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