com

SCIENCE DIRECT®

Intelligence 32 (2004) 509–537

INTELLIGENCE

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

Jelte M. Wicherts[a,*], Conor V. Dolan[a], David J. Hessen[a], Paul Oosterveld[a], G. Caroline M. van Baal[b], Dorret I. Boomsma[b], Mark M. Span[c]

[a]*Psychological Methods, Department of Psychology, University of Amsterdam, Roetersstraat 15, 1018 WB Amsterdam, The Netherlands*
[b]*Department of Biological Psychology, Free University of Amsterdam, Amsterdam, The Netherlands*
[c]*Test-Developer, Swets Test Publishers, Lisse, The Netherlands*

Received 11 June 2003; received in revised form 7 July 2004; accepted 8 July 2004

Abstract

The gains of scores on standardized intelligence tests (i.e., Flynn effect) have been the subject of extensive debate concerning their nature, causes, and implications. The aim of the present study 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 Wechsler Adult Intelligence Scale (WAIS) gains from 1967 to 1999, Dutch Differential Aptitude Test (DAT) gains from 1984 to 1995, gains on a Dutch children intelligence test (RAKIT) from 1982 to 1993, and reanalyses of results from Must, Must, and Raudik [*Intelligence 167* (2003) 1–11] and Teasdale and Owen [*Intelligence 28* (2000) 115–120]. The results of multigroup 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.

Keywords: Intelligence tests; Factorial invariance; Flynn effect

* Corresponding author. Tel.: +31 205256880.
E-mail address: J.M.Wicherts@uva.nl (J.M. Wicherts).

0160-2896/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2004.07.002

1. Introduction

Ever since Flynn (1984, 1987) documented worldwide increases in scores on standardized intelligence tests, there has been extensive debate about the nature, causes, and 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, 2000).

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 (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 United States). Based on his results, Flynn (1987, p. 189) questioned the validity of IQ tests and suggested that other between-group differences on IQ tests may not reflect true intelligence differences. Furthermore, Flynn (1998a, p. 40) states that "massive IQ gains add viability to an environmental hypothesis about the IQ gap between Black and White Americans." 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 because it is primarily concerned with measured intelligence or IQ.

The purpose of the present paper 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 tests 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, Dolan, Kelderman, & Mellenbergh, 2003; Meredith, 1993). In addition, factorial invariance implies measurement invariance with respect to cohort (Meredith, 1993), which, in turn, means that the

intelligence test is unbiased with respect to cohort (Mellenbergh, 1989). We use multigroup 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. (2003) and Little (1997). MGCFA addresses within- (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 cohort (between group) differences are differences on the same underlying constructs (Lubke et al., 2003). 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., 2003).

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, 2000). 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 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 (pp. 320–321), Rushton, and Must et al. 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 (B–W) differences, obtained contradictory results. In addition, Colom et al. (2001) report high positive correlations using the standardization data of the Spanish Differential Aptitude Test (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, Dolan, & Kelderman, 2001). Indeed, this correlation may assume values that are interpreted in support of the importance of *g*, while in fact, MGCFA indicates that factorial invariance is not tenable (Dolan, Roorda, & Wicherts, 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 B–W 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.

2. Testing factorial invariance with MGCFA

MGCFA can be applied to address the question whether differences in IQ test score between groups reflect true, that is, latent, differences in ability (Lubke et al., 2003). We now present in detail the confirmatory factor model that can be used to this end (c.f. Bollen, 1989; Lubke et al., 2003; Sörbom, 1974).

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

$$\boldsymbol{y}_{ij} = \boldsymbol{\nu}_i + \boldsymbol{\Lambda}_i \boldsymbol{\eta}_{ij} + \boldsymbol{\varepsilon}_{ij} \ , \tag{1}$$

where $\boldsymbol{\eta}_{ij}$ is a q-dimensional random vector of correlated common factor scores ($q<p$) and $\boldsymbol{\varepsilon}_{ij}$ is a p-dimensional vector of residuals that contain both random error and unique measurement effects. The $(p\times q)$ matrix $\boldsymbol{\Lambda}_i$ contains factor loadings, and the $(p\times 1)$ matrix $\boldsymbol{\nu}_i$ contains measurement intercepts. It is generally assumed that $\boldsymbol{\varepsilon}_{ij}$ is p-variate normally distributed, with zero means and a diagonal covariance matrix $\boldsymbol{\Theta}_i$; that is, residual terms are mutually uncorrelated. Furthermore, the vector $\boldsymbol{\eta}_{ij}$ is assumed to be q-variate normally distributed, with mean α_i and $(q\times q)$ positive definite covariance matrix $\boldsymbol{\Psi}_i$. Given these assumptions, the observed variables are normally distributed $\boldsymbol{y}_{ij}\sim N_p(\mu_i,\Sigma_i)$, where, assuming the covariance between $\boldsymbol{\eta}_{ij}$ and $\boldsymbol{\varepsilon}_{ij}$ is zero:

$$\boldsymbol{\mu}_i = \boldsymbol{\nu}_i + \boldsymbol{\Lambda}_i\alpha_i \tag{2}$$

$$\boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}_i\boldsymbol{\Psi}_i\boldsymbol{\Lambda}_i^t + \boldsymbol{\Theta}_i. \tag{3}$$

Note that superscript t denotes transposition.

We identify a sufficient number of fixed zeroes in $\boldsymbol{\Lambda}_i$ to avoid rotational indeterminacy, given correlated common factors. In the same matrix $\boldsymbol{\Lambda}_i$, 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).

Factorial invariance can be investigated by fitting a series of increasingly restrictive models. These are presented in Table 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: $\boldsymbol{\Lambda}_1=\boldsymbol{\Lambda}_2$, factor loadings equal across cohorts; Horn & McArdle, 1992), and a model with equal factor loadings and equal residual variances (Model 3: $\boldsymbol{\Lambda}_1=\boldsymbol{\Lambda}_2$ and $\boldsymbol{\Theta}_1=\boldsymbol{\Theta}_2$). In the next two steps, we

Table 1
Summary of models in case of Cohorts 1 and 2

Number	Description	$\Sigma_1=$	$\Sigma_2=$	$\mu_1=$	$\mu_2=$
1	Configural invariance	$\boldsymbol{\Lambda}_1\boldsymbol{\Psi}_1\boldsymbol{\Lambda}_1^t+\boldsymbol{\Theta}_1$	$\boldsymbol{\Lambda}_2\boldsymbol{\Psi}_2\boldsymbol{\Lambda}_2^t+\boldsymbol{\Theta}_2$	$\boldsymbol{\nu}_1$	$\boldsymbol{\nu}_2$
2	Metric invariance	$\boldsymbol{\Lambda}\boldsymbol{\Psi}_1\boldsymbol{\Lambda}^t+\boldsymbol{\Theta}_1$	$\boldsymbol{\Lambda}\boldsymbol{\Psi}_2\boldsymbol{\Lambda}^t+\boldsymbol{\Theta}_2$	$\boldsymbol{\nu}_1$	$\boldsymbol{\nu}_2$
3	Equal residual variances	$\boldsymbol{\Lambda}\boldsymbol{\Psi}_1\boldsymbol{\Lambda}^t+\boldsymbol{\Theta}$	$\boldsymbol{\Lambda}\boldsymbol{\Psi}_2\boldsymbol{\Lambda}^t+\boldsymbol{\Theta}$	$\boldsymbol{\nu}_1$	$\boldsymbol{\nu}_2$
4a	Strict factorial invariance	$\boldsymbol{\Lambda}\boldsymbol{\Psi}_1\boldsymbol{\Lambda}^t+\boldsymbol{\Theta}$	$\boldsymbol{\Lambda}\boldsymbol{\Psi}_2\boldsymbol{\Lambda}^t+\boldsymbol{\Theta}$	$\boldsymbol{\nu}$	$\boldsymbol{\nu}+\boldsymbol{\Lambda}\delta$
4b	Strong factorial invariance	$\boldsymbol{\Lambda}\boldsymbol{\Psi}_1\boldsymbol{\Lambda}^t+\boldsymbol{\Theta}_1$	$\boldsymbol{\Lambda}\boldsymbol{\Psi}_2\boldsymbol{\Lambda}^t+\boldsymbol{\Theta}_2$	$\boldsymbol{\nu}$	$\boldsymbol{\nu}+\boldsymbol{\Lambda}\delta$

Except for Step 4b (nested under Step 2), each model is nested under the previous one.
Between-cohort differences in common factor means are expressed by δ (i.e., $\delta=\alpha_2-\alpha_1$).

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).[1] The latter involves the equality of intercepts ($\boldsymbol{\nu}_1=\boldsymbol{\nu}_2$), in addition to the equality of factor loadings and residual variances. Observed mean differences are then due to common factor mean differences: $m_2-m_1=\boldsymbol{\Lambda}(\alpha_2-\alpha_1)$. Strong factorial invariance does not include the equality constraint on the residual variances ($\boldsymbol{\Theta}_1\neq\boldsymbol{\Theta}_2$). 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., 2003; 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 ($\boldsymbol{\nu}_1=\boldsymbol{\nu}_2$), 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; Joreskog & Sörbom, 2002) using maximum likelihood (ML) estimation. We assess model fit by the χ^2 in relation to degrees of freedom and by other fit indices such as the RMSEA (Browne & Cudeck, 1993), the comparative fit index (CFI; Bentler, 1990), and the AIC and CAIC (c.f., Joreskog & Sörbom, 2002). The relative fit of the models in Table 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 0.05 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 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 multigroup 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

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

in degrees of freedom), we conclude that the particular constraint is untenable. If so, modification indices (MIs) can pinpoint the source, in terms of parameters, responsible for misfit. MIs are measures of how much the χ^2 is expected to decrease if a constraint on a given set of parameters is relaxed, and the model is refitted (Joreskog & Sörbom, 2002). We now turn to the confirmatory factor analyses of the five data sets.

3. Study 1: Dutch adults 1967/1968 and 1998/1999: WAIS

3.1. Samples

The Wechsler Adult Intelligence Scale (WAIS) was translated in Dutch more than 30 years ago (Stinissen, Willems, Coetsier, & Hulsman, 1970). Here, we compare the 1967/1968 standardization sample of the Dutch WAIS (N=2100) with 77 Dutch participants who completed the WAIS during the standardization of the WAIS-III in 1998 and 1999 (Wechsler, 2000). The mean age of the 1990s sample is 40.3 years (S.D.=14.0). In terms of the WAIS-III scores, this sample appears representative, with a mean WAIS-III IQ of 100.6 and an S.D. of 14.8 (Wechsler, 2000). However, it should be noted that the original Dutch WAIS-III standardization sample is slightly underrepresented with respect to participants from low-educational backgrounds (Swets Test Publishers, 2003; Tellegen, 2002). Therefore, these WAIS-III IQ's are an underestimation of approximately 2 IQ points (Swets Test Publishers, 2003).

In the 1998/1999 sample, the WAIS administration followed between 2 and 12 weeks after the 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 United States. 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 because 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.

3.2. Measures

The WAIS contains 11 subtests: Information (INF), Comprehension (COM), Arithmetic (ARI), Similarities (SIM), Digit Span (DSP), Vocabulary (VOC), Digit Symbol (DSY), Picture Completion (PCO), Block Design (BDE), Picture Arrangement (PAR), and 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, and SIM), perceptual organization (PCO, PAR, BDE, OAS, and DSY), and memory/freedom from distractibility (DSP, ARI, and DSY). This factor model is displayed in Fig. 1.

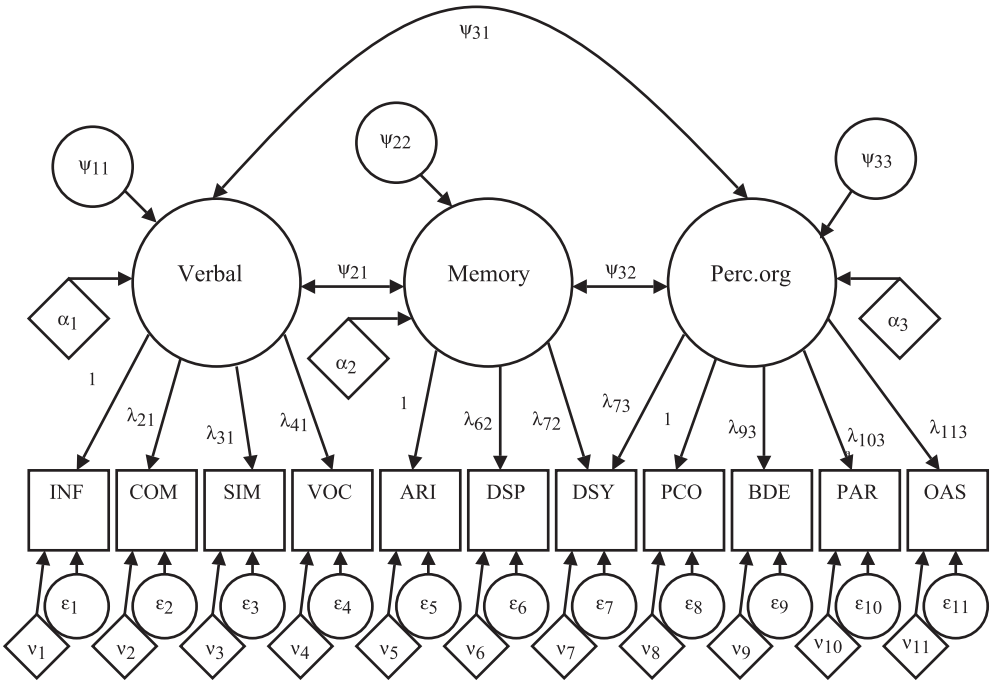

Fig. 1. Factor structure of WAIS.

3.3. Results and discussion

Correlations between subtests, as well as means and standard deviations of both cohorts, are reported in Table 2.[2] As can be seen from the mean differences between cohorts, the Flynn effect is present on all subtests, with effect sizes (in 1967/1968 S.D. 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, Performance, and Total IQs, respectively. These IQ gains are in line with gains on the WAIS(-R) found in the United States and in Germany (Flynn, 1998c; Satzger, Dragon, & Engel, 1996).

The fit indices of the factor models differing with respect to between-cohort equality constraints are reported in Table 3. The model with identical configuration of factor loadings in both cohorts (Model 1: configural invariance) fits poorly in terms of χ^2. However, the large χ^2 is due to the large standardization sample (χ^2 is highly sensitive to sample size; Bollen & Long, 1993), and the RSMEA and 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., $\mathbf{\Lambda}_1=\mathbf{\Lambda}_2$). All fit indices indicate that this 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: $\mathbf{\Theta}_1=\mathbf{\Theta}_2$) is not completely tenable because AIC and $\Delta\chi^2$ indicate a clear deterioration in fit as compared with the metric invariance model. However, RMSEA, CFI, and CAIC indicate that this restriction is tenable. In a formal sense, residual

[2] In this paper, we include summary statistics with which 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 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	
Mean 1967/1968	9.10	14.13	7.60	10.93	11.17	27.63	47.47	9.53	13.27	10.73	36.00
S.D. 1967/1968	5.44	5.52	3.80	5.55	3.33	12.10	12.83	3.54	6.42	4.45	14.88
Mean 1998/1999	13.78	20.84	11.10	19.14	12.88	40.22	58.58	13.51	20.41	14.38	44.65
S.D. 1998/1999	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

Correlations of the 1967/1968 sample (N=1100) below diagonal and 1998/1999 sample (N=77) above diagonal. Effect sizes are in 1967/1968 S.D. units.

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 (ν_1=ν_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 with 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 χ^2 comparing Models 4a and 4b to less restricted models is very large.[3] Both strong and strict factorial invariances 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 MIs (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 (χ^2=301.3, df=95, RMSEA=0.044, CFI=0.993, AIC=414, and CAIC=809). This enables a cautionary interpretation of the factor mean gains ($\alpha_2-\alpha_1$) thus found. The parameter estimates of the gains in this partial invariance model are memory/freedom from distractibility: 2.34 (S.E.=0.25, Z=9.20, P<.01); verbal comprehen-

[3] 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 χ^2. Even if, say, the χ^2 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.

Table 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** and **Θ**	332.3	100	3 vs. 2	52.6	11	0.044	0.993	421	782
4a	**Λ**, **Θ**, and **ν**	408.5	108	4a vs. 3	76.2	8	0.050	0.990	494	801
4b	**Λ** and **ν**	368.1	97	4b vs. 2	88.4	8	0.051	0.991	484	865

sion: 5.11 (S.E.=0.49, Z=10.53, P<.01); perceptual organization 3.69 (S.E.=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, whose 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 $\mathbf{\Psi}_1=\mathbf{\Psi}_2$; $\Delta\chi^2$=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 factor variance differences. This second-order model has the following fit indices: χ^2=321.5, *df*=102, RMSEA=0.044, CFI=0.993, AIC=423, and 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. Study 2: Danish draftees 1988 and 1998: Børge Prien's Prøve

4.1. 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 include all Danish draftees of 1988 (N=33.833) and 1998 (N=25.020), comprising about 90% to 95% of the Danish male population of 18-year-olds of those years (Teasdale & Owen, 1989, 2000).

4.2. 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) ML estimation, although the data are slightly negatively skewed (Teasdale & Owen, 2000) because the ML estimation is quite robust to mild skewness.

4.3. Results and discussion

The descriptive statistics of both cohorts are reported in Table 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 S.D. units. Furthermore, it is noteworthy that the standard deviations of all subtests, except the Geometric Figures Test, have decreased in the 10-year period. Teasdale and Owen 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.

Teasdale and Owen (2000, p. 117) state that the similarity of test intercorrelations across both cohorts is striking. We now use these data to test whether factorial invariance with respect to cohorts is tenable. This enables us to unravel whether the Danish gains reflect true (i.e., latent) gains in intelligence. Table 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 constrained to be cohort invariant (i.e., $\mathbf{\Lambda}_1=\mathbf{\Lambda}_2$). This step is accompanied by a relative improvement in fit, with all fit indices having better values in Model 2 than in 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: $\mathbf{\Theta}_1=\mathbf{\Theta}_2$) are somewhat inconsistent. The RMSEA indicates an improvement in the model fit from Model 2 to 3, while $\Delta\chi^2$, CFI, AIC, and CAIC show deterioration in fit. The highest modification

Table 4
Correlations and descriptive statistics of BPP 1988–1998

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

Correlations of 1988 sample (N=33,833) below diagonal and of 1998 sample (N=25,020) above diagonal. Effect sizes are in 1988 S.D. units.

Table 5
Fit indices test for factorial invariance of BPP 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 and Θ	547.1	11	3 vs. 2	71.6	4	0.040	0.949	565	735
4a	Λ, Θ, and ν	782.4	14	4a vs. 3	235.3	3	0.043	0.926	797	936
4b	Λ and ν	710.1	10	4b vs. 2	234.6	3	0.048	0.932	734	913

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 the cohort-invariant measurement intercepts (i.e., $\nu_1=\nu_2$) leads to a clear deterioration in fit, with all fit indices assigning poorer values in Models 4a and 4b as opposed to Models 3 and 2 (i.e., models without this mean restriction). Therefore, both strong (Model 4a) and strict factorial invariances (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 that 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 were 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 the case of 1000 participants 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 MIs 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 do. 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, and 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 $\alpha_2-\alpha_1$ is 0.17 (S.E.=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 increase in intelligence. Furthermore, the partial strict factorial invariance model shows that the (latent) factor variance in the second cohort is smaller (3.67, S.E.=0.047) than the factor variance of the first cohort (3.96, S.E.=0.046). The latter is consistent with earlier findings (Teasdale & 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 are 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 have 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 BPP 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.

5. Study 3: Dutch high school students 1984 and 1994/1995: DAT 1983

5.1. 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 participants were not sampled to be representative. Nevertheless, the latter data stem from 10 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 of 922 participants, of which 490 were females (of 11 participants, gender was unknown). Because Evers and Lucassen found large sex differences on the DAT, we randomly selected 93% of the females to equal the gender proportion of the three 1990s 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, 7 of the 10 schools in the 1990s cohort also participated in the 1984 standardization. Thus, the representativeness of the 1990s samples seems mainly to be compromised by the omission of participants from large-sized towns such as Amsterdam. The precise age of the participants during testing is unknown, but the mean would normally lie around 14.5 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.

5.2. Measures

The Dutch DAT 1983 (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). Because two subtests were not deemed informative by the school authorities, a significant part of the 1990s sample was not administered the Mechanical Reasoning (MR; 40% missing) and/or the Speed and Accuracy (SA) subtests (64% missing). This resulted in a shortening of the testing session for these participants, but this appears to not have resulted in higher scores on the remaining subtests. Probably because of the breaks in between subtests, the scores of these participants on the subtests that would have followed MR and SA did not

significantly differ from the corresponding scores of participants 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 (VO, SP, LU, VR, and NA) and an abstract factor (VR, AR, SR, and NA).

5.3. Results and discussion

We now present results for each educational level separately, beginning with the lowest level. The means, standard deviations, and inter-subtest correlations of both MAVO cohorts are reported in Table 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 1990s cohort, such as the aforementioned lack of participants 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). Because 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=.90; Spearman=.80) with the Spanish effect sizes found by Colom et al. (1998, 2001).

Because 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

Table 6
Correlations and descriptive statistics of DAT 1983 1984–1995 medium–low level (MAVO)

	VO	SP	LU	VR	AR	SR	NA
VO		.13	.51	.33	.13	.18	.11
SP	.23		.19	.11	.02	−.04	.10
LU	.55	.32		.26	.12	.19	.09
VR	.36	.17	.35		.34	.38	.24
AR	.27	.08	.27	.40		.58	.44
SR	.28	−.04	.21	.39	.52		.35
NA	.25	.16	.18	.32	.42	.38	
Mean 1984	42.5	59.2	29.4	18.7	33.3	30.7	17.7
S.D. 1984	10.2	8.4	6.9	7.2	7.3	9.3	6.0
Mean 1994/1995	39.89	58.97	27.05	17.32	32.61	30.76	15.14
S.D. 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

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

Table 7
Correlations and descriptive statistics of DAT 1983 1984–1995 medium–high level (HAVO)

	VO	SP	LU	VR	AR	SR	NA
VO		.29	.53	.32	.10	.13	.02
SP	.31		.35	.09	.03	−.09	.04
LU	.51	.36		.39	.20	.18	.03
VR	.28	.16	.33		.43	.38	.19
AR	.10	.04	.17	.36		.61	.42
SR	.18	.00	.13	.37	.53		.32
NA	.12	.13	.13	.23	.35	.29	
Mean 1984	49.8	64.5	34.5	23.6	37.5	35.9	22.2
S.D. 1984	9.7	8.3	6.5	8.6	6.2	10.3	6.0
Mean 1994/1995	48.78	66.68	35.55	23.87	37.80	37.05	20.07
S.D. 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

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

tenability of factorial invariance of the MAVO cohorts are in line with the following results of the HAVO cohorts.

The subtest correlations, as well as the descriptives, of both medium–high level (HAVO) cohorts are reported in Table 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 .82 (Spearman=.80). Because it has been suggested that the Spanish DAT gains are compatible with increases in g ("a Jensen effect"[4]; see Colom et al., 2001), it is interesting to check whether the HAVO gains can be considered factorially invariant with respect to cohort, because factorial invariance is a crucial aspect of the hypothesis that the manifest gains are due to gains in g.

Fit indices of the models leading up to factorial invariance in the HAVO comparison are reported in Table 8. The first model fits sufficiently, as judged by RMSEA and CFI. The step from the configural to the metric invariance model (Model 2: $\mathbf{\Lambda}_1=\mathbf{\Lambda}_2$) 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: $\mathbf{\Theta}_1=\mathbf{\Theta}_2$), 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 with Models 3 and 2, both factorial invariance models (4a and 4b) show a clear decline in all fit indices (although CAICs in Models 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 ($\boldsymbol{\nu}_1=\boldsymbol{\nu}_2$) is untenable and, therefore, that the Dutch increase in DAT test scores at this educational level cannot be explained by increases in latent intelligence.

[4] A Jensen effect occurs whenever g loadings of (sub)tests correlate significantly with the (sub)tests' correlations with other variables.

Table 8
Fit indices test for factorial invariance of DAT 1983 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** and **Θ**	86.2	36	3 vs. 2	16.2	7	0.045	0.978	154	366
4a	**Λ**, **Θ**, and **ν**	153.3	41	4a vs. 3	67.1	5	0.063	0.952	210	390
4b	**Λ** and **ν**	136.2	34	4b vs. 2	66.2	5	0.065	0.958	204	428

Here, we again consider the partial strong factorial invariance model and relax the intercepts associated with the largest MIs. 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, and CAIC=390). In this partial strong factorial invariance model, the factor mean of the verbal factor is significantly higher in the 1994/1995 as opposed to the 1984 sample (1.46, S.E.=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, S.E.=0.38, *Z*=0.97, *P*>.05).

Next, we turn to the highest educational level, denoted VWO. Descriptive statistics and subtest correlations of both VWO cohorts are reported in Table 9. As was the case in the medium–low educational level (MAVO) above, the 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=.79, Spearman=.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

Table 9
Correlations and descriptive statistics of DAT 1983 1984–1995 high level (VWO)

	VO	SP	LU	VR	AR	SR	NA
VO		.36	.57	.40	.11	.22	.16
SP	.39		.37	.25	.19	.12	.24
LU	.56	.45		.38	.15	.11	.15
VR	.38	.32	.44		.45	.42	.41
AR	.15	.14	.22	.35		.54	.38
SR	.19	.08	.15	.39	.53		.33
NA	.20	.22	.19	.29	.31	.33	
Mean 1984	56	70.9	39.9	30.1	40	40	26.3
S.D. 1984	9.4	8.7	7.2	8.8	5.3	9.7	5.8
Mean 1994/1995	51.12	69.37	37.18	24.45	40.03	38.93	23.86
S.D. 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

Correlations of 1984 sample (*N*=1100) below diagonal and of 1994/1995 sample (*N*=188) above diagonal. Effect sizes are in 1984 S.D. units.

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.

6. Study 4: Dutch children 1981/1982 and 1992/1993: RAKIT

6.1. Samples

In this study, we compare 5-year-olds from the 1981/1982 standardization sample of the RAKIT (Bleichrodt, Drenth, Zaal, & Resing, 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 socioeconomic 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; NCBS, 2003). Nevertheless, the 1990s 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. Because 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 2, the second twin then is assigned to the other twin sample. The twin data provide a useful opportunity to cross validate the results of model fitting, in which the 1982 cohort is compared with both Twin Samples 1 and 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.

6.2. Measures

The RAKIT (Bleichrodt et al., 1984) is an individually administered Dutch intelligence test for children (aged 4 to 11 years) composed of 12 subtests. RAKIT IQ has been shown to correlate .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 .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).

6.3. Results and discussion

The descriptive statistics of the standardization sample, as well as both twin samples, are reported in Table 10. 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 in 1992/1993 is reflected in average IQs of 102.6 (S.D.=13.7) and 103.0 (S.D.=12.6) in Twin Samples 1 and 2, respectively. Considering the somewhat higher SES of the 1990s sample, these gains appear small in comparison with gains found on the WISC-R in the U.S. (i.e., 5.3 IQ points from 1972 to 1989; Flynn, 1998c) and German WISC (20 IQ points from 1956 to 1983; Schallberger, 1987).

The fit indices of the various models for both twin samples are reported in Table 11. The first model (i.e., configural invariance) fits well in both the comparisons containing Twin Samples 1 and 2. With the exceptions of a minor decrease in the CFI values of both samples, and a small increase in RMSEA in the

Table 10
Correlations and descriptive statistics of RAKIT 1982–1992/1993

	EX	VM	DI	LN	HF	IP
EX		.34	.40	.34	.30	.14
VM	.34		.12	.52	.24	.30
DI	.39	.28		.15	.30	.10
LN	.24	.40	.06		.19	.33
HF	.39	.30	.28	.26		.08
IP	.13	.36	.19	.31	.24	
Mean 1982	15.01	15.17	14.95	14.97	15.37	14.94
S.D. 1982	5.02	5.10	4.99	4.97	5.06	4.99
Mean 1992-1	15.50	16.00	13.60	16.63	16.30	15.36
S.D. 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
EX		.35	.32	.19	.29	.15
VM			.10	.46	.26	.20
DI				.07	.18	.14
LN					.24	.26
HF						.10
Mean 1992-2	15.68	15.76	14.34	16.68	16.37	15.13
S.D. 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

Correlations of 1982 sample (N=207) below diagonal and of 1992/1993 sample (N=208) above diagonal. Effect sizes are in 1982 S.D. units.
Second sample N=207.

Table 11
Fit indices test for factorial invariance of RAKIT 1982–1993/1994

Model	Equality constraints	χ^2	*df*	Compare	$\Delta\chi^2$	Δdf	RMSEA	CFI	AIC	CAIC
1st sample										
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 and Θ	47.4	26	3 vs. 2	17.3	6	0.062	0.961	102	243
4a	Λ, Θ, and ν	68.5	30	4a vs. 3	21.2	4	0.078	0.929	116	236
4b	Λ and ν	50.7	24	4b vs. 2	20.7	4	0.072	0.952	109	260
2nd sample										
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 and Θ	38.7	26	3 vs. 2	8.6	6	0.049	0.973	95	236
4a	Λ, Θ, and ν	54.2	30	4a vs. 3	15.6	4	0.064	0.947	104	224
4b	Λ and ν	45.5	24	4b vs. 2	15.4	4	0.068	0.953	107	257

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: The 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., $\nu_1=\nu_2$) is untenable. Again, it appears that the 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: χ^2=31.43, *df*=22, RMSEA=0.042, CFI=0.985, AIC=94, and CAIC=255 and χ^2=31.09, *df*=22, RMSEA=0.045, CFI=0.982, AIC=95, and 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, S.E.=0.43, Z=1.61, P>.05 and 0.80, S.E.=0.44, Z=1.82, P>.05 for the nonverbal and the verbal factors, respectively). However, in the second twin sample, the factor mean of the nonverbal factor is significantly higher than the standardization sample (0.85, S.E.=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, S.E.=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 that 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.

7. Study 5: Estonian children 1934/1936 and 1997/1998: National Intelligence Test

7.1. Samples

The data from this last comparison stem from Must et al. (2003), who compared two Estonian data sets 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. found gains on most of the subtests, which were not consistent with the Jensen effect. It is interesting to submit these Estonian data to the MGCFA approach because 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 1930s and 1990s cohorts, respectively. For further information on the samples, the reader is referred to Must et al.

7.2. 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.

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, and AN). This model is used in the fitting of the subsequent models.

7.3. Results and discussion

Table 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, the 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. Because this decline may also be due to a decrease in the latent factor(s), we proceed with the analyses.

Table 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 does. All fit measures, except the CAIC, show deteriorating fit. Therefore, factor loadings cannot be considered cohort invariant (i.e., $\mathbf{\Lambda}_1 \neq \mathbf{\Lambda}_2$). 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

Table 12
Correlations and descriptive statistics of the 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	
Mean 1934/1937	16.92	24.45	27.21	25.26	35.60	25.65	26.62	13.86	24.28	27.10
S.D. 1934/1937	4.47	5.27	6.45	6.70	8.27	5.51	12.92	5.78	6.63	8.42
Mean 1997/1998	14.53	22.26	29.83	19.20	39.14	24.84	29.52	17.28	30.04	33.00
S.D. 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

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

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 nonuniform bias with respect to cohorts (Lubke et al., 2003). Consequently, we present the next steps for illustrative reasons only. In fitting Model 3 ($\boldsymbol{\Theta}_1=\boldsymbol{\Theta}_2$), 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., $\boldsymbol{\nu}_1\neq\boldsymbol{\nu}_2$). The latter is primarily caused by the Information subtest. Because of the large number of parameters that show large MIs in all nonfitting 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.

Table 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	$\boldsymbol{\Lambda}$	209.6	74	2 vs. 1	58.9	10	0.074	0.978	324	634
3	$\boldsymbol{\Lambda}$ and $\boldsymbol{\Theta}$	316.2	84	3 vs. 2	106.5	10	0.088	0.964	400	655
4a	$\boldsymbol{\Lambda}$, $\boldsymbol{\Theta}$, and $\boldsymbol{\nu}$	1147.5	92	4a vs. 3	831.3	8	0.185	0.831	1250	1460
4b	$\boldsymbol{\Lambda}$ and $\boldsymbol{\nu}$	1029.1	82	4b vs. 2	819.5	8	0.183	0.853	1120	1386

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.

8. General discussion

The present aim was to determine whether the 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 (ν_1=ν_2; Models 4a and 4b; see Table 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 that the 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 extent, 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. Because 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. The measurement invariance can also be investigated at the item level. Flieller (1988) compared two cohorts of French 8-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 found that two thirds of the 64 items were biased with respect to cohort. That is, the majority of the 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. (2003) have discussed how covariates can be incorporated in a multigroup 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 (e.g., Mellenbergh & Kok, 1991), such as correlational, quasi-experimental, and experimental researches. 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 participants nowadays more readily resort to guessing than participants in earlier times did, 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 because 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 .60 (WAIS data) and .58 (RAKIT data). We know that in both data sets, factorial invariance is not tenable. Yet, correlations of about .60 are invariably interpreted in support of the importance of *g*. For instance, the repeated application of the correlated vectors method to B–W differences in intelligence test scores resulted in a mean correlation of about .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 (2000) 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). Because the current approach (MGCFA) was previously applied in U.S. 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 Models 1 to 4a within each of the seven data sets to compare the results of the tests of factorial invariance on the Flynn effects and the racial groups. These standardized AIC values are reported in Fig. 2.

As can be seen, the relative AIC values of the five Flynn comparisons show a strikingly similar pattern. In these cohort comparisons, Models 1 and 2 have approximately similar standardized AICs, which indicates that the 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 the measurement invariance between Blacks and Whites is tenable because 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 B–W differences in the United States. 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

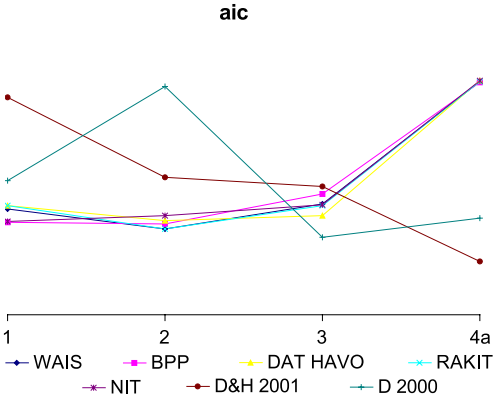

Fig. 2. Plot of standardized AIC values of data sets by stepwise models to achieve strict factorial invariance.

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 the 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 paper is that factorial invariance with respect to cohorts is not tenable. Clearly, this finding requires replication in other data sets. 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.

Acknowledgements

The preparation of this article was supported by a grant from the Netherlands Organization for Scientific Research (NWO). We thank Arne Evers and Jules Stinissen for providing valuable information on outdated standardization samples, and we express our appreciation for the valuable comments on previous drafts by Drs. Flynn, Widaman, and Jensen.

Appendix A. Description of the WAIS subtests

Information (INF): 22 open-ended questions measuring general knowledge concerning events, objects, people, and place names.

Comprehension (COM): 14 daily-life or societal problems that the participant has to understand, explain, or solve. For this, the subject needs to comprehend social rules and concepts.

Arithmetic (ARI): 16 arithmetic items that the participant has to solve without the use of paper and pencil.

Similarities (SIM): 13 word pairs about daily objects and concepts. The participant has to explain the similarities of the words.

Digit Span (DSP): 14 series of digits that participants have to recall verbally forwards (12 items) or backwards (2 items).

Vocabulary (VOC): 30 words, of which the participant has to give the meaning.

Digit Symbol (DSY): 115 items containing pairs of numbers and symbols. The participant uses a key to write down the symbol related to a number.

Picture Completion (PCO): 20 incomplete pictures of everyday events and objects about which the participant has to name the missing parts.

Block Design (BDE): 13 two-dimensional geometric figures that the participant has to copy by arranging two-colored blocks.

Picture Arrangement (PAR): 10 items in which pictures have to be arranged in a logical order.

Object Assembly (OAS): five puzzles of everyday objects that the participant has to assemble.

Source: Wechsler (1955, 2000) and Stinissen et al. (1970)

Appendix B. Description of the subtests from BPP

Letter Matrices (LEM): 19 items (15 min) in a 3×3 matrix format, with cells containing series of letters conforming to a pattern. The participant has to give the letter series that conforms to this pattern.

Verbal Analogies Test (VAT): 24 verbal analogies that the participant has to complement (5 min). The answers have to be chosen from a two lists of 100 possible responses.

Number Series Test (NST): 17 series of four numbers that the participant has to complement (15 min).

Geometric Figures Test (GFT): 18 items (10 min) containing complex geometric figures that have to be composed by five simple figures.

Source: Teasdale and Owen (1987, 1989, 2000)

Appendix C. Description of the DAT 1983 subtests

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 a 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.

Source: Evers and Lucassen (1992)

Appendix D. Description of the RAKIT subtests

Exclusion (EX): contains 30 items in which the child has to choose one out of four figures that is deviant. Measures inductive reasoning.

Discs (DI): contains 12 items in which the child has to put discs with holes on sticks. 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. Measures transformation of a visual field.

Verbal Meaning (VM): contains 40 words whose meaning the child has to denote by pointing out one out of four pictures. Measures passive verbal learning.

Learning Names (LN): contains 10 pictures of animals whose names the child has to learn. Measures active learning.

Idea Production (IP): contains five items in which the child has to produce names of objects and situations that belong to a broadly described category. Measures verbal fluency.

Source: Bleichrodt et al. (1984)

Appendix E. Description of the National Intelligence Test subtests

Arithmetic (AR): 16 arithmetic problems that require a solution for an unknown quantity.

Computation (CT): 22 items requiring addition, subtraction, multiplication, and division of both integers and fractions.

Sentence Completion (SC): 20 items requiring filling in missing words to make sentences understandable and correct.

Information (IN): 40 items about general knowledge.

Concepts (CC): 24 items requiring selecting two characteristic features from among those given.

Vocabulary (VO): 40 items requiring knowledge about the qualities of different objects.

Synonyms–Antonyms (SA): 40 items requiring the evaluation of whether the words presented mean the same or opposite.

Analogies (AN): 32 items requiring transferring the relation between two given words to other presented words.

Symbol–Number (SN): 120 items in which the correct digit must be assigned to a presented symbol from a key.

Comparisons (CP): 50 items requiring same or different judgments about sets of numbers, family names, and graphic symbols presented in two columns.

Source: Must et al. (2003)

References

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

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 and Zeitlinger.

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). Thousand Oaks, CA: Sage Publications.

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.

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). Thousand Oaks, CA: Sage Publications.

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.

Catron, D. W., & Thompson, C. C. (1979). Test–retest gains in WAIS scores after four retest intervals. *Journal of Clinical Psychology*, *35*, 352–357.

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.

Davenport, E. C. (1990). Significance testing of congruence coefficients: A good idea? *Educational and Psychological Measurement*, *50*, 289–296.

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

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., & 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 Frank Columbus (Ed.), *Advances in Psychology Research*, *vol. 6* (pp. 31–59). Huntington, NY: Nova Science Publishers.

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.

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.

Evers, A., & Lucassen, W. (1992). Handleiding DAT'83 (DAT '83 Manual). Lisse, The Netherlands: Swets and Zeitlinger.

Flieller, A. (1988). Application du modele de Rasch a un probleme de comparaison de generations [Applications 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 *g*s 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.

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*, *vol. 233* (pp. 202–227). Chichester, UK: Wiley.

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, US: American Psychological Association.

Horn, J. L., & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research*, *18*, 117–144.

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.

Husén, T., & Tuijnman, A. (1991). The contribution of formal schooling to the increase in intellectual capital. *Educational Researcher*, *20*, 17–25.

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.

Jöreskog, K. G. (1993). Testing structural equation models. In K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models* (pp. 294–316). Thousand Oaks, CA: Sage Publications.

Jöreskog, K. G., & Sörbom, D. (2002). *LISREL 8: Structural equation modeling with the SIMPLIS command language*. Chicago, IL: Scientific Software International.

Little, T. D. (1997). Mean and covariance structures (MACS) analyses of cross-cultural data: Practical and theoretical issues. *Multivariate Behavioral Research*, *32*, 53–76.

Lubke, G. H., & Dolan, C. V. (2003). Can unequal residual variances across groups mask measurement invariance 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. (2003). On the relationship between sources of within- and between-group differences and measurement invariance in the common factor model. *Intelligence*, *31*, 543–566.

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., & Hampson, S. (1986). The rise of national intelligence: Evidence from Britain, Japan and the USA. *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.

Martorell, R. (1998). Nutrition and the worldwide rise in IQ scores. In Ulric Neisser (Ed.), *The rising curve: Long term gains in IQ and related measures* (pp. 183–206). Washington, DC, US: 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.

Mellenbergh, G. J. (1989). Item bias and item response theory. *International Journal of Educational Research*, *13*, 127–143.

Mellenbergh, G. J., & Kok, F. G. (1991). Finding the biasing traits. In J. M. Collins (Ed.), *Advances in computer-based human assessment* (pp. 291–306). Dordrecht, the Netherlands: Kluwer Academic Publishers.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika*, *58*, 525–543.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods*, *2*, 248–260.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied Psychological Measurement*, *17*, 297–334.

Mingroni, M. A. (2004). The secular rise in IQ: Giving heterosis a closer look. *Intelligence*, *32*, 65–83.

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.

Muthén, B. O. (1989). Factor structure in groups selected on observed scores. *British Journal of Mathematical and Statistical Psychology*, *42*, 81–90.

NCBS. (2003). Beroepsbevolking naar onderwijsniveau 1993. Retrieved May 14th, 2003, from http://statline.cbs.nl

Neisser, U. (Ed.). (1998). *The rising curve: Long-term gains in IQ and related measures*. Washington, DC, US: American Psychological Association.

Oosterveld, P. (1996). Questionnaire design methods. Amsterdam, The Netherlands: University of Amsterdam.

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.

Rodgers, J. L. (1998). A critique of the Flynn effect: Massive IQ gains, methodological artifacts or both? *Intelligence*, *26*, 337–356.

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. (2000). Flynn effects not genetic and unrelated to race differences. *American Psychologist*, *55*, 542–543.

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.

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, US: American Psychological Association.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical & Statistical Psychology*, *27*, 229–239.

Spitz, H. H. (1989). Variations in Wechsler interscale IQ disparities at different levels of IQ. *Intelligence*, *13*, 157–167.

Steenkamp, J.B.E.M., & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–90.

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)*. Lisse, The Netherlands: Swets and Zeitlinger.

Swets Test Publishers. (2003). Aanvullend normonderzoek WAIS-III. Retrieved May 7th, 2003, from http://www.swetstest.nl/info/WAIS-III/

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.

Tellegen, P. J. (2002). De kwaliteit van de normen van de WAIS-III [Quality of the WAIS-III norms]. *De Psycholoog*, *37*, 463–465.

Tuddenham, R. D. (1948). Soldier intelligence in World Wars I and II. *American Psychologist*, *3*, 54–56.

Wechsler, D. (1955). *WAIS manual*. New York: The Psychological Corporation.

Wechsler, D. (2000). *WAIS-III.* Nederlandstalige bewerking. Technische handleiding. Lisse, The Netherlands: Swets Test Publishers.

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 (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.

Zajonc, R. B., & Mullally, P. R. (1997). Birth order: Reconciling conflicting effects. *American Psychologist*, *52*, 685–699.