Available online at www.sciencedirect.com
ScienceDirect

Intelligence 35 (2007) 283–300

INTELLIGENCE

Score gains on *g*-loaded tests: No *g*[☆]

Jan te Nijenhuis [a,*], Annelies E.M. van Vianen [a], Henk van der Flier [b]

[a] *Work and Organizational Psychology, University of Amsterdam, Amsterdam, The Netherlands*
[b] *Work and Organizational Psychology, Free University, Amsterdam, The Netherlands*

Received 10 December 2005; received in revised form 23 June 2006; accepted 11 July 2006
Available online 14 August 2006

Abstract

IQ scores provide the best general predictor of success in education, job training, and work. However, there are many ways in which IQ scores can be increased, for instance by means of retesting or participation in learning potential training programs. What is the nature of these score gains? Jensen [Jensen, A.R. (1998a). *The g factor: The science of mental ability*. London: Praeger] argued that the effects of cognitive interventions on abilities can be explained in terms of Carroll's three-stratum hierarchical factor model. We tested his hypothesis using test–retest data from various Dutch, British, and American IQ test batteries combined into a meta-analysis and learning potential data from South Africa using Raven's Progressive Matrices. The meta-analysis of 64 test–retest studies using IQ batteries (total $N=26{,}990$) yielded a correlation between *g* loadings and score gains of −1.00, meaning there is no *g* saturation in score gains. The learning potential study showed that: (1) the correlation between score gains and the *g* loadedness of item scores is −.39, (2) the *g* loadedness of item scores decreases after a mediated intervention training, and (3) low-*g* participants increased their scores more than high-*g* participants. So, our results support Jensen's hypothesis. The generalizability of test scores resides predominantly in the *g* component, while the test-specific ability component and the narrow ability component are virtually non-generalizable. As the score gains are not related to *g*, the generalizable *g* component decreases and, as it is not unlikely that the training itself is not *g*-loaded, it is easy to understand why the score gains did not generalize to scores on other cognitive tests and to *g*-loaded external criteria.

Keywords: *g*; IQ testing; *g* loading; Training; Coaching; Learning potential; Dynamic testing; South Africa; Test–retest; Score gains

1. Training and score gains

Scores on cognitive tests are the best general predictors of accomplishments in school and in the workplace, and it is predominantly the *g* component of the IQ tests that is responsible for this criterion-related validity (Ree & Earles, 1991; Ree, Earles, & Teachout, 1994; Thorndike, 1985). At the same time, IQ test scores can be increased by various forms of training. Kulik, Bangert-Drowns, and Kulik's (1984) meta-analysis on test preparation studies resulted in effect sizes on intelligence tests for practice and additional coaching of 0.25 S.D. and 0.51 S.D., respectively. Dynamic testing (Grigorenko & Sternberg, 1998) focuses on what children learn in a special training in an attempt to go beyond IQ scores. A general finding is that scores go up by 0.5 to 0.7 S.D. after a dynamic training (Swanson & Lussier, 2001). Ericsson and Lehmann (1996) report immense score increases after intensive training, for

[☆] Jan te Nijenhuis and Annelies van Vianen, Work and Organizational Psychology, University of Amsterdam, Amsterdam, the Netherlands, and Henk van der Flier, Work and Organizational Psychology, Free University, Amsterdam, the Netherlands.

[*] Corresponding author. Gouden Leeuw 746, 1103 KR Amsterdam, The Netherlands.

E-mail address: JanteNijenhuis@planet.nl (J. te Nijenhuis).

0160-2896/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2006.07.006

instance on a memory task very similar to the subtest Forward Digit Span of the WISC. It is clear that IQ scores can be increased by training. The question is what inferences can be drawn from these gains. Do they represent true increases in mental ability or simply in performance on a particular test instrument?

2. Jensen's hypothesis: score gains can be summarized in the hierarchical intelligence model

Jensen (1998a, ch. 10) hypothesized that the effects of training on abilities can be summarized in terms of Carroll's (1993) three-stratum hierarchical factor model of cognitive abilities. At the highest level of the hierarchy (stratum III) is general intelligence or *g*. One level lower (stratum II) are the broad abilities, Fluid Intelligence, Crystallized Intelligence, General Memory and Learning, Broad Visual Perception, Broad Auditory Perception, Broad Retrieval Ability, and Broad Cognitive Speediness or General Psychomotor Speed. One level lower still (stratum I) are the narrow abilities, such as Sequential Reasoning, Quantitative Reasoning, Verbal Abilities, Memory Span, Visualization, and Perceptual Speed. At the lowest level of the hierarchy are large numbers of specific tests and subtests. Some tests, despite seemingly very different formats, have been demonstrated empirically to cluster into one narrow ability (Carroll, 1993).

It is hypothesized that a training effect is most clearly manifested at the lowest level of the hierarchy of intelligence, namely on specific tests that most resemble the trained skills. One hierarchical level higher, the training effect is still evident for certain narrow abilities, depending on the nature of the training. However, the gain virtually disappears at the level of broad abilities and is altogether undetectable at the highest level, *g*. This implies that the transfer of training effects is strongly limited to tests or tasks that are all dominated by one particular narrow skill or ability. There is virtually no transfer across tasks dominated by different narrow abilities, and it disappears completely before reaching the level of *g*. Thus, there is an increase in narrow abilities or test-specific ability that is independent of *g*. Test-specific ability is defined as that part of a given test's true-score variance that is not common to any other test; i.e., it lacks the power to predict performance on any other tasks except those that are highly similar. Gains on test specificities are therefore not generalizable, but 'empty' or 'hollow'. Only the *g* component is highly generalizable. Jensen (1998a, ch. 10) gives various examples of empty score gains, including a detailed analysis of the Milwaukee project, claiming IQ scores rose, but not *g* scores. Another example of empty score gains is given by Christian, Bachnan, and Morrison (2001) who state that increases due to schooling show very little transfer across domains.

It is hypothesized that the *g* loadings of the few tests that are most similar to the trained skills and therefore most likely to reflect the specific training diminish after training. That is, after training, these particular tests reflect the effect of the specific training rather than the general ability factor.

It is one of the most firmly established facts in the social sciences that IQ tests have a high degree of predictive validity for educational criteria (Jensen, 1980; Schmidt & Hunter, 1998), meaning that high-*g* persons learn virtually always more than low-*g* persons. For instance, Kulik, Kulik, et al.'s (1984) meta-analysis reported practice effects on intelligence tests of 0.80 S.D., 0.40 S.D., and 0.17 S.D. for subjects of high, middle, and low ability, respectively. In industrial psychology, the more complex the training or job, the higher the correlation of performance with *g* (Schmidt & Hunter, 1998). This means that training or job situations, and also educational settings, vary in the degree to which they are *g*-loaded (Gottfredson, 1997, 2002). However, Ackerman (1987) cites several classical studies on the acquisition of simple skills through often repeated exercise where low-*g* persons made the most progress. These findings could be interpreted as an indication that this specific skill acquisition process is not *g*-loaded. It may be that some of the various forms of training referred to above also show the largest gains for low-*g* persons.

There are many ways to test Jensen's hypothesis. Below, we address (1) studies on repeated testing and *g* loadedness, (2) studies on practice and coaching, and (3) studies on learning potential. The practice studies used a pretest–posttest design, where both the coaching and learning potential studies used a pretest–intervention–posttest design.

3. First test of Jensen's hypothesis: studies on repeated testing and *g* loadedness

What do we find after repeated test taking? In a classic study by Fleishman and Hempel (1955) as subjects were repeatedly given the same psychomotor tests, the *g* loading of the tests gradually decreased and each task's specificity increased. Neubauer and Freudenthaler (1994) showed that after 9 h of practice the *g* loading of a modestly complex intelligence test dropped from .46 to .39. Te Nijenhuis, Voskuijl, and Schijve

(2001) showed that after various forms of test preparation the *g* loadedness of their test battery decreased from .53 to .49. Based on the work of Ackerman (1986, 1987), it can be concluded that through practice on cognitive tasks part of the performance becomes overlearned and automatic; the performance requires less controlled processing of information, which is reflected in lowered *g* loadings.

4. Second test of Jensen's hypothesis: studies on practice and coaching

Three studies on practice and coaching have shown increases in test scores that are not related to the *g* factor. This suggests that the gains are 'empty' or 'hollow'. In the first study, Jensen (1998a, ch. 10) analyzed the effect of practice on the General Aptitude Test Battery (GATB). He found negative correlations ranging from −.11 to −.86 between effect sizes on practice and the tests' *g* loadings. Therefore, the gains were largest on the least cognitively complex tests. In the second study, te Nijenhuis et al. (2001) found a small correlation of −.08 for test practice, and large negative correlations of −.87 for both of their two test coaching conditions. Jensen carried out a factor analysis of the various GATB score gains and found two large factors that did not correlate with the *g* factor extracted from the GATB. Most likely, the score gains are not on the *g* factor or the broad abilities, but on the test specificities, since te Nijenhuis et al. showed that practice and coaching reduce the *g*-loadedness of their tests. In a third study (Coyle, 2006), factor analysis demonstrated that the change in aptitude test scores had a zero loading on the *g* factor.

So, the studies on practice and coaching appear to support the theory. However, since there are only a few empirical studies that have tested the link (or absence thereof) between gains in test score from practice and coaching and *g* loadings, replications are required before the conclusion can be firmly established. Therefore, we combined several such studies with various Dutch, British, and American test batteries into a meta-analysis.

5. Third test of Jensen's hypothesis: studies on learning potential

Jensen hypothesizes that the effects of training are not on *g*, but that the gains are empty and training should therefore not lead to increased predictive validity. Based on learning potential theory, one would come to an opposite prediction, namely that training leads to higher predictive validity. The fact that the theoretical framework of learning potential does not include the *g* factor is of no importance here; we solely focus on a prediction based on learning potential theory that is opposite to a prediction based on Jensen's theory based on a hierarchical intelligence model. Some learning potential training studies report predictive validities of pre- and posttest scores. Based on Jensen's theory, one would predict (1) no higher predictive validity for learning potential tests in comparison with classical cognitive tests and (2) no increase in predictive validity due to training when using posttest scores instead of pretest scores. However, based on learning potential theory, one would predict a substantial increase in predictive validity in both cases. So, studies on learning potential constitute a test of Jensen's hypothesis.

A large number of studies have been carried out to check for learning potential beyond IQ scores, generally showing that scores go up substantially after mediation. Apart from theoretical considerations, dynamic tests should show higher criterion-related validities than classical IQ tests to justify the time-consuming procedure. Based on a lengthy review of most of the literature, Grigorenko and Sternberg (1998) concluded that the empirical data do not consistently show higher predictive power of dynamic tests compared with traditional tests. Murphy (2002) did an excellent and detailed review of all South African studies on learning potential, including virtually all missed by Grigorenko and Sternberg (probably due to difficulty of access). Many studies (Boeyens, 1989; de Villiers, 1999; Engelbrecht, 1999; Gaydon, 1988; Haeck, Yeld, Conradie, Robertson, & Shall, 1997; Lipson, 1992; Nel, 1997; Shochet, 1986; Skuy et al., 2002; Yeld, & Haeck, 1997; Zaaiman, 1998; Zaaiman, van der Flier, & Thijs, 2001; Zolezzi, 1992, 1995) used data from the numerous South African university entrance programs that had adopted a dynamic framework for assessing disadvantaged, underprepared students. The aim of these programs was to give underprepared applicants an optimal chance to prove that they have the ability to succeed with further study. Again it was found that while some South African studies show higher criterion-related validities for learning potential tests, the effect was not consistent.

However, in these studies, the learning potential tests were compared against individual tests or an unweighted combination of a limited number of tests, but generally not against a full test battery and in no case against *g* scores. *g* scores have been shown to yield higher predictive validities than individual tests or an unweighted score sum (Ree, & Earles, 1991; Ree et al., 1994; Thorndike, 1985). So, these were comparisons where the cognitive predictor with the highest predictive validity was not used, but where the dynamic tests were pitted against predictors with substantially lower predictive validities than *g*. As no direct comparisons were made between learning potential tests and *g* it is not possible to draw the conclusion that *g* had higher predictive validity. However, since a comparison of a learning potential test with one test or a combination of a limited number of tests generally results in comparable predictive validities, and *g* scores clearly have higher predictive validities than one test or a combination of a limited number of tests, it not unlikely that a *g* score will have a higher predictive validity than a learning potential test score. This also suggests that the findings might best be interpreted as tentative support for Jensen's theory.

So, the studies on learning potential appear to support the theory that score gains can be summarized in the hierarchical intelligence model. However, more direct tests of the theory are required and, therefore, a learning potential study was reanalyzed.

6. Research questions

The research question of this study is whether score gains from test–retest studies and mediated interventions can be summarized in terms of Carroll's three-stratum hierarchical intelligence model. We examined whether (1) correlations between score gains and the *g* loadedness of the scores are negative in sign, (2) the *g* loadedness of scores decreases after mediation, and (3) low-*g* persons show the largest gains after the mediation training. We carried out a meta-analysis to be able to provide a convincing answer to the first research question. In a more explorative study on learning potential in South Africa, we tried to find support for all three research questions.

7. Test–retest studies

To test whether there is a negative correlation between *g* loading of tests and score gains, we carried out a meta-analysis of all test–retest studies of Dutch, British, and American test batteries available in the Netherlands. All studies were simple practice studies–no intervention such as additional coaching took place–and used well-validated tests.

8. Method

Psychometric meta-analysis (Hunter & Schmidt, 1990) aims to estimate what the results of studies would have been if all studies had been conducted without methodological limitations or flaws. The results of perfectly conducted studies would allow a less obstructed view of the underlying construct-level relationships (Schmidt & Hunter, 1999). One of the goals of the present meta-analysis is to have a reliable estimate of the true correlation between standardized test–retest score gains (*d*) and *g*. Although the construct of *g* has been thoroughly studied, the construct underlying score gains is less well understood. One of the aims of the present study is to have a clearer understanding of the construct underlying score gains by linking it to the *g* nexus. Carrying out a complete meta-analysis on the relationship between *d* and *g* would require the collection of a very large number of datasets. However, applying meta-analytical techniques to a sufficiently large number of studies will also lead to a reliable estimate of the true correlation between *d* and *g*. We therefore collected a large number of studies heterogeneous across various possible moderators.

To get a reliable correlation between *g* and *d*, we focused on batteries with a minimum of seven subtests. Libraries and test libraries of universities were searched and several members of the Dutch Testing Commission and test publishers were contacted. We limited ourselves to non-clinical samples, without health problems. Only a minority of test manuals report test–retest studies; especially before 1970 they are rare. The search yielded virtually all test–retest studies available in the Netherlands. The GATB manual (1970, ch. 20) reports very large datasets on secondary school children who took the GATB with respectively 1-, 2-, and 3-year intervals. At the time of the first test, large samples of children that had the same age as the test–retest children at the time of the second test also took the test. Through a comparison of the scores, the maturation effects could be separated from the test–retest effects, so we included the data in the present study.

Standardized score gains were computed by dividing the raw score gain by the S.D. of the pretest. In general, *g* loadings were computed by submitting a correlation matrix to a principal axis factor analysis and using the loadings of the subtests on the first unrotated factor. In some cases, *g* loadings were taken from studies where other procedures were followed; these procedures have been shown empirically to lead to highly comparable results. Pearson correlations between the standardized score gains and the *g* loadings were computed.

Psychometric meta-analytical techniques (Hunter & Schmidt, 1990, 2004) were applied to the resulting 64 r_{gd}'s using the software package developed by Schmidt and Le (2004). Psychometric meta-analysis is based on the principle that there are artifacts in every dataset and that most of these artifacts can be corrected. In the present study, we corrected for five artifacts that alter the value of outcome measures listed by Hunter and Schmidt (1990): (1) sampling error, (2) reliability of the vector of g loadings, (3) reliability of the vector of score gains, (4) restriction of range of g loadings, and (5) deviation from perfect construct validity.

8.1. Correction for sampling error

In many cases, sampling error explains the majority of the variation between studies, so the first step in a psychometric meta-analysis is to correct the collection of effect sizes for differences in sample size between the studies.

8.2. Correction for reliability of the vector of g loadings

The value of r_{gd} is attenuated by the reliability of the vector of g loadings for a given battery. When two samples have a comparable N, the average correlation between vectors is an estimate of the reliability of each vector. The collection of datasets in the present study included no g vectors for the same battery from different samples and therefore artifact distributions were based upon other studies reporting g vectors for two or more samples. So, the effect sizes and the distribution of reliabilities of the g vector were based upon different samples. When two g vectors were compared the correlation between them was used, and when more than two g vectors were compared the average correlation for the various combinations of two vectors was used. The combined N from the samples on which the g vector was based was taken as the weight of one data point.

Several samples were compared that differed little on background variables. For the comparisons using children, we chose samples that were highly comparable with regard to age and, for the comparisons of adults, we chose samples that were roughly comparable with regard to age. In a study on young children, Schroots and van Alphen de Veer (1979) report correlation matrices for the Leidse Diagnostische Test for eight age groups between 4 and 8 years of age. The average correlation between the adjacent age groups is .75 (combined N=1169). Several studies report data on both younger and older children. The Dutch/Flemish WISC-R (van Haasen et al., 1986) has samples with comparable N of Dutch and Flemish children, so the 11 age groups between 6 and 16 could be compared. This resulted in an average correlation of .78 (combined N=3018). Jensen (1985) reports g loadings of the 12 subtests of the WISC-R obtained in three large independent representative samples of Black and White children. The average correlation between the g vectors obtained for each sample is .86 for the Black children (combined N=1238) and .93 for the White children (combined N=2868). In a study on older children, Evers and Lucassen (1991) report the correlation matrices of the Dutch DAT. The average correlation between the g vectors of three educational groups is .88 (combined N=3300). The US GATB manual (1970, chapter 20) gives correlation matrices for large groups of boys and girls in secondary school. The average correlation between the g vectors of the same-age boys and girls is .97 (combined N=26,708) Several studies report data on adults. g loadings of the eight subtests of the GATB are reported by te Nijenhuis and van der Flier (1997) for applicants at Dutch Railways and by de Wolff and Buiten (1963) for seamen at the Royal Dutch Navy, resulting in a correlation of .90 (combined N=1306). The US GATB manual (1970) gives correlation matrices for two large groups of adults, which yields a correlation between g vectors of .94 (combined N=4519). Johnson, Bouchard, Krueger, McGue, and Gottesman (2004) report g loadings for a sample that took the WAIS, and Wechsler (1955) reports the correlation matrices of the WAIS for adults of comparable age, so g loadings could be computed. The correlation between the g vectors for the two studies is .72 (combined N=736). So, it appears that g vectors are quite reliable, especially when the samples are very large.

The number of tests in the batteries in the present study varied from 7 to 14. The number of tests does not necessarily influence the size of r_{gd}, but clearly has an effect upon its variability. Because variability in the values of the artifacts influences the amount of variance artifacts explain in observed effect sizes, we estimated this variability using data from the samples described in the previous paragraph.

8.3. Correction for reliability of the vector of score gains

The value of r_{gd} is attenuated by the reliability of the vector of score gains for a given battery. When two samples have a comparable N, the average correlation between vectors is an estimate of the reliability of each vector. The reliability of the vector of score gains was estimated using the present datasets, comparing samples that took the same test and that differed little on background variables. For the comparisons using children, we choose samples that were highly comparable with regard to age and for the comparisons of adults we choose samples that were roughly comparable with regard to age.

In the GATB manual (1970, ch. 15), 13 combinations of two studies are described where large samples of men and women that are comparable with respect to age and background took the same GATB subtests. The average unweighted correlation between the *d* vectors of men and women is .83 (total $N=3760$). In the GATB manual (1970, ch. 20), three combinations of three studies are described where very large samples of boys and girls that are in the same grade in secondary school took the same GATB subtests. This yielded correlations between the *d* vectors of, respectively, .99, .98, and .94 (total $N=20{,}541$). Together, van Geffen (1972) and Bosch (1973) report three Dutch GATB test–retest studies on children in secondary school, resulting in three comparisons between *d* vectors. The average *N*-weighted correlation between the *d* vectors is .47 (total $N=127$). Vectors of score gains from two different datasets on the WISC-R were compared. Tuma and Appelbaum (1980) tested children with an average age of 10, and Wechsler (1974) tested 10- and 11-year-olds. The correlation between the two *d* vectors is .71 (total $N=147$). Comparison of vectors of score gains from datasets on the DAT (Bennett, Seashore, & Wesman, 1974) resulted in correlations of, respectively, .78 and .73, so an average *r* of .76 (total $N=254$). So, it appears that *d* vectors are quite reliable, especially when the samples are very large. We estimated the reliabilities of the *d* vectors in the database using data from the samples described in this paragraph.

8.4. Correction for restriction of range of g loadings

The value of r_{gd} is attenuated by the restriction of range of *g* loadings in many of the standard test batteries. The most highly *g*-loaded batteries tend to have the smallest range of variation in the subtests' *g* loadings. Jensen (1998a, pp. 381–382) shows that restriction in *g* loadedness strongly attenuates the correlation between *g* loadings and standardized group differences. Hunter and Schmidt (1990, pp. 47–49) state that the solution to range variation is to define a reference population and express all correlations in terms of that reference population. The Hunter and Schmidt meta-analytical program computes what the correlation in a given population would be if the standard deviation were the same as in the reference population. The standard deviations can be compared by dividing the study population standard deviation by the reference group population standard deviation, that is $u=\mathrm{S.D.}_{\mathrm{study}}/\mathrm{S.D.}_{\mathrm{ref}}$. As the reference we took the tests that are broadly regarded as exemplary for the measurement of the intelligence domain, namely the various versions of the Wechsler tests for children. The average standard deviation of *g* loadings of the various Dutch and US versions of the WISC-R and the WISC-III was 0.128. So, the S.D. of *g* loadings of all test batteries was compared to the average S.D. in *g* loadings in the Wechsler tests for children. This resulted in some batteries–such as the GATB–having a value of *u* larger than 1.00.

8.5. Correction for deviation from perfect construct validity

The deviation from perfect construct validity in *g* attenuates the value of r_{gd}. In making up any collection of cognitive tests, we do not have a perfectly representative sample of the entire universe of all possible cognitive tests. So any one limited sample of tests will not yield exactly the same *g* as any other limited sample. The sample values of *g* are affected by psychometric sampling error, but the fact that *g* is very substantially correlated across different test batteries implies that the differing obtained values of *g* can all be interpreted as estimates of a "true" *g*. The value of r_{gd} is attenuated by psychometric sampling error in each of the batteries from which a *g* factor has been extracted.

The more tests and the higher their *g* loadings, the higher the *g* saturation of the composite score. The Wechsler tests have a large number of subtests with quite high *g* loadings resulting in a highly *g*-saturated composite score. Jensen (1998a, pp. 90–91) states that the *g* score of the Wechsler tests correlate more than .95 with the tests' IQ score. However, shorter batteries with a substantial number of tests with lower *g* loadings will lead to a composite with a somewhat lower *g* saturation. Jensen (1998a, ch. 10) states that the average *g* loading of an IQ score as measured by various standard IQ tests is in the +.80 s. When we take this value as an indication of the degree to which an IQ score is a reflection of "true" *g*, we can estimate that a tests' *g* score correlates about .85 with "true" *g*. As *g* loadings are the correlations of tests with the *g* score, it is most likely that most empirical *g* loadings will underestimate "true" *g* loadings; so, empirical *g* loadings correlate about .85 with "true" *g* loadings. As the Schmidt and Le computer program only includes corrections for the first four artifacts the correction for deviation from perfect construct validity was carried out on the value of r_{gd} after correction for the first four artifacts. To limit the

Table 1
Dutch, British, and US studies of correlations between *g* loadings and gain scores

Reference	Test	*r*	*N*	Information
Drenth et al. (1968)	AKIT	−.57	100	Primary-school children
van Geffen (1972)	GATB	−.45	42	Secondary-school children
		−.21	42	
Bosch (1973)	GATB	−.07	43	Secondary-school children
Schroots and van Alphen de Veer (1979)	LDT	−.42	96	Pre-school and secondary-school children
Bleichrodt et al. (1987)	RAKIT	.09	49	Pre-school children
		−.25	51	Primary-school children
		−.21	49	Primary-school children
van der Doef et al. (1989)	WISC-R	−.69	22	Primary-school children with learning problems
Mulder et al. (2004)	KAIT	−.23	46	Secondary-school children + young adults
		−.42	25	Adults
Kort et al. (2005)	WISC-III	−.15	42	Primary-school children
		−.26	67	Primary-school children
		−.46	39	Secondary-school children
Luteijn and Barelds (2005)	GIT2	−.51	44	Adults
Kooij et al. (2005)	WAIS-III	−.63	60	Adults
Elliott (1983)	BAS	−.65	60	Primary-school children
Wechsler (1967)	WPPSI	−.46	50	Pre-school children
United States Department of Labor (1970)	GATB	−.35	156	Office applicants
		−.66	605	Male high school seniors
		−.70	554	Female high school seniors
		−.58	223	Males 1-day interval
		−.41	186	Females 1-day interval
		−.50	202	Males 2-week interval
		−.52	152	Females 2-week interval
		−.67	156	Males 6-week interval
		−.61	168	Females 6-week interval
		−.43	176	Males 13-week interval
		.02	149	Females 13-week interval
		−.62	157	Males 26-week interval
		−.32	136	Females 26-week interval
		−.69	119	Males 1-year interval
		−.31	183	Females 1-year interval
		−.96	118	Males 2-year interval
		−.75	170	Females 2-year interval
		−.75	123	Males 3-year interval
		−.48	183	Females 3-year interval
		−.92	3398	Boys secondary school
		−.92	3680	Girls secondary school
		−.91	3348	Boys secondary school
		−.91	3491	Girls secondary school
		−.84	3229	Boys secondary school
		−.87	3395	Girls secondary school
Wechsler (1974)	WISC-R	−.48	97	Primary-school children
		−.66	102	Primary-school children
		−.21	104	Secondary-school children
Bennett et al. (1974)	DAT	−.79	92	Boys secondary school
		−.53	81	Girls secondary school
		−.29	81	Boys secondary school
		−.62	100	Girls secondary school
Covin (1977)	WISC-R	−.57	30	Primary-school children with learning problems
Tuma and Appelbaum (1980)	WISC-R	−.08	45	Primary- and secondary-school children
Matarazzo et al. (1980)	WAIS	−.10	29	Young males
Wechsler (1981)	WAIS-R	−.64	71	Adults
		−.48	48	Adults

(continued on next page)

Table 1 (*continued*)

Reference	Test	*r*	*N*	Information
McCormick et al. (1983)	ASVAB	−.73	57	adults
Kaufman and Kaufman (1983)	K-ABC	−.27	46	Pre-school children
		−.18	36	Pre- and primary-school children
		−.22	70	Primary-school children
Wechsler (1997)	WAIS-III	−.45	100	Young adults
		−.57	102	Adults
		−.51	104	Adults
		.03	88	Adults
Reeve and Lam (2005)	EAS	−.34	123	Undergraduate students

In general, the *g* loadings were based on the correlation matrix taken from the manuals containing the test–retest studies or from the correlation matrix based on the largest sample size we could find. What follows is a list of the sources of the *g* loading, when not taken from the manuals containing the test–retest study.

van Geffen (1972) and Bosch (1973): de Wolff and Buiten (1963), see also Johnson, te Nijenhuis, and Bouchard (in press); Bleichrodt et al. (1987): te Nijenhuis et al. (2004), who used the same data on which the RAKIT manual is based; van der Doef, Kwint, and van der Koppel (1989): Dutch WISC-R manual; Elliott (1983): Table 9.8: Age 9:0–9:11 years; U.S. Dept. of Labor's GATB (1970): Jensen (1985, p. 214) using the largest correlation matrix in the GATB manual; Wechsler (1974), Covin (1977), and Tuma and Appelbaum (1980): Jensen (1985, p. 214, first study); Bennett et al. (1974): average of four highly similar correlation matrices; Matarazzo et al. (1980): Wechsler's (1955, p. 17) Table 8 for ages 25–34; McCormick et al. (1983): Ree and Carretta (1994); Reeve and Lam (2005) utilize SEM analyses and use item parcels instead of full scale scores to compute *g* loadings. The average *g* loading of all the item parcels for a specific subtest was taken as the *g* loading of that specific subtest.

risk of overcorrection, we conservatively chose the value of .90 for the correction.

9. Results

The results of the studies on the correlation between *g* loadings and gain scores are shown in Table 1. The table gives data derived from 64 studies, with participants numbering a total of 26,990. The table gives the reference for the study, the cognitive ability test used, the correlation between *g* loadings and gain scores, the sample size, and background information on the study. It is clear that virtually all correlations are negative and that the size of the few positive correlations is very small.

Table 2 shows the results of the psychometric meta-analysis of the 64 data points. It shows (from left to right): the number of correlation coefficients (*K*), total sample size (*N*), the mean observed correlations (*r*) and their standard deviation (S.D.*r*), the true correlations one can expect once artifactual error from unreliability in the *g* vector and the *d* vector and range restriction in the *g* vector has been removed (ρ) and their standard deviation (S.D.ρ). The next two columns present the percentage of variance explained by artifactual errors (% VE) and the 95% credibility interval (95% CI). This interval denotes the values one can expect for ρ in 19 out of 20 cases.

The large number of data points and the very large sample size indicate that we can have confidence in the outcomes of this meta-analysis. The estimated true correlation has a value of −.95 and 81% of the variance in the observed correlations is explained by artifactual errors. However, Hunter and Schmidt (1990) state that extreme outliers should be left out of the analyses, because they are most likely the result of errors in the data. They also argue that strong outliers artificially inflate the S.D. of effect sizes and thereby reduce the amount of variance that artifacts can explain. We chose to leave out three outliers–more than 4 S.D. below the average *r* and more than 8 S.D. below ρ–comprising 1% of the research participants. This resulted in no changes in the value of the true correlation, a large decrease in the S.D. of ρ with 74%, and a large increase in the amount of variance explained in the observed correlations by artifacts by 22%. So, when the three outliers are excluded, artifacts explain virtually all of the variance in the observed correlations. Finally, a correction for deviation from perfect construct validity in *g*

Table 2
Meta-analysis results for correlations between *g* loadings and gain scores after corrections for reliability and restriction of range

Studies included	*K*	*N*	*r*	S.D.*r*	ρ	S.D.ρ	%VE	95% CI
All	64	26,990	−.80	.20	−.95	.11	81	−0.74 to 1.16
All minus 3 outliers	61	26,704	−.81	.18	−.95	.03	99	−0.91 to 1.00

K=number of correlations, *N*=total sample size, *r*=mean observed correlation (sample size weighted), S.D.*r*=standard deviation of observed correlation, ρ=true correlation (observed correlation corrected for unreliability and range restriction), S.D.ρ=standard deviation of true correlation, %VE=percentage of variance accounted for by artifactual errors, 95% CI=95% credibility interval.

took place, using a conservative value of .90. This resulted in a value of −1.06 for the final estimated true correlation between *g* loadings and score gains. Applying several corrections in a meta-analysis may lead to correlations that are larger than 1.00 or −1.00, as is the case here. Percentages of variance accounted for by artifacts larger than 100% are also not uncommon in psychometric meta-analysis. They also do occur in other methods of statistical estimation (see Hunter & Schmidt, 1990, pp. 411–414 for a discussion).

10. Discussion

A large-scale meta-analysis of 64 test–retest studies shows that after corrections for several artifacts there is an estimated true correlation of −1.06 between *g* loading of tests and score gains and virtually all of the variance in observed correlations is attributable to these artifacts. As several artifacts explain virtually all the variance in the effect sizes, other dimensions on which the studies differ, such as age of the test takers, test–retest interval, test used, average-IQ samples, or samples with learning problems, play no role at all.

The estimated true correlation of −1.06 is the result of various corrections for artifacts that attenuate the correlations. The estimated values of the artifacts may underestimate or overestimate the population values of the artifacts. Therefore, estimates of true effect sizes may overestimate or underestimate the population values of the effect size. As a solution to this problem, Hunter and Schmidt (2004) suggest carrying out several meta-analyses on the same construct and taking the average estimated effect size of all meta-analyses. The general idea is that meta-analysis is a powerful research tool, but does not give perfect outcomes.

A correlation of −1.06 falls outside the range of acceptable values of a correlation, but one has to make a distinction between the meta-analytical estimate of the true correlation between *g* and *d*, and the true correlation between *g* and *d*. We interpret the value of −1.06 for the meta-analytical estimate as meaning that the true correlation between *g* and *d* is −1.00. A correlation of −1.00 means that there is an inverse relationship between *g* and score gains. So, the tests with the highest *g* loadings show the smallest gains. The most straightforward interpretation of this very large negative correlation is that there is no *g* saturation in test–retest gain scores.

11. The South African learning potential study

In a carefully carried-out study, Skuy et al. (2002) used a dynamic testing procedure to see whether it would improve the scores of Black South African students on Raven's Standard Progressive Matrices (RSPM). The Bantu Education Act of 1954 established a discriminatory educational system characterized by poorly qualified teachers, sparsely equipped and funded schools, and generally poor quality. Most Black students in the sample had not received the same quality of education as White students. Black, White, Indian, and Colored research participants took the RSPM on two occasions and, in between, randomly constituted experimental groups were exposed to the Mediated Learning Experience. Both the Black South African group and the group consisting of White, Indian, and Colored South Africans improved over their baseline on the RSPM, and the Black group showed greater improvement.

The value of these cognitive interventions increases when the score gains are transferred to other tests and to external criteria, such as school or work achievement. Therefore, the research participants also took Feuerstein's Representational Stencil Design Test as a transfer measure. The subject is presented with a stencil of a geometric design and then asked to point to which stencils need to be used and in what sequence in order to construct an identical design. Like the RSPM, the Stencils test also requires representational/abstract thinking, but the training on the RSPM showed little transfer to it. Moreover, the correlation of the RSPM scores with performance in the end-of-year psychology examination did not significantly improve after mediation. Once again, the score gains were empty; they did not generalize. Skuy et al. go on to ask the question what it is that was improved by their interventions. Professor Skuy made his data accessible to the present authors, so we could perform additional analyses.

12. Sample

The data from Skuy et al. (2002) were used, with the exception of data from three research participants because their pretest IQ scores were extremely low (more than 3 S.D.s below the group mean). Ninety-five university students in psychology aged 16 to 29 (mean age = 20, S.D. = 2.3; 25 males, 70 females) participated in this study. They were 66 Black students (20 males, 46 females) and 29 White (20), Indian (6), and Colored (3) students (5 males, 24 females). The mean age of the Black group was 20 (S.D. = 2.5) and of the White, Indian, and Colored group 19 years (S.D. = 1). Subjects were randomly assigned to the experimental group ($n = 55$) and to the control group ($n = 40$).

13. Procedure

The students participated in pre- and posttest phases, with a group intervention in between. The study focused on improvement in scores on the RSPM, using the Set Variations II of the Learning Propensity Assessment Device as the mediation task. Mediation training took 3 h and was conducted by three experienced psychologists with the assistance of six postgraduate psychology students. A detailed description is given in Skuy et al. (2002).

14. Measures and cognitive intervention

The Raven's Standard Progressive Matrices consists of 60 items (divided into 5 sets of 12 items) designed to measure the ability to form comparisons, to reason by analogy, and to organize spatial information into related wholes. It has been established as one of the purest measures of *g* (Jensen, 1998a). Skuy et al. (2002) found no evidence for test bias against Blacks in South African education. Rushton, Skuy, and Bons (2004) showed that the Raven's gave comparable predictive validities for students from various groups. Cross-cultural testing research has clearly shown that unsufficient proficiency in the language of the test can lead to biased assessments in tests with a strong verbal component. However, the Raven's is a nonverbal test.

The Learning Propensity Assessment Device consists of 14 exercises. Each exercise contains an initial mediation task. Subsequent tasks increase in complexity and novelty and aim to assist the learner to achieve mastery over the task. The purpose of mediation is to assist the learner to develop the appropriate cognitive strategies and functions needed for the successful completion of the task. The Set Variations II of the Learning Propensity Assessment Device consists of five sets of items, which comprise variations of Sets C, D, and E of the RSPM test. Each set of variations contains a learning task for the purpose of initial mediation followed by a series of progressively more difficult variations to which the skills learned must be applied. Mediation involves discussing with groups how to define the problem to be solved, focus on the task, set rules, regulate problem solving behavior, and identify the correct sequence of logical steps needed to solve the task. Mediation also involves helping the subject to develop appropriate concepts, verbal tools, and insights in relation to the task. A detailed description is given in Skuy et al. (2002).

15. Statistical analyses

Although the Skuy et al. study is among the South African learning potential studies with the largest sample size, the *N* is not large. We therefore chose basic statistical analyses.

15.1. Descriptive statistics

Means, standard deviations, and reliabilities were computed for the various groups. With regard to measures of effect size, Hunter and Schmidt (1990, p. 271) advise choosing estimates of variance with the least error. Because repeated test takings tend to change the size of the S.D. (Ackerman, 1987), we chose the S.D. of the pretests for the denominator. The correlation between scores before and after the training was computed to see whether the training had an effect on the rank order of individual's scores.

15.2. Correlation between score gains and g loadedness

Because our sample was not large and quite specific, estimates of *g* loadedness were taken from Lynn, Allik, and Irwing's (2004) item analysis of RSPM in Estonia using a large ($N=2735$), nationally representative sample. The same reasoning as in psychometric meta-analysis applies, namely that larger samples give better estimates of *g* loadings than smaller samples. In a hierarchical factor analysis of the items using structural equations modeling, Lynn et al. computed *g* loadings of 52 of the 60 items. In the present study, Pearson correlations were calculated between the *g* loadings of these 52 items and the effect sizes on these items.

15.3. g loadings

The RSPM consists of dichotomous items, so we computed a correlation matrix of polychoric correlations (Nunnally & Bernstein, 1994). A principal axis factor analysis was carried out. The percentage variance explained by the first unrotated factor was taken as an estimate of *g* loadedness. Because sample size was limited, we collapsed the experimental and the control group.

15.4. Correlation between sum scores and score gains

We tested whether individuals with low-*g* improved their scores more than those with high-*g* by correlating gain scores with pretest RSPM scores for each of the four research groups. As gain scores tend to be

Table 3
Proportion of sample selecting the correct answer on items of Raven's Standard Progressive Matrices by group

Set A			Set B			Set C			Set D			Set E		
Item	Black	Other[a]	Item	Black	Other	Item	Black	Other	Item	Black	Other	Item	Black	Other
1	1.00	1.00	13	1.00	1.00	25	1.00	.97	37	1.00	1.00	49	.74	.90
2	.97	1.00	14	1.00	1.00	26	.96	1.00	38	.99	1.00	50	.64	.90
3	.97	1.00	15	1.00	1.00	27	.96	1.00	39	.89	1.00	51	.79	.97
4	1.00	.97	16	.91	.97	28	.86	.93	40	.92	1.00	52	.56	.83
5	1.00	1.00	17	.96	.97	29	.94	.97	41	.96	1.00	53	.52	.83
6	.99	1.00	18	.85	1.00	30	.76	.83	42	.92	1.00	54	.35	.76
7	.94	.97	19	.77	.66	31	.88	.97	43	.77	1.00	55	.42	.79
8	.91	.93	20	.79	.97	32	.50	.79	44	.76	.93	56	.21	.69
9	1.00	.97	21	.83	.97	33	.74	.90	45	.71	.97	57	.30	.41
10	.91	.97	22	.92	1.00	34	.61	.79	46	.79	.93	58	.12	.41
11	.83	.90	23	.80	.90	35	.53	.69	47	.29	.41	59	.02	.17
12	.68	.83	24	.59	.83	36	.06	.35	48	.26	.38	60	.11	.21

[a] Other = White, Indian, and Colored.

negatively correlated with pretest scores as a function of unreliability (see Cronbach, 1990; Nunnally & Bernstein, 1994), we corrected the correlations using Tucker, Damarin, and Messick's (1966) formula 63. Using the formula, one adds to each correlation the term (S.D. pretest/S.D. gain score) * (1 − reliability pretest).

16. Results

16.1. Descriptive statistics

Internal consistencies (Cronbach α's) on the RSPM ranged from .76 to .86 for the pre- and posttests, respectively. Table 3 shows the proportion of each of the groups, which selected the correct answer on each of the 60 items of the pretest. Across the 60 items, the order of the p values was almost identical for Blacks and White/Indian/Coloreds (r = .92, p = .00).

Table 4 shows the means and standard deviations for the total RSPM scores for the four groups, along with the d effect sizes representing the difference between pre- and posttest scores (Cohen, 1988). First, we examined whether there was an effect of race (Black vs. White/Indian/Colored) and group (experimental vs. control) on the pretest scores. There was a significant effect due to race ($F(1, 91) = 24.13$, $p = .00$, $\eta^2 = .21$), but not group ($F(1, 91) = 2.28$, $p = .14$, $\eta^2 = .02$). This means that mean pretest scores of Blacks ($M = 44.44$, S.D. = 6.65) were lower than those of White/Indian/Coloreds ($M = 51.41$, S.D. = 5.05), and that mean pretest scores of experimental and control groups were comparable ($M = 45.53$, S.D. = 7.04 and $M = 48$, S.D. = 6.7, respectively).

Secondly, we investigated the effects of training on the posttest scores by performing a two-way ANCOVA on the total posttest scores with race and group as factors and the total pretest scores as the covariate. There was a significant effect for group ($F(1, 95) = 13.81$, $p = .00$, $\eta^2 = .13$) and for race ($F(1, 90) = 3.99$, $p = .05$, $\eta^2 = .04$), but not for the two-way interaction of group and race ($F(1, 90) = 0.28$, $p = .60$, $\eta^2 = .00$). These results indicate that the training was equally effective for both the Black and White/Indian/Colored students. Posttest scores of Blacks ($M = 49.41$, S.D. = 5.91), however, remained

Table 4
Pre- and posttest mean raven's scores, standard deviations, and mean effect sizes for Black and White/Indian/Colored students

	Black experimental (n = 40)		Black control (n = 26)		Other[a] experimental (n = 15)		Other control (n = 14)	
	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest	Pretest	Posttest
Raw scores								
M	43.78	50.10	45.46	48.35	50.20	55.80	52.71	55.36
S.D.	6.64	5.31	6.69	6.71	6.05	3.76	3.45	3.43
Percentile	14	41	16	31	41	75	55	68
Effect size	0.95		0.43		0.93		0.77	

Percentiles are based on U.S. adult norms; see Raven, Raven, and Court's (2000), Table SPM13.

[a] Other = White, Indian, and Colored.

significantly lower ($F(1, 91)=28.33$, $p=.00$) than those of Whites/Indians/Coloreds ($M=55.59$, S.D.$=3.55$). Although posttest scores of the experimental group ($M=51.65$, S.D.$=5.53$) were higher than those of the control group ($M=50.8$, S.D.$=6.65$), differences between both groups were nonsignificant ($F(1, 91)=0.85$, $p=.36$).

The correlation between scores before and after the training was .84 ($p=.00$) for the experimental group and .90 ($p=.00$) for the control group, showing that the training had only limited effect on the rank order of individual's scores. This means that the test strongly, but not perfectly measures the same constructs on both occasions.

16.2. Correlation between score gains and g loadedness

We estimated effect sizes for each of the four groups (race by condition) by computing the difference between mean pretest scores and posttest scores, divided by the standard deviation of the pretest scores of Black and White/Indian/Colored students, respectively. Finally, we calculated the correlations between effect sizes and the *g* loadings taken from Lynn et al. Correlations were −.24 ($p=.10$) for the Black experimental group, −.21 ($p=.20$) for the White/Indian/Colored experimental group, −.08 ($p=.59$) for the Black control group, and −.41 ($p=.01$) for the White/Indian/Colored control group. Small sample sizes usually attenuate correlations (Hunter & Schmidt, 1990). Collapsing the groups indeed resulted in higher average correlations: −.39 for the complete experimental group and −.26 for the complete control group.

16.3. g loadings

Using the combined experimental and control group, a principle axis factor analysis on the pretest and posttest scores, respectively, resulted in a first unrotated factor explaining 22% of the variance in the pretest scores and 18% of the variance in the posttest scores. These findings suggest that the *g* loadedness of the RSPM decreased substantially after Mediated Learning Experience.

16.4. Correlation between score gains and sum score

Correlating score gains with RSPM total scores resulted in values of −.60 ($p=.00$) for the Black experimental group, −.18 ($p=.38$) for the Black control group, −.82 ($p=.00$) for the White/Indian/Colored experimental group, and −.48 ($p=.08$) for the White/Indian/Colored control group. After the use of the correction formula of Tucker et al. (1966), these correlations became −.39, −.08, −.61, and −.35, respectively. Overall, these correlations show that low-*g* persons improved their scores more strongly than high-*g* persons.

17. Discussion

Skuy et al. (2002) hypothesized that the low-quality education of Blacks in South Africa would lead to an underestimate of their cognitive abilities by IQ tests. Groups of Black and White/Indian/Colored students took the Raven's Progressive Matrices twice, and in between received Feuerstein's Mediated Learning Experience. The test scores went up substantially in all groups. Evidence for an authentic change in the *g* factor requires broad transfer or generalizability across a wide variety of cognitive performance. However, Skuy et al. show that the gains did not generalize to scores on an other, highly similar test and to external criteria, and were therefore hollow. As the score gains were in some cases quite large–14 IQ points for the Black experimental group–the question becomes what is it that improved.

The findings show that the correlations between score gains and *g* loadedness of the items were −.39 for the complete experimental group and −.26 for the complete control group. However, because the *g* loadings and gain scores are measured at the item level their reliabilities are not high, resulting in substantial attenuation of the correlation between *g* and *d*. Moreover, RSPM does not measure *g* perfectly: Jensen (1998a, p. 91) estimates its *g* loading at .83. When we estimate the reliability of the *g* vector at .70 and the reliability of the gain score vector at .50, corrections for unreliability and deviation from perfect construct validity of *g* only would result in estimated true correlations of, respectively, −.80 and −.53. These values should be taken as underestimates; controlling for additional artifacts will bring them closer to the very strong negative correlation found in the meta-analysis.

The findings suggest that after training the *g* loadedness of the test decreased substantially. We found negative, substantial correlations between gain scores and RSPM total scores. Table 4 shows that the total score variance decreased after training, which is in line with low-*g* subjects increasing more than high-*g* subjects. Since, as a rule, high-*g* individuals profit the most from training–as is reflected in the ubiquitous positive correlation between IQ scores and training performance (Jensen, 1980; Schmidt & Hunter, 1998)–

these findings could be interpreted as an indication that Feuerstein's Mediated Learning Experience is not *g*-loaded, in contrast with regular trainings that are clearly *g*-loaded. Substantial, negative correlations between gain scores and RSPM total scores are no definite proof of this hypothesis, but are in line with it. Additional substantiation of our hypothesis that the Feuerstein training has no or little *g* loadedness is that Coyle (2006) showed that gain scores loaded virtually zero on the *g* factor. Moreover, Skuy et al. reported that the predictive validity of their measure did not increase when the second Raven score was used. The fact that individuals with low-*g* gained more than those with high-*g* could be interpreted as an indication that the Mediated Learning Experience was not *g*-loaded. It should be noted, however, that Feuerstein most likely did not intend his intervention to be *g*-loaded. He was interested in increasing the performance of low scorers on both tests and external criteria.

18. General discussion

IQ scores are by far the best general predictor of success in education, job training, and work. However, there are many ways in which these IQ scores can be increased, for instance by means of retesting or participating in a learning potential training program. What conclusions can be drawn from such score gains? Jensen's (1998a) hypothesis that the effects of training on abilities can be summarized in terms of Carroll's three-stratum hierarchical factor model was tested in a meta-analysis on test–retest data using Dutch, British, and American test batteries, and with learning potential data from South Africa using Raven's Progressive Matrices. The meta-analysis convincingly shows that test–retest score gains are not *g*-loaded. The findings from the learning potential study are clearly in line with this: when the attenuation caused by unreliability and other artifacts is taken into account the correlation between *g* loadings of items and gains on items has a value that is somewhat comparable to the one found in the meta-analysis for test batteries. The data suggest that the *g* loadedness of item scores decreases after the intervention training. Te Nijenhuis et al.'s (2001) finding that practice and coaching reduced the *g*-loadedness of their test scores strengthens the present findings using item scores. The findings show that not the high-*g* participants increase their scores the most–as is common in training situations–but it is the low-*g* persons showing the largest increases of their scores. This suggests that the intervention training is not *g*-loaded.

Our findings fit quite well with the hierarchical model of intelligence. The generalizability of test scores resides predominantly in the *g* component, whereas the test-specific ability component and the narrow ability component are virtually non-generalizable. This is, for instance, evidenced by the earlier finding that adding verbal tests to a *g* score or numerical tests to a *g* score resulted in only a very small incremental validity (Ree & Earles, 1991; Ree et al., 1994). Additionally, Ericsson and Lehmann (1996) reported immense gains for a memory task focusing on one narrow ability, but did not find any improvement for comparable memory tasks focusing on another narrow ability. As the score gains are not related to *g*, the generalizable *g* component decreases, and since it is not unlikely that the Feuerstein training itself is not *g*-loaded it is easy to understand why the score gains did not generalize to scores on the cognitively loaded Representational Stencil Design Test. For a similar reason, the score gains did not generalize to *g*-loaded external criteria, as the correlation of the RSPM scores with performance in the end-of-year psychology examination did not significantly improve after mediation. Reeve and Lam (2005) claimed that retesting does not change the nature of what is being tested, but our findings suggest the opposite.

19. Limitations of the studies

Our meta-analysis and our analysis of the South African study are strongly based on the method of correlated vectors (MCV), and recently it has been shown to have limitations. Dolan and Lubke (2001) have shown that when comparing groups substantial positive vector correlations can still be obtained even when groups differ not only on *g*, but also on factors uncorrelated with *g*. Ashton and Lee (2005) show that associations of a variable with non-*g* sources of variance can produce a vector correlation of zero even when the variable is strongly associated with *g*. They suggest that the *g* loadings of a subtest are sensitive to the nature of the other subtest in a battery, so that a specific sample of subtests may cause a spurious correlation between the vectors. Notwithstanding these limitations, studies using MCV continue to appear (see, for instance, Colom, Haier, & Jung, in press; Hartmann, Kruuse, & Nyborg, in press; Lee et al., 2006). The outcomes of our meta-analysis of a large number of studies using the method of correlated vectors may make an interesting contribution to the discussion on the limitations of the method of correlated vectors.

A principle of meta-analysis is that the amount of information contained in one individual study is quite modest. Therefore, one should carry out an analysis of all studies on one topic and correct for artifacts, leading to a strong increase of the amount of information. The fact that our meta-analytical value of $r=-1.06$ is virtually identical to the theoretically expected correlation between g and d of -1.00 holds some promise that a psychometric meta-analysis of studies using MCV is a powerful way of reducing some of the limitations of MCV. An alternative methodological approach is to limit oneself to the rare datasets enabling the use of structural equations modeling. However, from a meta-analytical point of view, these studies yield only a quite modest amount of information.

Additional meta-analyses of studies employing MCV are necessary to establish the validity of the combination of MCV and psychometric meta-analysis. Most likely, many would agree that a high positive meta-analytical correlation between measures of g and measures of another construct implies that g plays a major role, and that a meta-analytical correlation of -1.00 implies that g plays no role. However, it is not clear what value of the meta-analytical correlation to expect from MCV when g plays only a modest role. After the present meta-analysis on a construct that clearly has an inverse relationship with g, it would be informative to carry out meta-analyses of studies on variables that are strongly linked to g and variables that are modestly linked to g. An example of the latter would be secular score gains, which, according to Lynn's (1990) nutrition theory, should be modestly g-loaded.

The sample sizes in the South African study are not large, but still larger than those in many other studies of learning potential, where an $N\approx 10$ is not unusual. The results of a reanalysis of the many existing studies on dynamic testing could lead to a meta-analysis with a large combined N. The mean posttest score was quite high, so a ceiling effect may have taken place for the White/Indian/Colored group, leading to an underestimation of the experimental score gain for this group.

Instead of testing the hypothesis with a strongly unidimensional test such as the RSPM it would be better to use a multidimensional test. Moreover, a large sample size would allow the use of more rigorous data-analytical techniques leading to more definitive results. However, to the best of our knowledge, datasets meeting these requirements do not exist, and the Skuy et al. study is arguably the best South African learning potential study.

20. Score gains as low-quality measures of motivation?

As criterion-related validity is strongly dependent on g, te Nijenhuis et al.'s finding of lowered g loadings after training should result in lowered criterion-related validity. However, the empirical findings show the opposite: virtually all test–retest and test preparation studies on cognitive tests and scholastic aptitude tests that reported both criterion-related validities demonstrate small to modest increases in criterion-related validity for the second or third test score (see Allalouf & Ben-Shakhar, 1998; Bashi, 1976; Coyle, 2006; Hausknecht, Trevor, & Farr, 2002; Jones, 1986; Linn, 1977; Olsen & Schrader, 1959; Ortar, 1960; Powers, 1985; Reeve & Lam, 2005). In the carefully designed study by Allalouf and Ben-Shakhar (1998) of a university entrance test, the experimental group received an intensive 40-h test coaching program, while the control group did not. The criterion-related validity for the retest increased for both groups. Most importantly, the increase was the same—it was not larger for the experimental group.

In a little-known, but carefully designed, large-scale learning potential study by Resing (1990; see Table 4.23), she compared an experimental group that received a pretest, a learning potential training and a posttest against a control group that received only the pretest and the posttest. The mean criterion-related validity of the various second scores was .62 for both the experimental and the control group. Learning potential training did not result in incremental criterion-related validity over and above the validity resulting from simply retesting. The findings from both Resing and Allalouf and Ben-Shakhar suggest that cognitive interventions do not increase criterion-related validity more than simple retesting.

g and the personality measure conscientiousness have been shown to make an excellent combination of predictors (Schmidt & Hunter, 1998). Conscientiousness represents, among other characteristics, persistence, a will to achieve, and the ability to focus effort on the goal. A field study on test preparation using actual job applicants (Clause, Delbridge, Schmitt, Chan, & Jennings, 2001) showed that motivation to perform well on the test correlated .25 with test performance. One could speculate that score increases do not reflect a true cognitive component but rather become low-quality measures of motivation. Further, since the increase in validity due to retesting and learning potential training is modest in comparison to the large increase obtainable from the use of personality questionnaires personality testing might provide a less expensive and more accurate alternative.

21. Effectiveness of various training formats

Components of the mediation training used by Skuy et al. (2002) are similar to the test training used in te

Nijenhuis et al. (2001). Both the Dutch training and the South African training took 3 h, but whereas in the Dutch training the focus was on two different test formats, the South African training dealt only with one test format. The test training by Lloyd and Pidgeon (1961) took even less time, namely two half-hour segments, each focusing on one test format. The effect sizes in all studies were roughly comparable. This suggests that the methodologies employed by te Nijenhuis et al. and Lloyd and Pidgeon were more efficient than those used by Skuy et al. It is possible that the components of the mediation training that are not present in the other two training formats are not effective in raising test scores and could therefore be left out. If true, it might be possible to increase the scores on the RSPM by one S.D. with a relatively simple 1-h training.

22. Generalizability of findings

Can these findings of hollow score gains after test–retest, test practice, and Mediated Learning Experience Training be generalized to other studies where training-induced score gains were found? Ericsson and Lehmann (1996) reported tremendous score increases after intensive training on numeric memory tests, but these gains did not generalize in the least to verbal memory tests. Such gains on one narrow ability do not generalize to another narrow ability clustering under the same broad ability and are therefore hollow. Similarly, Jensen (1998b) showed that score gains due to adoption were not on the *g* factor and were, therefore, most likely hollow.

Rushton (1999) argued that intergenerational score gains are not linked to *g*, suggesting the Flynn effects may be empty, but he was strongly criticized by Flynn (1999, 2000). In studies on the Flynn effect, score gains found in cross-sectional studies are largest on the RSPM (Flynn, 1987). It has been suggested by Lynn (1998) that a substantial part of these intergenerational score gains on the RSPM are generalizable–they do reflect higher *g*–but the remaining part is hollow and should be interpreted as schooling effects. The RSPM does require the application of the mathematical principles of addition, subtraction, progression, and the distribution of values. In the three decades (1950s–1980s) over which these increases in RSPM scores have occurred, increasing proportions of 15- to 18-year-olds have remained in schools, where they have learned math skills that they have applied to the solution of matrices problems. Our findings could be interpreted as support for Lynn's hypothesis of the partial hollowness of score gains on the RSPM. Notwithstanding the high *g* loading of the sum score of the RSPM, it is quite sensitive to test–retest effects and training effects. Some studies on the Flynn effect (Lynn & Hampson, 1986; Teasdale & Owen, 1989) show that the increase in scores is largely concentrated in the lower segments of the IQ distribution. Our finding that low scorers show the largest gains after training may additionally support the notion that a part of the Flynn effect on the RSPM is hollow. Finally, Wicherts et al.'s (2004) findings show that in some of their datasets the secular score gains are most strongly linked to broad-, narrow-, and test-specific abilities, showing that an important part of the gains are non-generalizable.

Ceci (1991) showed that increased schooling leads to higher IQ scores, but are these gains highly specific or predominantly generalizable? It would be interesting to apply the techniques we used in this study to the findings from previous intervention studies. It may be that biological interventions (such as diet, vitamin supplements, vaccination against infectious disease) rather than psychological or educational interventions, are the most cost-effective method of producing true changes in *g* and broad abilities. It may be that there is a biological barrier between the first stratum and the second stratum that restricts the effects of behavioral interventions to narrow abilities and test specificities.

Acknowledgement

We like to thank Mervyn Skuy for his permission to use his dataset.

Thanks to Marié de Beer, Raegan Murphy, Welko Tomic, Art Jensen, and Frank Schmidt for feedback on previous versions of this paper.

Thanks to Arne Evers, Wilma Resing (Dutch Test Committee), and Andress Kooij (Harcourt) for also helping in locating test–retest studies.

References

Ackerman, P. L. (1986). Individual differences in information processing. An investigation of intellectual abilities. *Intelligence*, *10*, 101−139.

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing skills. *Psychological Bulletin*, *102*, 3−27.

Allalouf, A., & Ben-Shakhar, G. (1998). The effect of coaching on the predictive validity of scholastic aptitude tests. *Journal of Educational Measurement*, *35*(1), 31−47.

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

Bashi, Y. (1976). *Verbal and non-verbal abilities of 4th, 6th and 8th grade students in the Arab educational system in Israel.* Jerusalem: Hebrew University School of Education.

Bleichrodt, N., Resing, W. C. M., Drenth, P. J. D., & Zaal, J. N. (1987). *Intelligentie-meting bij kinderen. Empirische en methodologische verantwoording van de geReviseerde Amsterdamse Kinder Intelligentie Test [Measuring the intelligence of children. Empirical and methodological justification of the Revised Amsterdam Children Intelligence Test].* Lisse, the Netherlands: Swets.

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1974). *Differential Aptitude Tests (5th ed.): Manual.* New York: The Psychological Corporation.

Boeyens, J. C. A. (1989). *Learning potential: An empirical investigation.* Pretoria, South Africa: Human Science Research Council.

Bosch, F. (1973). *Inventarisatie, beschrijving en onderzoek m.b.t. de wijzigingen van de G.A.T.B.; incl. test-hertest onderzoek (No. Pz3b.Rp.0120) [Stock-taking, description, and research concerning the modifications of the GATB; includes test–retest study].* Utrecht, the Netherlands: Nederlandse Spoorwegen.

Carroll, J. B. (1993). Human cognitive abilities. *A survey of factor analysis studies.* Cambridge: University Press.

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components? A reassessment of the evidence. *Developmental Psychology*, *27*, 703−722.

Christian, K., Bachnan, H. J., & Morrison, F. J. (2001). Schooling and cognitive development. In R. J. Sternberg, & E. L. Grigorenko (Eds.), *Environmental effects on cognitive abilities* (pp. 287−335). Mahwah, NJ: Erlbaum.

Clause, C. S., Delbridge, K., Schmitt, N., Chan, D., & Jennings, D. (2001). Test preparation activities and employment test performance. *Human Performance*, *14*, 149−167.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale: Lawrence Erlbaum.

Colom, R., Jung, R. E., & Haier, R. J. (in press). Finding the *g*-factor in brain structure using the method of correlated vectors. *Intelligence*.

Covin, T. A. (1977). Stability of the WISC-R for 9-year-olds with learning difficulties. *Psychological Reports*, *40*, 1297−1298.

Coyle, T. R. (2006). Test–retest changes on scholastic aptitude tests are not related to *g*. *Intelligence*, *34*, 15−27.

Cronbach, L. J. (1990). *Essentials of psychological testing.* New York: HarperCollins.

de Villiers, A.B. (1999). *Disadvantaged students' academic performance: Analysing the zone proximal development.* Unpublished D. Phil. thesis, University of Cape Town, South Africa.

de Wolff, C. J., & Buiten, B. (1963). Een factoranalyse van vier testbatterijen [A factor analysis of four test batteries]. *Nederlands Tijdschrift Voor Psychologie*, *18*, 220−239.

Dolan, C. V., & Lubke, G. (2001). Viewing Spearman's hypothesis from the perspective of multigroup PCA: A comment on Schonemann's criticism. *Intelligence*, *29*, 231−245.

Drenth, P. J. D., Petrie, J. F., & Bleichrodt, N. (1968). *Handleiding bij de Amsterdamse Kinder Intelligentie Test [Manual of the Amsterdam Children Intelligence Test].* Amsterdam: Vrije Universiteit.

Elliott, C. D. (1983). *British Ability Scales. Manual 2: Technical Handbook.* Windsor, Great-Britain: NFER-Nelson.

Engelbrecht, M. (1999). *Leerpotensiaal as voorspeller van akademiese sukses van universiteitsstudente* [Learning potential as predictor of the academic success of university students]. Unpublished D. Phil. thesis, Potchefstroom University for Christian Higher Education, South Africa.

Ericsson, K. A., & Lehmann, A. C. (1996). Expert and exceptional performance. Evidence of maximal adaptation to task constraints. *Annual Review of Psychology*, *47*, 273−305.

Evers, A., & Lucassen, W. (1991). *Handleiding DAT '83: Differentiële Aanleg Testserie [Manual DAT'83: Differential Aptitude Test series].* Amsterdam: Swets.

Fleishman, E. A., & Hempel, W. E. (1955). The relation between abilities and improvement with practice in a visual discrimination reaction task. *Journal of Experimental Psychology*, *49*, 301−312.

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

Flynn, J. R. (1999). 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. (2000). IQ gains, WISC subtests and fluid *g*: *g* theory and the relevance of Spearman's hypothesis to race. In G. R. B. J. Goode (Ed.), *The nature of intelligence* (pp. 202−227). New York: Wiley.

Gaydon, V.P. (1988). *Predictors of performance of disadvantaged adolescents on the Soweto/Alexandra gifted child programme.* Unpublished M. Ed. dissertation, University of the Witwatersrand, South Africa.

Gottfredson, L. S. (1997). Why g matters. The complexity of everyday life. *Intelligence*, *24*(1), 79−132.

Gottfredson, L. S. (2002). *g*: Highly general and highly practical. In R. J. Sternberg, & E. L. Grigorenko (Eds.), *The general intelligence factor: How general is it?* (pp. 331−380). Mahwah, NJ: Erlbaum.

Grigorenko, E. L., & Sternberg, R. J. (1998). Dynamic testing. *Psychological Bulletin*, *124*, 75−111.

Haeck, W., Yeld, N., Conradie, J., Robertson, N., & Shall, A. (1997). A developmental approach to mathematics testing for university admissions and course placement. *Educational Studies in Mathematics*, *33*, 71−91.

Hartmann, P., Kruuse, N.H.S., & Nyborg, H. (in press). Testing the cross-racial generality of Spearman's hypothesis in two samples. *Intelligence.*

Hausknecht, J. P., Trevor, C. O., & Farr, J. L. (2002). Retaking ability tests in a selection setting: Implications for practice effects, training performance, and turnover. *Journal of Applied Psychology*, *87*(2), 243−254.

Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis.* London: Sage.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis* (2nd ed.). London: Sage.

Jensen, A. R. (1980). *Bias in mental testing.* London: Methuen.

Jensen, A. R. (1985). The nature of the black–white difference on various psychometric tests. Spearman's hypothesis. *Behavioral and Brain Sciences*, *8*, 193−263.

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

Jensen, A. R. (1998b). Adoption data and two *g*-related hypotheses. *Intelligence*, *25*, 1−6.

Johnson, W., Bouchard, T. J., Krueger, R. F., Jr., McGue, M., & Gottesman, I. I. (2004). Just one *g*: Consistent results from three test batteries. *Intelligence*, *32*, 95−107.

Johnson, W., te Nijenhuis, J., & Bouchard, T.J., Jr. (in press). Replication of the hierarchical visual-perceptual-image rotation model in de Wolff and Buiten's (1963) battery of 46 tests of mental ability. *Intelligence.*

Jones, R. J. (1986). A comparison of the predictive validity of the MCAT for coached and uncoached students. *Journal of Medical Education*, *61*, 335−338.

Kaufman, A. S., & Kaufman, N. L. (1983). K-ABC: Kaufman Assessment Battery for Children. *Interpretive manual.* Circle Pines, MN: AGS.

Kooij, A. P., Rolfhus, E., Wilkins, C., Yang, Z., & Zhu, J. (2005). *WAIS-III Nederlandstalige bewerking. Technisch rapport hernormering [WAIS-III adoptation in Dutch. Technical report renorming].* Amsterdam: Harcourt.

Kort, W., Schittekatte, M., Dekker, P. H., Verhaeghe, P., Compaan, E. L., Bosmans, M., & Vermeir, G. (2005). *WISC-IIINL: Wechsler Intelligence Scale for Children, Derde Editie NL. Handleiding en verantwoording [The Dutch WISC-III: Wechsler Intelligence Scale for Children, Third Edition for the Netherlands. Manual and justification].* Amsterdam: NIP.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin*, *95*, 179−188.

Kulik, J. A., Kulik, C. C., & Bangert, R. L. (1984). Effects of practice on aptitude and achievement test scores. *American Educational Research Journal*, *21*, 435−447.

Lee, K. H., Choi, Y. Y., Gray, J. R., Cho, S. H., Chae, J. -H., Lee, S., et al. (2006). Neural correlates of superior intelligence: Stronger recruitment of posterior parietal cortex. *NeuroImage*, *29*(2), 578−586.

Linn, R. L. (1977). On the treatment of multiple scores for Law School Admission Test repeaters (Report #LSAC-77-4). *In Law School Admission Council, Reports of LSAC Sponsored Research: Volume III, 1975-1977.* Princeton, NJ: Law School Admission Council.

Lipson, L.E. (1992). *Relationship of static and dynamic measures to scholastic achievement of black pupils*. Unpublished M.Ed. dissertation, University of Witwatersrand, South Africa.

Lloyd, F., & Pidgeon, D. A. (1961). An investigation into the effects of coaching on non-verbal test material with European, Indian and African children. *British Journal of Educational Psychology*, *31*, 145−151.

Luteijn, F., & Barelds, D. P. H. (2005). *GIT2: Groninger Intelligentie Test 2 [GIT2: Groningen Intelligence Test 2].* Amsterdam: Harcourt.

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

Lynn, R. (1998). In support of the nutrition theory. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 207−215). Washington, DC: American Psychological Association.

Lynn, R., Allik, J., & Irwing, P. (2004). Sex differences on three factors identified in Raven's Standard Progressive Matrices. *Intelligence*, *32*, 411−424.

Lynn, R., & Hampson, S. (1986). The rise of national intelligence: Evidence from Britain, Japan and the U.S.A. *Personality and Individual Differences*, *7*, 23−32.

Matarazzo, J. D., Carmody, T. P., & Jacobs, L. D. (1980). Test–retest reliability and stability of the WAIS: A literature review with implications for clinical practice. *Journal of Clinical Neuropsychology*, *2*(2), 89−105.

McCormick, B.K., Dunlap, W.P., Kennedy, R.S., & Jones, M.B. (1983). The effects of practice on the Armed Forces Vocational Aptitude Test Battery. US Army Research Institute for the Behavioral and Social Sciences, Technical Report 602.

Mulder, J. L., Dekker, R., & Dekker, P. H. (2004). *Kaufman Intelligentietest voor adolesecenten en volwassenen (KAIT): Handleiding [Kaufman Intelligence test for adolescents and adults (KAIT): Manual].* Leiden, the Netherlands: PITS.

Murphy, R. (2002). *A review of South African research in the field of dynamic assessment.* Unpublished MA dissertation. University of Pretoria. (available online from: http://upetd.up.ac.za/thesis/available/etd-05042002-161239/).

Nel, A. (1997). *Die voorspelling van akademiese sukses binne kontekst van 'n alternatiewe universiteitstoelatingsbeleid* [The prediction of academic success within the context of an alternative policy of university admission]. Unpublished M.A. dissertation, Rand Afrikaans University, South Africa.

Neubauer, A. C., & Freudenthaler, H. H. (1994). Reaction time in a sentence-picture verification test and intelligence: Individual strategies and effects of extended practice. *Intelligence*, *19*, 193−218.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Olsen, M., & Schrader, W. B. (1959). *The use of preliminary and final Scholastic Aptitude Test scores in predicting college grades (College Entrance Examination Board Research and Development Reports, and Statistical Reports #59-19.* Princeton, NJ: Educational Testing Service.

Ortar, G. R. (1960). Improving test validity by coaching. *Educational Research*, *2*, 137−142.

Powers, D. E. (1985). Effects of test preparation on the validity of Graduate Admission Test. *Applied Psychological Measurement*, *9*, 179−190.

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

Ree, M. J., & Carretta, T. R. (1994). The correlation of general cognitive ability and psychomotor tracking tests. *International Journal of Selection and Assessment*, *2*, 209−216.

Ree, M. J., & Earles, A. A. (1991). Predicting training success: Not much more than *g*. *Personnel Psychology*, *44*, 321−332.

Ree, M. J., Earles, J. A., & Teachout, M. S. (1994). Predicting job performance: Not much more than *g*. *Journal of Applied Psychology*, *79*, 518−524.

Reeve, C. L., & Lam, H. (2005). The psychometric paradox of practice effects due to retesting: Measurement invariance and stable ability estimates in the face of observed score changes. *Intelligence*, *33*, 535−549.

Resing, W. C. M. (1990). *Intelligentie en leerpotentieel: Een onderzoek naar het leerpotentieel van jonge leerlingen uit het basis-en speciaal onderwijs [Intelligence and learning potential: A study into the learning potential of young students in basic and special education].* Amsterdam, the Netherlands: Swets.

Rushton, J. P. (1999). Secular gains in IQ are 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., Skuy, M., & Bons, T. A. (2004). Construct validity of Raven's Advanced Progressive Matrices for African and non-African engineering students in South Africa. *International Journal of Selection and Assessment*, *12*(3), 220−229.

Schmidt, F. L., & Hunter, J. E. (1998). The validity and utility of selection methods in personnel psychology. Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, *124*(2), 262−274.

Schmidt, F. L., & Hunter, J. E. (1999). Theory testing and measurement error. *Intelligence*, *27*(3), 183−198.

Schmidt, F. L., & Le, H. (2004). Software for the Hunter-Schmidt meta-analysis methods. *University of Iowa, Department of Management and Organization, IOWA City, IQ. 42242*.

Schroots, J. J. F., & van Alphen de Veer, R. J. (1979). *LDT: Leidse Diagnostische Test. Deel 1: Handleiding [LDT: Leiden Diagnostic Test. Part 1: Manual].* Lisse, the Netherlands: Swets.

Shochet, I. M. (1986). *Manifest and potential performance in advantaged and disadvantaged students.* Unpublished D.Phil. dissertation, University of the Witwatersrand, South Africa.

Skuy, M., Gewer, A., Osrin, Y., Khunou, D., Fridjon, P., & Rushton, J. P. (2002). Effects of mediated learning experience on Raven's Matrices scores of African and non-African university students in South Africa. *Intelligence*, *30*, 221−232.

Swanson, H. E., & Lussier, C. M. (2001). A selective synthesis of the experimental literature on dynamic assessment. *Review of Educational Research*, *71*, 321−363.

Teasdale, T. W., & Owen, D. R. (1989). Continuing secular increase in intelligence and a stable prevalence of high intelligence levels. *Intelligence*, *13*, 255−262.

Tuma, J. M., & Appelbaum, A. S. (1980). Reliability and practice effects of WISC-R IQ estimates in a normal population. *Educational and Psychological Measurement*, *40*, 671−678.

te Nijenhuis, J., Tolboom, E., Resing, W., & Bleichrodt, N. (2004). Does cultural background influence the intellectual performance of children from immigrant groups?: Validity of the RAKIT intelligence test for immigrant children. *European Journal of Psychological Assessment*, *20*, 10−26.

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

te Nijenhuis, J., Voskuijl, O. F., & Schijve, N. B. (2001). Practice and coaching on IQ tests: Quite a lot of *g*. *International Journal of Selection and Assessment*, *9*, 302−308.

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

Tucker, L. R., Damarin, F., & Messick, S. (1966). A base-free measure of change. *Psychometrika*, *31*(4), 457−473.

van der Doef, M. P., Kwint, J. M., & van der Koppel (1989). Wat leren moeilijk lerende kinderen van de WISC-R? [What do children who have difficulties in learning learn from the WISC-R?] *Kind en Adolescent*, *10*, 136−141.

United States Department of Labor. (1970). *Manual for the USTES General Aptitude Test Battery. Section III. Development.* Washington, DC: United States Department of Labor.

van Geffen (1972). *De betrouwbaarheid van de GATB 1002-B op brugklasniveau [The reliability of the GATB 1002 B for the first class at secondary school].* Catholic University Nijmegen, the Netherlands: Psychology of Work and Organisation.

van Haasen, P. P., de Bruyn, E. E. J., Pijl, Y. J., Poortinga, Y. H., Lutje Spelberg, H. C., Vander Steene, G., et al. (1986). *WISC-R: Wechsler Intelligence Scale for Children-Revised; Nederlandstalige uitgave [WISC-R: Wechsler Intelligence Scale for Children-Revised; Dutch edition].* Lisse, the Netherlands: Swets.

Wechsler, D. (1955). *Manual for the Wechsler Adult Intelligence Scale.* New York: The Psychological Corporation.

Wechsler, D. (1967). *Manual for the Wechsler Preschool and Primary Scale of Intelligence.* New York: The Psychological Corporation.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children-Revised.* New York: The Psychological Corporation.

Wechsler, D. (1981). *WAIS-R manual: Wechsler Adult Intelligence Scale-Revised.* New York: The Psychological Corporation.

Wechsler, D. (1997). *WAIS-III: Wechsler Adult Intelligence Scale-third edition and WMS-III: Wechsler Memory Scale-third edition, Technical manual.* New York: The Psychological Corporation.

Wicherts, J. W., Dolan, C. V., Oosterveld, P., van Baal, G. C. V., Boomsma, D. I., & Span, M. M. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*(5), 509−537.

Yeld, N., & Haeck, W. (1997). Educational histories and academic potential: Can tests deliver? *Assessment and Evaluation in Higher Education*, *22*, 5−16.

Zaaiman, H. (1998). *Selecting students for Mathematics and Science: The challenge facing higher education in South Africa.* South Africa, Pretoria: HSRC Publishers.

Zaaiman, H., van der Flier, H., & Thijs, G. D. (2001). Dynamic testing in selection for an educational programme: Assessing South African performance on the Raven Progressive Matrices. *International Journal of Selection and Assessment*, *9*, 258−269.

Zolezzi, S. A. (1992). *Alternative selection measures for university undergraduate admissions.* Unpublished M.Ed dissertation, University of the Witwatersrand, South Africa.

Zolezzi, S. A. (1995). *The effectiveness of dynamic assessment as an alternative aptitude testing strategy.* Unpublished D.Phil. dissertation, University of South Africa, South Africa.