Intelligence 41 (2013) 791–801

Contents lists available at ScienceDirect

Intelligence

Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypothesis perhaps

Jakob Pietschnig [a,b,*,1], Ulrich S. Tran [b,1], Martin Voracek [b,c,1]

[a] School of Health & Social Sciences, Middlesex University Dubai, United Arab Emirates
[b] Department of Basic Psychological Research and Research Methods, Faculty of Psychology, University of Vienna, Austria
[c] Georg Elias Müller Department of Psychology, Georg August University of Göttingen, Germany

ARTICLE INFO

Article history:
Received 28 January 2013
Received in revised form 4 June 2013
Accepted 5 June 2013
Available online 4 July 2013

Keywords:
Flynn effect
Measurement invariance
Guessing behavior
Decreasing IQ variability
Item response theory

ABSTRACT

Potential explanations for generational intelligence test score gains continue to be subject to intense debate and scrutiny in the scientific community. However, the explanatory value of some of the proposed causes remains difficult to determine, since only little empirical evidence is available. To clarify the role of two scarcely investigated theories accounting for the Flynn effect, this study set out to examine the role of changing test-taking behavior (Brand's hypothesis) and of a narrowing of the IQ ability distribution (Rodgers' hypothesis). Archival records of crystallized intelligence test performance over a time-span of 17 years of a large number of psychiatric inpatients and outpatients in Austria were investigated ($N = 5445$; 1978–94). This sample was particularly suitable to investigate our hypotheses since participants were under no pressure to perform which makes observed changes in test taking behavior attributable to personal style and ability rather than differential performance in pressure situations. Analytical approaches of both classical test theory and item response theory (IRT) yielded gains of 1.0 to 2.4 IQ points per decade. Test-taking behavior indicative of guessing and decreasing population IQ variability appeared to contribute both to IQ test score gains. IRT-based analyses showed that gains were largely preserved when controlling for highest educational qualification, while the test instrument showed measurement invariance between cohorts. However, IRT-based results also suggested that changes in test-taking behavior might not necessarily reflect increased guessing, but item drift instead. In all, this evidence emphasizes better performance of individuals of the lower tail of the IQ ability distribution in more recent years as one important contributing factor for generational IQ test score gains.

1. Introduction

Following the seminal studies of James Flynn (1984, 1987) and Richard Lynn (1982), many researchers became interested in the investigation of generational intelligence test score gains in the general population, a phenomenon that has since become known as the Flynn effect. Evidence for such gains had already been published in the first half of the twentieth century (e.g., Merrill, 1938; Tuddenham, 1948; for an overview, see Lynn, 2013), but performance differences were mainly attributed to differences in sampling between cohorts. Possibly, the importance of these gains was not recognized at this time because there were no theories accounting for what might have caused such gains. In contrast, one theory actually predicted that intelligence test performance would decrease over time (Cattell, 1937).

As in more recent years the topic became subject to intense scrutiny, numerous hypotheses aiming to explain the

* Corresponding author at: School of Health & Social Sciences, Middlesex University Dubai, Block 16, Knowledge Village, P.O. Box 500697, United Arab Emirates.
E-mail address: j.pietschnig@mdx.ac (J. Pietschnig).
[1] All three authors contributed equally; author names are listed alphabetically.

0160-2896/$ – see front matter
http://dx.doi.org/10.1016/j.intell.2013.06.005

Flynn effect have been offered. In general, better education (Husén & Tuijnman, 1991; Teasdale & Owen, 2005), improved nutrition (Lynn, 1989, 2009b), reduced pathogen stress (Eppig, Fincher, & Thornhill, 2010), and social multiplier effects (Dickens & Flynn, 2001) appear to have been the most intensely discussed theories. Recently, hybrid vigor (Mingroni, 2004, 2007) has been shown to be theoretically sound, but practically yields too small effects to wholly explain the IQ gains (Woodley, 2011). In one recent experimental investigation, the beneficial effects of more frequent use of advanced technology (Neisser, 1997) could not be evidenced (Sigal & McKelvie, 2012). Still other theories comprise influences of a more demanding everyday environment (Schooler, 1998), decreasing family size (Zajonc & Mullally, 1997), slower life-history (Woodley, 2012), or less frequently cited speculations such as genomic imprinting effects due to visual stimulation (Storfer, 1999) and even effects of the collective subconscious (Mahlberg, 1997).

Intelligence test score gains have been shown to be noticeably differentiated across countries (Voracek, 2006). The strongest gains have been found for France, Israel, Japan, Kenya, the Netherlands, and Spain, while gains in nations like Australia, Brazil, Great Britain, Ireland, and New Zealand have been moderate (Colom, Flores-Mendoza, & Abad, 2007; Colom, Lluis-Font, & Andres-Pueyo, 2005; Daley, Whaley, Sigman, Espinosa, & Neumann, 2003; Flynn, 2009). Of note, in some Scandinavian countries with available data (Norway and Sweden), gains appear to be stagnating (Sundet, Barlaug, & Torjussen, 2004), and average intelligence test performance has even been observed to be decreasing in Denmark during more recent years (Teasdale & Owen, 2005).

Another important variable moderating the Flynn effect is the measured intelligence domain itself. Gains have been observed to be typically considerably higher on test measures of fluid intelligence than of crystallized intelligence (e.g., Flynn, 2009; Pietschnig, Voracek, & Formann, 2011; Voracek, 2006). While in Anglo-American countries there appear to be virtually no gains on measures of crystallized intelligence (Lynn, 2009a), evidence from German-speaking countries intriguingly shows gains on measures of crystallized intelligence to be of similar magnitude as typically observed for fluid intelligence (Pietschnig, Voracek, & Formann, 2010; Voracek, 2006).

This pattern of differential gains could in principle be due to differential causal factors for the Flynn effect operating on different intelligence domains. It has also been shown that cognitive ability gains can even be observed among infants on developmental tests, thus rendering educational factors unlikely to fully account for the Flynn effect (Lynn, 2009b; Thompson, 2012). However, in general there exists a consensus that performance on crystallized test measures is more strongly associated with schooling than performance on tasks assessing fluid intelligence (e.g., McArdle, Ferrer-Caja, Hamagami, & Woodcock, 2002). Therefore, it is conceivable that gains on the crystallized intelligence domain may be more strongly associated with improved education than corresponding gains on fluid intelligence.

Two additional hypotheses that have rarely been investigated until now (and therefore were more closely investigated in the present study) attribute IQ gains to changes in test-taking behavior (Brand, 1987a, 1987b, 1990; Brand, Freshwater, & Dockrell, 1989) and to a narrowing of the ability distribution (Rodgers, 1999; Rowe & Rodgers, 2002), respectively. The former account (henceforth, the Brand hypothesis) proposes that Western societies, where most of the accounts of the Flynn effect stem from, became more permissive over time during the 20th century, which in turn may have caused individuals to take chances more often and to guess the answer for an item rather than skipping it. Multiple-choice response formats of items make guessing easy and time limits of some cognitive test measures (e.g., some administrations of the Raven matrices tests) would effectuate that quick indiscriminate guessing leads to the strongest effects and therefore to the largest IQ gains. Thus, the resulting improved scores would then merely reflect a personality facet (i.e., risk-taking behavior), rather than an expression of improved cognitive performance.

Flynn (1990) criticized Brand's hypothesis as untenable, since substantial IQ test score gains of Scottish students on the verbal subscales of the Wechsler intelligence test batteries could be observed. Because these subscales comprise open-format answers and are administered without time constraints, Flynn argued that a main prediction of Brand's assumption was unfulfilled. However, the value of Brand's account to contribute to explanations of the Flynn effect remains largely unaddressed, because of Flynn's reliance on increases of item pass-rates of the unaltered items in WISC and WISC-R for the calculations of IQ gains, rather than on empirically observed samples (Brand, 1990).

The hypothesis of Rodgers suggests that improved test performance reflects decreasing population variance and a resulting narrowing of the ability distribution with respect to cognitive abilities. This would mean that higher scoring of individuals within the lower tail of the ability distribution in turn leads to a shift of the lower tail upwards to the mean and overall to higher observed mean test scores.

The value of IRT-based (item response theory) assessments of IQ test score changes has been previously demonstrated (Beaujean, 2006; Beaujean & Osterlind, 2008; Beaujean & Sheng, 2010), although applications of such methods generally have remained rare instances in the respective research, because item-level data from appropriate subject pools frequently are unavailable for researchers. While analysis of sum scores is more straightforward and interpretation may be intuitively easier, IRT-based examination of data on the item level presents several advantages compared to analyses of sum scores, as has been previously pointed out (e.g., Beaujean & Sheng, 2010). IRT methods make it feasible to directly and unequivocally examine the unidimensionality of the test measure (i.e., ensuring all items reflect performance related to the same latent ability) and to equate group scores on the same scale.

Moreover, it has been proposed that intelligence test score gains may be due to changes in the constructs underlying the psychometric instruments scrutinized (Beaujean & Sheng, 2010; Wicherts et al., 2004). In this case, differences in test scores between different cohorts would not reflect true changes in the latent ability (i.e., the test would not be measurement invariant), but instead would rather be due to other systematic differences between the examined cohorts. While measurement invariance has been found to be tenable in at least one study (Beaujean & Sheng, 2010), the assumption of measurement invariance could not be retained in two other accounts