Intelligence xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Intelligence

journal homepage:

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 xxxx

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 © 2013 Elsevier Inc. All rights reserved.
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

Please cite this article as: Pietschnig, J., et al., Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypoth..., *Intelligence* (2013), http://dx.doi.org/10.1016/j.intell.2013.06.005

(Beaujean & Osterlind, 2008; Wicherts et al., 2004). One major advantage of the application of IRT-based methods, such as the two-parameter normal ogive model, is that this allows for direct examination of measurement invariance.

In the present study, we investigated intelligence test score changes on a well-established measure of crystallized intelligence in a large sample of psychiatric inpatients and outpatients over a time-span of 17 years. This sample was particularly suitable to investigate changes in test taking behavior since participants were instructed not to guess if unsure about the correct solution and there were no foreseeable consequences of performance for participants. Accordingly, observed changes most likely reflect genuine changes in ability or personal answering style. We aimed to provide a detailed examination of two proposed causes for the Flynn effect, namely changed test-taking behavior (i.e., more guessing; Brand, 1987a, 1987b, 1990; Brand et al., 1989) and a narrowing of the ability distribution (Rodgers, 1999; Rowe & Rodgers, 2002), by means of standard approaches of classical test theory. Additionally, IRT-based analyses allowed the examination of measurement invariance in the present sample, thus making it possible to assess whether potentially present IQ test score gains are due to genuine changes of the measured latent ability or whether factors different from latent abilities are more likely to be responsible. Moreover, in the IRT-based analyses, we also controlled for differences in the educational level of test-takers across time, in order to assess whether improved education may in fact be responsible for any observed performance gains.

2. Methods

2.1. Participants

Item-level test protocols of the psychometric assessments of a total of 5815 inpatients and outpatients of the University Clinic of Psychiatry in Vienna, Austria, who (i) between 1978 and 1994 were referred to the clinical psychological assessment unit within the university clinic and (ii) there completed at least one form of a crystallized intelligence measure, were obtained from archival records and compiled by author M.V. These item-level test protocols were then linked with basic demographic information of the patients, but never to psychiatric diagnoses or other clinical information. Institutional review board approval by the university clinic for the study was granted to author M.V. This dataset has first been analyzed, foremost within the framework of classical test theory, by Voracek (2002). Initial results of this, concerning the Flynn effect, have briefly been summarized by Voracek (2006), and the evidence subsequently has also been included in one meta-analysis of the Flynn effect on vocabulary tests (Pietschnig et al., 2010). In the present analysis, only data of patients who at least solved one item of a measure of crystallized intelligence (the Multiple-Choice Vocabulary Test, MWT-B; Lehrl, 1977) correctly were included, yielding data of 5445 patients (2390 women; sample mean age = 35.1 years). For IRT-based analyses, participants were categorized in one of three educational levels according to completion of lower secondary education or less (low level), vocational education and training (medium level), and upper secondary education or university degree.

2.2. Materials

The Multiple-Choice Vocabulary Test (MWT-B; Lehrl, 1977) is a well-established and widely used measure of crystallized intelligence in German language. It is a rather simple task, requiring individuals in each of the 37 items to single out an existing word placed among four made-up distractor words, i.e., neologisms. This comparatively simple item structure makes it suitable for administration to individuals with impaired cognitive functions, as it has been shown to provide a good estimate of premorbid IQ (Lehrl, 1977, p. 23). Particularly the short administration time (given as 3 to 5 min in the test manual) has made this test measure attractive in clinical contexts.

The MWT-B has been demonstrated to be reliable and valid (Lehrl, Triebig, & Fischer, 1995). An important detail for the present study is the fact that the MWT-B had not been restandardized over the whole time-span of our investigation, thus allowing examination of potential changes on test performance of individuals over time.

2.3. Procedure

The source and sampling contexts of the data analyzed here are described above (Section 2.1). Over the investigated time-span of 17 years, the catchment area, structure, and supply area of the University Clinic of Psychiatry were fairly the same, so it can be assumed that the composition of patients referred within the clinic to its clinical psychological assessment unit did not change substantially, although there is no direct empirical evidence for this assumption available. As the test was administered routinely among those who were referred to the clinical psychological assessment unit within the psychiatric clinic and the test had no obvious implications for patients (just used as a quick evaluation of patients' vocabulary-based, and thus premorbid, IQ), it can be assumed that patients' response behavior should have been largely unaffected from test anxiety. Importantly, participants were explicitly instructed (both verbally and in written form) not to guess, but instead to leave items unattempted when they were not really sure about the correct answer. Consequently, such a pressure-free test setting is particularly suitable to investigate guessing behavior, since in situations without pressure to perform, guessing behavior can be seen as the genuine readiness to guess on a test without influences of potential gains or losses. Testing took place individually, invariably right at the beginning of the psychological assessment session the patients were referred to, and the MWT-B was administered without time limit by experienced staff of the clinical psychological assessment unit.

We used both the straightforward approaches of classical test theory (i.e., analyzing raw scores) and the modeling of person and item parameters within the framework of item response theory (IRT). We first provide the results from the classical test–theoretical analyses and then report the findings from the IRT-based approach.

Please cite this article as: Pietschnig, J., et al., Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypoth..., *Intelligence* (2013), http://dx.doi.org/10.1016/j.intell.2013.06.005

3. Classical test–theoretical approach

3.1. Data analysis

3.1.1. Test score changes

Initially, we aimed to clarify whether there is evidence for test score changes over time in the data. This was achieved by two straightforward, well-established approaches: First, following an approach by Flynn (1998), weighted mean IQs of the initial three years of data collection were subtracted from the weighted mean IQ of the last three years of data collection. Differences were subsequently transformed to test for score changes in units of IQ per decade (Jensen, 1998). Second, year of data collection was regressed on individual test scores and on mean test score by year (see Pietschnig et al., 2010). Thus, the resulting slope of the regression equation can be interpreted as change in IQ points per year.

3.1.2. Test-taking behavior

In order to investigate changes of participants' test-taking behavior over time, we regressed test year on indicators of guessing behavior (the number of items that had been omitted in each test protocol, and the number of test protocols where all items had been attempted). Accordingly, averaged variables per year were treated as the criterion (mean n per year = 342; min = 201; max = 526) and year of assessment as the predictor.

In a second step of analysis, the last (and most difficult) six items of the test were examined separately, because the initial 31 items appeared to be too easy for most participants (mean correct responses per participant on these 31 initial items = 92%) and because guessing behavior should be reflected most markedly on the most difficult items (mean correct responses per participant on the remainder of 6 items = 36%).

Finally, measures of reliability were examined. Test year was regressed on interitem correlations and on Cronbach α per year.

3.1.3. Variability in task performance

Variability of test results over time was examined. For this purpose, test scores of participants were transformed to IQ scores and, based on the observed IQ distribution within each year, percentiles were determined. This approach serves to give a straightforward illustration of the trajectory of performance variability.

Subsequently, we examined changes in the shape of the performance distribution. Test year was regressed on skewness and on kurtosis per year.

3.2. Results

3.2.1. Flynn effect

Considerable intelligence test performance gains were observed for the weighted calculation, as introduced by Flynn (1998), as well as for the regressional approaches, although they somewhat varied in magnitude. The smallest gains were observed when test year was regressed on individual test results (1.69 IQ points per decade), middling gains when using Flynn's approach (1.98 IQ points per decade), and the largest gains when test year was regressed on mean test results per year (2.40 IQ points per decade). Clearly, a Flynn effect is observable in these data regardless of the method of assessment. However, here we do not focus on the absolute magnitude of the gains, since the goal of this study was the evaluation of the explanatory potential of variables that have been proposed as meaningful causes of the Flynn effect.

3.2.2. Test-taking behavior

Single regression models of test year on the mean and median number of omitted items in each test protocol were significant, showing a marked decrease of omitted items over time (bs = −0.09 and −0.27 respectively, ps < .001; Fig. 1). Another regression of test year on the percentage of test protocols wherein all items had been attempted to be solved yielded a significant positive slope (b = 0.01 for all items, b = 0.03 for Items 32 to 37, ps < .001; Fig. 2A). Results of both these regressional analyses could be interpreted as an increase in guessing behavior of test-takers. However, a decrease of omitted items is also a necessary side condition for that test performance gains may occur at all in this case of a test, of which most items are rather easy (see above). Thus, if test performance really increased over time, the number of omitted responses necessarily needed to decrease as well. Hence, the results on guessing behavior appear less conclusive than desired.

We next examined linear regressions of test year on the percentage of correct answers given for the six most difficult test items separately. With the exception of results for one item (Item 32: b = −0.003, p < .05), all regressions yielded positive slopes, indicating an increase of correct answers, although analyses for the two most difficult items failed to reach significance (bs = 0.001 to 0.012; Fig. 2B).

In a last step, we examined indicators of reliability over the investigated period. Single regressions yielded significant negative slopes for both interitem correlations (b = −0.003, p = .04) and Cronbach α (b = −0.003, p = .005; Fig. 3). Notwithstanding the significant decrease, reliability was satisfactory for all years (all Cronbach α > .87). These decreasing reliability indices over time can be interpreted as an expression of either more uniform answering patterns (due to increasing numbers of items answered correctly by increasing numbers of individuals) or more inconsistent and erratic answering patterns (due to generally increased guessing behavior of test-takers).

3.2.3. Variability in task performance

Changes in the variance of individuals' task performance over time were assessed. A regression of study year on standard deviations of the test scores yielded a negative slope, indicating decreases of test score variance over time, but test year failed to reach significance (b = −0.03, p > .05; Fig. 4A). Visual inspection of the trajectory of percentile ranks (Fig. 4B), however, appears to indicate a narrowing of the percentile ranks over time.

Examination of the yearly performance distribution showed increasing positive skewness (b = 0.013, p = .07; Fig. 5A). Although the regression coefficient was nonsignificant, there was a trend of a shifting of the lower end of the performance distribution towards the mean. Similarly, a regression of test year on kurtosis showed a nonsignificant yearly decrease in the excess of the kurtosis of the performance distribution (b = −0.010, p = .22; Fig. 5B). Although nonsignificant, the

Please cite this article as: Pietschnig, J., et al., Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypoth..., *Intelligence* (2013), http://dx.doi.org/10.1016/j.intell.2013.06.005

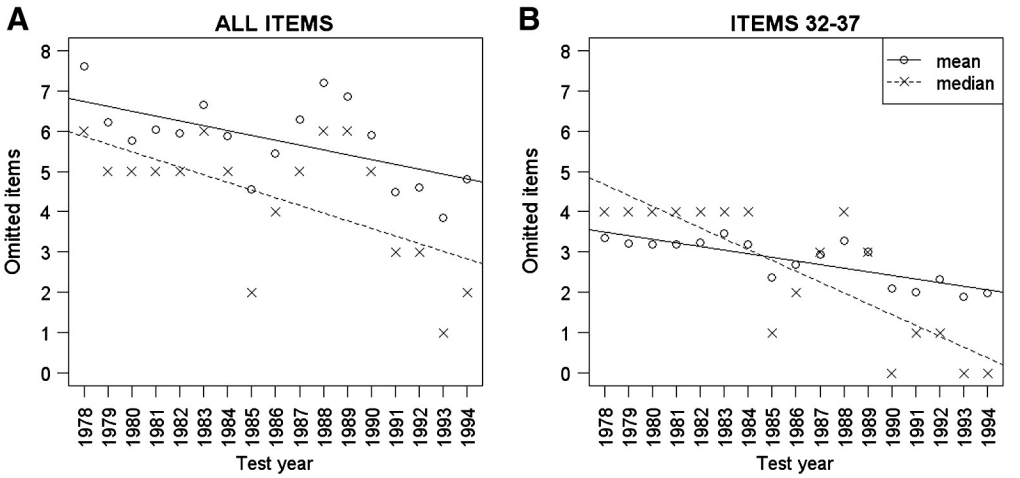

Fig. 1. Average number of unattempted items in (A) all items and (B) the most difficult items.

negative sign of the regression coefficient indicates a shift towards more platykurtic distributions over time.

Importantly, these changes of the distribution were not due to a ceiling effect. Of the total 5445 participants, only 81 managed to solve 36 and only 18 managed to solve all 37 items (i.e., about 1 in 70 and 1 in 300 participants respectively). Moreover, there was no indication for an accumulation of such exceptionally high test results towards the end of the data collection period (about 6 participants per year; minimum 2 and maximum 14 participants in 1988 and 1985 respectively).

3.3. Discussion

These classical test–theoretical findings show an increase of about 1.7 to 2.4 IQ points per decade on a widely-used measure of crystallized intelligence from 1978 to 1994 in a large patient sample. These findings are consistent with evidence that has been reported previously for German-speaking countries (Pietschnig et al., 2010, 2011), although the gains in the current sample appear to be somewhat smaller. Of note, the IQ gains can even be observed on the level of items, as the percentage of correct solutions of the six most difficult items over time generally increased.

Results from regression analyses on different potential indicators of guessing may be interpreted as support for the contention that has been put forward that the Flynn effect may be due to increased risk-taking behavior on behalf of the test-takers (Brand, 1987a, 1987b, 1990; Brand et al., 1989). In later test years, test-takers in this sample skipped less items, more frequently left no item unattempted, and attempted to solve the most difficult items more frequently. This pattern was obtained regardless of whether parametric or non-parametric results were entered as the dependent variables in the regression models. Moreover, interitem correlations

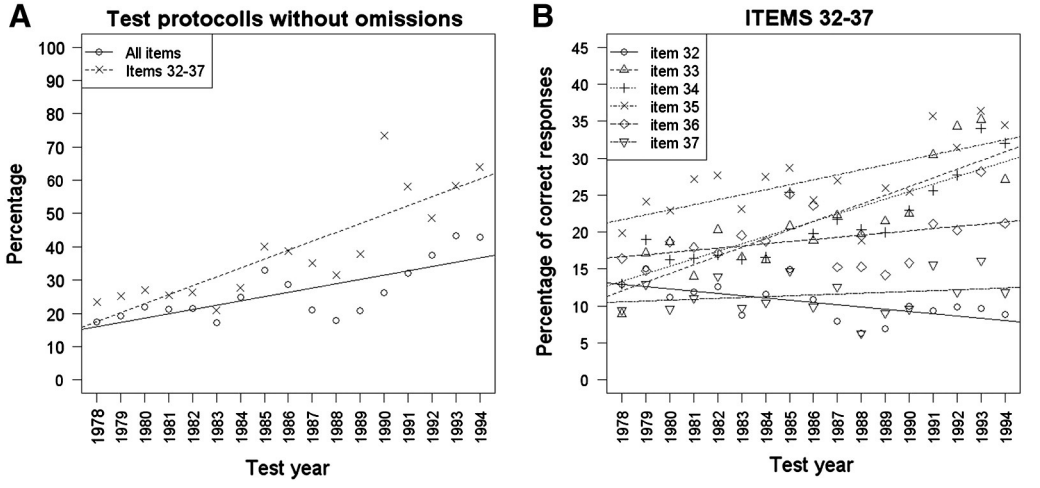

Fig. 2. (A) Percentage of test protocols in which all items have been attempted to be solved; (B) correct responses on the most difficult items.

Please cite this article as: Pietschnig, J., et al., Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypoth..., *Intelligence* (2013), http://dx.doi.org/10.1016/j.intell.2013.06.005

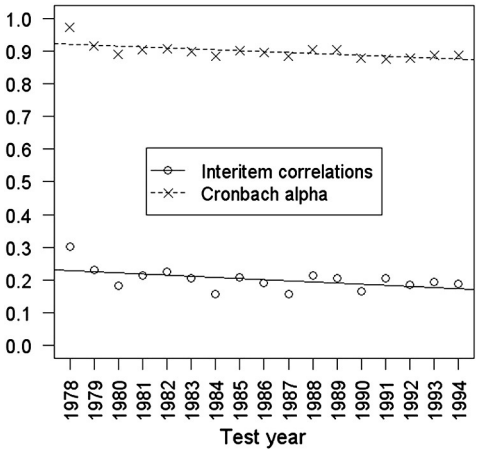

Fig. 3. Regression of test year on yearly interitem correlations and Cronbach α.

and measures of internal consistency decreased, indicating lower reliabilities that may have been due to increased guessing behavior of test-takers.

Examination of test performance variability over time did not reach nominal significance, but the sign of the regression coefficient pointed towards a decrease of variability in the performance distribution. A regression analysis of test year on the standard deviations of scores did not reach significance, although the sign was in the expected direction and visual inspection of the trajectory of percentile ranks appears to be showing a narrowing of the performance distribution. Additionally, although nominal significance was not reached, the signs of regression coefficients indicated increasing positive skewness and decreasing excess of the kurtosis of the performance distribution over time, thus supporting a narrowing of the shape of the distribution. This suggested that an improved performance of individuals from the lower tail of the ability distribution and the smaller overall performance variability resulting thereof may be responsible for higher mean test scores (Rodgers, 1999; Rowe & Rodgers, 2002). In conclusion, results from the classical test–theoretical approach indicated changed guessing behavior as well as a narrowing of the intelligence ability distribution as factors contributing to observed IQ gains.

4. Item response theory approach

4.1. Data analysis

The item response theory (IRT) analysis relied on the two-parameter model, as in previous studies of Beaujean and Osterlind (2008) and Beaujean and Sheng (2010). In this model, the probability of a correct item response depends on the person's latent ability and two item parameters, namely item discrimination (a_i) and item difficulty (b_i). Items may differ in their difficulty, but also in their discrimination, i.e., some items may discriminate more strongly between persons of different ability levels than other items (i.e., higher vs. lower discrimination). Item discrimination and item difficulty are equivalent to factor loadings and item thresholds in one-factor models in the factor-analytic framework (see Kamata & Bauer, 2008, for conversion formulae) which framework was therefore used in the current study (see below). Furthermore, the two-parameter model may either use the logistic function to model response probabilities (two-parameter logistic [2PL] model) or the normal ogive (two-parameter normal ogive [2PNO] model). 2PL and 2PNO are nearly equivalent. Parameters in the metric of the 2PN may be converted into the metric of the 2PL, using a scaling factor of 1.7. The 2PNO was used in this study.

The analysis proceeded in two steps and was based on the aggregation of data across years. Generally, fitting IRT models and testing for measurement invariance (see below) demands large samples ($n > 1000$ is often recommended). Hence, data were grouped into four cohorts, each covering assessments of four or five consecutive years: 1978–81 ($n = 1243$), 1982–85 ($n = 1376$), 1986–89 ($n = 1511$), and 1990–94 ($n = 1315$), respectively. For comparison, Beaujean and Sheng (2010) investigated the Flynn effect in 10-year cohorts.

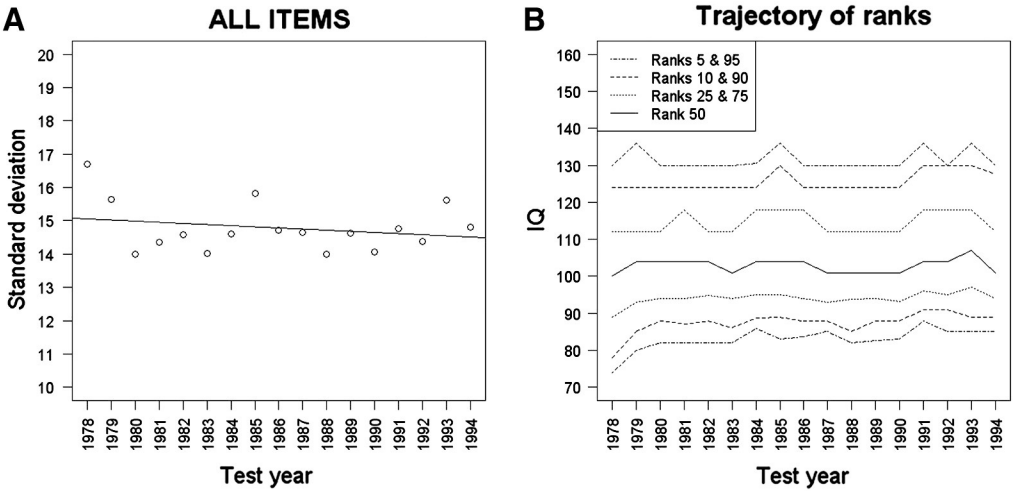

Fig. 4. (A) Regression of test year on standard deviations of mean test performance; (B) percentile ranks based on observed IQ distribution within year.

Please cite this article as: Pietschnig, J., et al., Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypoth..., *Intelligence* (2013), http://dx.doi.org/10.1016/j.intell.2013.06.005

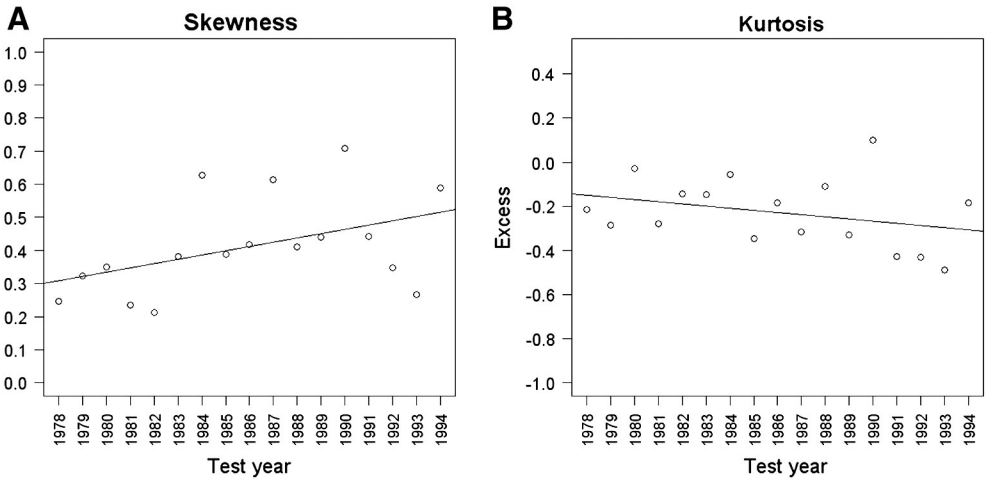

Fig. 5. (A) Regression of test year on yearly skewness and (B) kurtosis of IQ distributions.

The primary IRT analysis paralleled Beaujean and Sheng (2010) and followed procedures regarding the assessment of measurement invariance as outlined by Bontempo and Hofer (2007) and Glöckner-Rist and Hoijtink (2003). Across cohorts, we assessed (i) whether the MWT-B was unidimensional and the 2PNO fitted the data (i.e., testing for configural invariance). Given that (i) held, we further examined whether (ii) the item loadings and thresholds were invariant across cohorts (i.e., testing for full measurement invariance). In the case that full measurement invariance did not hold, we aimed to identify (iii) a subset of invariant items (i.e., testing for partial measurement invariance). These analyses served for placing respondents' latent ability estimates on the same scale in order to be able to assess differences in cohorts' latent means and variances.

The secondary IRT analysis directly investigated the impact of differences in educational level across the four cohorts. Table 1 provides a cross-tabulation of highest educational qualification and cohort. It can be seen that proportions of participants with low and medium educational levels decreased over time, whereas the proportion of participants with high educational level increased. Using the final model of the primary analysis, we re-estimated latent means and variances in the four cohorts, correcting for unequal sampling probabilities in the levels of educational level as the stratification variable.

Table 1
Highest educational qualification of individuals in the four cohorts.

Educational level	Cohorts			
	1978–81	1982–85	1986–89	1990–94
Low	392 (31.8%)	330 (24.9%)	366 (25.2%)	264 (21.7%)
Medium	583 (47.3%)	641 (48.3%)	692 (47.7%)	533 (43.9%)
High	257 (20.9%)	356 (26.8%)	393 (27.1%)	417 (34.3%)

Note. $N = 5224$ due to missing values. Low = lower secondary education or less; medium = vocational education and training; high = upper secondary education or university degree.

MPlus 6.11 and its weighted least-squares estimator with a mean- and variance-adjusted chi-squared test statistic (WLSMV) were used for all IRT analyses, fitting one-factor models on the data. WLSMV estimation is based on the items' polychoric correlation matrix and is especially suited for items with an ordered categorical item format (Beauducel & Herzberg, 2006) and was also used by Beaujean and Sheng (2010). In one-factor models, WLSMV estimation effectively fits 2PNO models on the data.

For assessment of measurement invariance, multi-group analyses were performed, using the Delta parameterization, as recommended by Muthén and Muthén (2008, p. 485) and as utilized by Beaujean and Sheng (2010). For the test of configural invariance (i.e., unidimensionality and fit of the 2PNO in all cohorts; Model 1), item loadings and thresholds were estimated freely across groups. Latent means and variances were fixed to 0 and 1 in all cohorts in order to keep the model identified. For the test of full measurement invariance (i.e., equal item discrimination and difficulty parameters in all cohorts; Model 2), item loadings and thresholds were estimated freely in the 1978–81 cohort and restricted to equality in the other cohorts. The latent mean and variance were fixed to 0 and 1 in the 1978–81 cohort, but estimated freely in the other cohorts. Using overall goodness-of-fit statistics (see below) and modification indices (MIs), we then identified item parameters that needed to be freed because of lack of invariance across groups. In a stepwise procedure, parameters were freed until a good fit was achieved. This final model (Model 3) was used for estimating latent means and variances in the four cohorts, fixing again the latent mean and variance at 0 and 1 in the 1978–81 cohort in order to keep the model identified. In a last stage of analysis, Model 3 was also contrasted with a further model (Model 4) wherein all latent variances were fixed at 1 (i.e., factor variance invariance), similar to Beaujean and Sheng (2010). For analyses of all models, all item scaling factors were fixed to unity.

In the secondary analysis, we re-estimated latent means and variances with the final model of the primary analysis. This time, information on respondents' educational level was incorporated into the model, using the stratification option of Mplus and using sampling weights that compensated for

Please cite this article as: Pietschnig, J., et al., Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypoth..., *Intelligence* (2013), http://dx.doi.org/10.1016/j.intell.2013.06.005

Table 2
Goodness-of-fit in tests of measurement invariance.

Model	χ^2	*df*	CFI	TLI	RMSEA
Primary analysis					
1. Configural invariance	5145.53	2516	.969	.968	.028 [.027, .029]
2. Full measurement invariance	4868.26	2732	.975	.976	.024 [.023, .025]
3. Partial measurement invariance	4731.31	2716	.977	.977	.023 [.022, .024]
4. Factor variance invariance	4534.40	2719	.979	.979	.022 [.021, .023]
Secondary analysis					
3. Correcting for stratification and omitting Item 1 (see text)	4699.25	2570	.973	.973	.025 [.024, .026]

Note. CFI = Comparative Fit Index; TLI = Tucker–Lewis Index; RMSEA = Root Mean Squared Error of Approximation.

differences in sampling probabilities across cohorts. Due to partly incomplete data, the overall sample size decreased from the full sample of $N = 5445$ in the primary analyses to $N = 5224$ (minus 4%) for this secondary analysis.

Goodness-of-fit was assessed with CFI (Comparative Fit Index), TLI (Tucker–Lewis Index), and RMSEA (Root Mean Squared Error of Approximation), using common cutoffs of >.95 (CFI, TLI) and of <.06 (RMSEA) to evaluate whether model fit was satisfactory (Hu & Bentler, 1999). We also tested for differences in the fit of Models 2 and 3, and of Models 3 and 4 using the *difftest* option of Mplus that provides chi-squared tests of difference in model fit for nested models with WLSMV estimation.

4.2. Results

4.2.1. Assessment of measurement invariance

One-factor models had a good fit in all cohorts, see Table 2 (Model 1). Hence, it could be safely assumed that the MWT-B measured one and the same construct across all four cohorts and that the 2PNO fitted the data overall well.

Tests of full measurement invariance (Model 2) appeared also favorably with regard to overall goodness-of-fit. However, MIs suggested that parameters of a number of items were not invariant across cohorts and needed to be freed. In a stepwise procedure, we freed item parameters that appeared to affect model fit substantially (MI > 10), until remaining MIs were <10. We freed item parameters in tandem (i.e., both the loading and threshold parameters), because response probabilities in the two-parameter model depend on both item parameters. Item parameter estimates for the four cohorts can be found in the supplementary material (Table S1). Interestingly, freeing item parameters appeared necessary specifically with regard to some of the more difficult items. Items 25 and 31 lost discriminative power over time, while Items 28, 33, and 34 became easier. Item 32 was more difficult for the 1986–89 cohort than for the other cohorts. Moreover, it was evident that the most difficult items (Items 32 to 37) generally had rather low discriminating power (Table S1). The resulting model of partial measurement invariance (Model 3) fitted the data well with regard to overall goodness-of-fit. Moreover, it also fitted significantly better on the data than Model 2 (*difftest*: $\chi^2 = 115.48$, $df = 16$, $p < .001$).

Model 4 (factor variance invariance) also appeared to fit the data well with regard to overall goodness-of-fit. However, Model 3 still fitted significantly better on the data than Model 4 (*difftest*: $\chi^2 = 9.17$, $df = 3$, $p = .027$). Hence, Model 3 was kept as the final model.

4.2.2. Correcting for differences in education

Using Model 3, we re-estimated latent means and variances, correcting for the different sampling probabilities of educational levels in the four cohorts. Item 1 was found to affect parameter estimation and overall model fit negatively in this analysis and was therefore excluded. Due to its easiness, nearly perfect correlations (e.g., $r = .995$) with other items emerged in some strata of the sample which spuriously lowered model fit.

Table 3
Effect sizes for average changes in scores across the four cohorts.

MWT-B sum scores	Cohorts			
	1978–81	1982–85	1986–89	1990–94
Mean	0.00[a]	0.74	0.25	1.16
SD	6.48	5.79	5.79	5.41
$d(t-1)$		0.12**	−0.08*	0.19***
$d(t-2)$			0.04	0.07
$d(t-3)$				0.19***
Latent scores	**1978–81**	**1982–85**	**1986–89**	**1990–94**
Mean	0.00[a]/0.00[a]	0.16/0.07	0.10/<0.01	0.23/0.11
SD	1.00[b]/1.00[b]	0.98/0.98	0.96/0.97	0.95/0.94
$d(t-1)$		0.16***/0.07	−0.07/−0.07	0.14***/0.11**
$d(t-2)$			0.10*/<0.01	0.08*/0.04
$d(t-3)$				0.24***/0.11**

Note. $d(t-i)$: Cohen *d* for the cohort in column *t* compared to the cohort *i* columns preceding it; with regard to the latent scores: cell entries pertain to results without correcting for stratification (left), and correcting for stratification (right).

[a] Reference group mean fixed to 0.
[b] Reference group variance fixed to 1.
* $p < .05$.
** $p < .01$.
*** $p < .001$.

Please cite this article as: Pietschnig, J., et al., Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypoth..., *Intelligence* (2013), http://dx.doi.org/10.1016/j.intell.2013.06.005

The fit of Model 3 without Item 1 in the secondary analysis was also good (Table 2).

4.2.3. Differences in MWT-B sum scores and in latent scores

Mean scores, standard deviations of the four cohorts, and effect sizes of the differences between cohorts are given in Table 3. The five upper lines pertain to the results in simple MWT-B sum scores which are presented as deviation scores with respect to the reference group (i.e., the 1978–81 cohort) by subtracting the mean of the reference group from the means of the respective cohorts. The five bottom lines pertain to the results in the latent scores, both uncorrected and corrected for differences in education.

For both MWT-B sum scores and uncorrected latent scores, an overall increase of scores across cohorts could be observed in similar magnitude. However, increases were not linear, as in the 1986–89 cohort both sum scores and latent scores decreased, compared to the 1982–85 cohort.

In corrected latent scores, this overall pattern of non-monotonous increases could be observed as well. However, effects appeared to be considerably weaker than for uncorrected scores. The largest observed difference, between the 1982–85 and 1990–94 cohorts, was diminished from $d = 0.24$ for uncorrected latent scores to $d = 0.11$ (minus 54%) for corrected latent scores. Differences between corrected and uncorrected latent scores did not depend on the inclusion or exclusion of Item 1. Excluding Item 1 in Model 3 in the unweighted analysis led to estimates of latent means and variances that differed by at most 0.001.

4.3. Discussion

Results from IRT-based analyses corroborated the evidence from the classical test–theoretical approach for IQ test score gains over time in the present sample. The observed increase of task performance amounted to 1.7 IQ points per decade for sum scores and to 2.1 IQ points per decade for IRT-based ability estimates (i.e., when transforming the effect size of the difference between the first and the last cohorts to the IQ metric). It should be kept in mind that all these calculations are based on the assumption of linearity of gains which however cannot be upheld in our data, since we observed considerable differences in the magnitude of changes and even a (nonsignificant) reversal of the sign of changes between adjacent cohorts (namely, between the 1982–85 and the 1986–89 cohorts). However, compared to the reference cohort (1978–81), observed changes were consistently positive across all subsequent cohorts.

Importantly, evidence for significant test score gains even held up when the results were controlled for test-takers' educational level, which, as expected, increased noticeably across cohorts. Although the latent change scores were somewhat smaller, increases of about 1.0 IQ points per decade were still observed. Accordingly, as expected, highest educational qualification moderated IQ test score gains, thus accounting for a portion of observed increases. Of further importance, measurement invariance could be shown for these data, thus indicating that test score changes reflected genuine and veridical changes of the measured latent ability itself.

Moreover, IRT-based analyses provided even stronger evidence for a narrowing of the shape of the ability distribution than the classical test–theoretical analyses: the variance of the latent scores decreased monotonically across the four cohorts. Thus, the IRT-based results clearly support the hypothesis that test performance gains are accompanied by a decreasing variability in the population.

5. General discussion

The present study shows evidence in support for changes in test-taking behavior as well as decreasing variability of test scores due to an upward shift of the lower tail of the ability distribution as two factors conceivably contributing for the Flynn effect in crystallized intelligence. Of further note, IRT-based analysis indicated measurement invariance of test scores across cohorts.

We observed considerable test score gains on a measure of crystallized intelligence, amounting to gains of 1.0 to 2.4 IQ points per decade. These results conform to previous evidence from the German-speaking area (Pietschnig et al., 2010). Contrary to most previous investigations, our analyses were not solely based on mean scores of test samples. Instead, we utilized item-level test protocols of all individuals, thus making it possible to analyze the data on the level of individual item responses, similar to a few reports in the Flynn effect research literature (see Beaujean, 2006; Beaujean & Osterlind, 2008; Beaujean & Sheng, 2010; Wicherts et al., 2004).

Analyses by standard means of classical test theory showed that the proportion of unattempted items in test protocols decreased, while the proportion of completely filled-in test protocols and attempts to solve the six most difficult items increased over time. This initial examination is suggestive of individuals over time becoming less inclined to skip items, probably because of their willingness to venture a lucky guess. Reliability indicators such as interitem correlations and internal consistency decreased significantly over time, suggesting either more uniform or more erratic answering patterns of test-takers, which would be consistent with an increase of guessing behavior. Of note, however, reliability indices maintained satisfactory values notwithstanding the significant decrease of left out items. These findings may be seen in support of Brand's statement that changes in test-taking behavior (i.e., an increased willingness to guess, due to generally more permissive attitudes in societies) may be responsible for the Flynn effect (Brand, 1987a, 1987b, 1990; Brand et al., 1989). However, in the case of an overall rather easy measure of crystallized intelligence, as used in this study, a decrease of omitted items is also a necessary side condition for the observation that test performance gains may occur at all. Hence, the results of the classical test theory analyses bearing on Brand's hypothesis appear less conclusive than desired. On the other side, keeping in mind that individuals in this sample evidently were not in a situation where they would have been under pressure to achieve high test results (i.e., there was no obvious gain to be gotten from performing well), increased guessing behavior in the present sample could reflect genuine changes of test-taking behavior.

Narrowing of the ability distribution due to decreasing test score variability (Rodgers, 1999; Rowe & Rodgers, 2002) appeared to be one factor clearly associated with intelligence test score changes. Although the regression analyses failed to reach nominal significance, regression coefficients consistently

Please cite this article as: Pietschnig, J., et al., Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypoth..., *Intelligence* (2013), http://dx.doi.org/10.1016/j.intell.2013.06.005

indicated a decreasing variability of test scores, a decreasing excess of the kurtosis, as well as an increasing positive skewness of the performance distribution over time. Moreover, visual inspection of the trajectory of percentiles associated with IQ scores showed decreasing performance variability. All of these findings suggest better performance of individuals from the lower end of the ability distribution, thus supporting the hypothesis of a narrowing of the IQ ability distribution due to an upward shift of the lower tail of the ability distribution towards the mean as one factor contributing to the Flynn effect. Importantly, this narrowing was not due to ceiling effects, since only few participants managed to solve all but one or all 37 items (1.5% or 0.3% of all participants respectively) and the number of high-scorers remained comparatively stable over the investigated period. These results are consonant with previous findings from a Spanish sample (Colom et al., 2005).

IRT-based analyses showed IQ test score gains in line with findings from classical test–theoretical approaches. These gains could be observed in analyses of sum scores as well as for the latent ability. Although the IRT-based estimates for the IQ gains appeared to be somewhat smaller than those derived from classical test–theoretical methods, they still were notable and of the same order of magnitude. Importantly, measurement invariance could be assumed in the data analyzed here, indicating that IQ gains are not merely artifactual or entirely spurious, but rather are associated with some veridical changes on the latent dimension. These findings differ from previous results for Dutch and Estonian samples, for which measurement invariance could not be upheld for a considerable number of subscales of IQ test batteries (Wicherts et al., 2004), but conform to evidence from data from the USA (Beaujean & Sheng, 2010).

When latent scores of cohorts were controlled for highest educational qualification, gains were observed to decrease in strength, albeit they remained significant. Such moderating effects of highest educational qualification are not surprising, as higher scores on crystallized intelligence test measures have frequently and foremost been linked exactly to higher education (e.g., McArdle et al., 2002). Previously, it has been suggested that education is unlikely to play a major role for generational intelligence test score gains, since such gains can already be observed among preschoolers (Lynn, 2009b; Thompson, 2012). However, our results show that although educational factors have a role regarding test score gains, they are unlikely to fully account for those gains. Rather, a considerable part appears to be due to changes of the shape of the ability distribution: a decrease of performance variability was clearly observable in our IRT-based analyses, regardless of whether highest educational qualification was controlled for or not.

Moreover, the IRT-based analyses suggested that changes in test-taking behavior, as were apparent in the classical test–theoretical analyses, were equally likely attributable to item drift (i.e., lack of invariance of item parameters across time; Goldstein, 1983). Two of the more difficult items (Items 33 and 34) and one less difficult item (Item 28) became easier over time, while two others (Items 25 and 28) lost discriminative power. Language is subject to change over time. Thus, indicators of crystallized intelligence that rely on the recognition of existing words (i.e., vocabulary tests) may likely also be subject to change over time, as is suggested by the current results. To sum up, the results of the IRT-based analyses do not provide strong and unequivocal support for changes in test-taking behavior, but instead suggest item drift of some subset of MWT-B items as a viable alternative explanation.

Different from most other studies investigating the Flynn effect, our results are not based on a representative sample of the general population, but rather on a psychiatric patient sample. However, since the intelligence test measure used and analyzed here has been shown to be a reliable and valid measure of premorbid IQ (Lehrl, 1977, p. 23; Lehrl et al., 1995), this should not impact results. On the contrary, even if impaired cognitive processing indeed impacted task performance, this would further corroborate the current findings, since any significant effects self-evidently would have been even more difficult to detect.

The major strength of this study is the large number of unobtrusively collected data on the item level over a sufficiently long time-span and stemming from a well-defined geographical (Vienna and its Eastern Austrian environs), cultural (German language), and institutional context (referrals to a clinical psychological assessment unit within a psychiatric clinic). This serendipity enabled us to examine two under-researched hypotheses for the Flynn effect, using standard means of classical test theory as well as IRT-based methods.

In all, our evidence suggests measurement invariance across cohorts and that the narrowing of the ability distribution contributes primarily to test score changes over time, while the contribution of increased guessing behavior over time seems less certain and is subject to alternative explanations. In a nutshell: Rodgers' hypothesis yes, Brand's hypothesis perhaps.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.intell.2013.06.005.

References

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, *13*, 186–203.

Beaujean, A. A. (2006). *Using item response theory to assess the Lynn–Flynn effect.* Unpublished doctoral dissertation. Columbia: University of Missouri-Columbia.

Beaujean, A. A., & Osterlind, S. J. (2008). Using item response theory to assess the Flynn effect in the National Longitudinal Study of Youth 79 children and adults data. *Intelligence*, *36*, 455–463.

Beaujean, A. A., & Sheng, Y. (2010). Examining the Flynn effect in the General Social Survey Vocabulary test using item response theory. *Personality and Individual Differences*, *48*, 294–298.

Bontempo, D. E., & Hofer, S. M. (2007). Assessing factorial invariance in cross-sectional and longitudinal studies. In A. D. Ong, & M. van Dulmen (Eds.), *Handbook of methods in positive psychology* (pp. 153–175). New York: Oxford University Press.

Brand, C. (1987a). Intelligence testing: Bryter still and Bryter? *Nature*, *328*, 110.

Brand, C. (1987b). British IQ: Keeping up with the times. *Nature*, *328*, 761.

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

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

Cattell, R. B. (1937). *The fight for our national intelligence.* London: P.S. King & Son.

Colom, R., Flores-Mendoza, C. W., & Abad, F. J. (2007). Generational changes on the draw-a-man test: A comparison of Brazilian urban and rural children tested in 1930, 2002, and 2004. *Journal of Biosocial Science*, *39*, 79–89.

Colom, R., Lluis-Font, J. M., & Andres-Pueyo, A. (2005). The generational intelligence gains are caused by decreasing variance in the lower half of the distribution: Supporting evidence for the nutrition hypothesis. *Intelligence*, *33*, 83–91.

Please cite this article as: Pietschnig, J., et al., Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypoth..., *Intelligence* (2013), http://dx.doi.org/10.1016/j.intell.2013.06.005

Daley, T. C., Whaley, S. W., Sigman, M. D., Espinosa, M. P., & Neumann, C. (2003). IQ on the rise: The Flynn effect in rural Kenyan children. *Psychological Science*, *14*, 215–219.
Dickens, W. T., & Flynn, J. R. (2001). Heritability estimates versus large environmental effects: The IQ paradox resolved. *Psychological Review*, *108*, 346–369.
Eppig, C., Fincher, C. L., & Thornhill, R. (2010). Parasite prevalence and the worldwide distribution of cognitive ability. *Proceedings of the Royal Society B*, *277*, 3801–3808.
Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin*, *95*, 29–51.
Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*, 171–191.
Flynn, J. R. (1990). Massive IQ gains on the Scottish WISC: Evidence against Brand et al.'s hypothesis. *Irish Journal of Psychology*, *11*, 41–50.
Flynn, J. R. (1998). Israeli military IQ tests: Gender differences small; IQ gains large. *Journal of Biosocial Science*, *30*, 541–553.
Flynn, J. R. (2009). *What is intelligence? Beyond the Flynn effect* (expanded paperback ed.)Cambridge, UK: Cambridge University Press.
Glöckner-Rist, A., & Hoijtink, H. (2003). The best of both worlds: Factor analysis of dichotomous data using item response theory and structural equation modeling. *Structural Equation Modeling*, *10*, 544–565.
Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement*, *20*, 369–377.
Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.
Husén, T., & Tuijnman, A. (1991). The contribution of formal schooling to the increase in intellectual capital. *Educational Researcher*, *20*, 17–25.
Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.
Kamata, A., & Bauer, D. J. (2008). A note on the relation between factor analytic and Item Response Theory models. *Structural Equation Modeling*, *15*, 136–153.
Lehrl, S. (1977). *MWT-B: Multiple-Choice Vocabulary Test [in German].* Erlangen, Germany: Straube.
Lehrl, S., Triebig, G., & Fischer, B. (1995). Multiple Choice Vocabulary Test MWT as a valid and short test to estimate premorbid intelligence. *Acta Neurologica Scandinavica*, *91*, 335–345.
Lynn, R. (1982). IQ in Japan and the United States shows a growing disparity. *Nature*, *297*, 222–223.
Lynn, R. (1989). Positive correlation between height, head size and IQ: A nutrition theory of the secular increases in intelligence. *British Journal of Educational Psychology*, *59*, 372–377.
Lynn, R. (2009a). Fluid intelligence but not vocabulary has increased in Britain, 1979–2008. *Intelligence*, *37*, 249–255.
Lynn, R. (2009b). What has caused the Flynn effect? Secular increases in the development quotient of infants. *Intelligence*, *37*, 16–24.
Lynn, R. (2013). Who discovered the Flynn effect? A review of early studies of the secular increase of intelligence. *Intelligence*, http://dx.doi.org/10.1016/j.intell.2013.03.008.
Mahlberg, A. (1997). The rise in IQ scores. *American Psychologist*, *52*, 71.
McArdle, J. J., Ferrer-Caja, E., Hamagami, F., & Woodcock, R. W. (2002). Comparative longitudinal structural analyses of the growth and decline of multiple intellectual abilities over the life span. *Developmental Psychology*, *38*, 115–142.
Merrill, M. A. (1938). The significance of IQ's on the revised Stanford–Binet scales. *Journal of Educational Psychology*, *29*, 641–651.
Mingroni, M. A. (2004). The secular rise in IQ: Giving heterosis a closer look. *Intelligence*, *32*, 65–83.
Mingroni, M. A. (2007). Resolving the IQ paradox: Heterosis as a cause of the Flynn effects and other trends. *Psychological Review*, *114*, 806–829.
Muthén, B., & Muthén, L. (2008). *Mplus user's guide*. Los Angeles, CA: Muthén & Muthén.
Neisser, U. (1997). Rising scores on intelligence tests. *American Scientist*, *85*, 440–447.
Pietschnig, J., Voracek, M., & Formann, A. K. (2010). Pervasiveness of the IQ rise: A cross-temporal meta-analysis. *PLoS One*, *5*, e14406.
Pietschnig, J., Voracek, M., & Formann, A. K. (2011). Female Flynn effects: No sex differences in generational IQ gains. *Personality and Individual Differences*, *50*, 759–762.
Rodgers, J. L. (1999). A critique of the Flynn effect: Massive IQ gains, methodological artifacts, or both? *Intelligence*, *26*, 337–356.
Rowe, D. C., & Rodgers, J. L. (2002). Expanding variance and the case of historical changes in IQ means: A critique of Dickens and Flynn (2001). *Psychological Review*, *109*, 759–763.
Schooler, C. (1998). Environmental complexity and the Flynn effect. In U. Neisser (Ed.), *The rising curve: Long-term gains in IQ and related measures* (pp. 67–79). Washington, DC: American Psychological Association.
Sigal, M. J., & McKelvie, S. J. (2012). Is exposure to visual media related to cognitive ability? Testing Neisser's hypothesis for the Flynn effect. *Journal of Articles in Support of the Null Hypothesis*, *9*, 23–50.
Storfer, M. (1999). Myopia, intelligence, and the expanding human neocortex: Behavioral influences and evolutionary implications. *International Journal of Neuroscience*, *98*, 153–276.
Sundet, J. M., Barlaug, D. G., & Torjussen, T. M. (2004). The end of the Flynn effect? A study of secular trends in mean intelligence test scores of Norwegian conscripts during half a century. *Intelligence*, *32*, 349–362.
Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn effect in reverse. *Personality and Individual Differences*, *39*, 837–843.
Thompson, J. (2012). Richard Lynn's contributions to personality and intelligence. *Personality and Individual Differences*, *53*, 157–161.
Tuddenham, R. D. (1948). Soldier intelligence in World Wars I and II. *American Psychologist*, *3*, 54–56.
Voracek, M. (2002). *Three studies on the Multiple-Choice Vocabulary Test (MWT): Test reanalysis, Lynn–Flynn effect, and phase-IV study [in German].* Unpublished doctoral dissertation. : University of Vienna.
Voracek, M. (2006). Phlogiston, fluid intelligence, and the Lynn–Flynn effect. *The Behavioral and Brain Sciences*, *29*, 142–143.
Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., & Span, M. M. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*, 509–537.
Woodley, M. A. (2011). Heterosis doesn't cause the Flynn effect: A critical examination of Mingroni (2007). *Psychological Review*, *118*, 689–693.
Woodley, M. A. (2012). A life history model of the Lynn–Flynn effect. *Personality and Individual Differences*, *53*, 152–156.
Zajonc, R. B., & Mullally, P. R. (1997). Birth order: Reconciling conflicting effects. *American Psychologist*, *52*, 685–699.

Please cite this article as: Pietschnig, J., et al., Item-response theory modeling of IQ gains (the Flynn effect) on crystallized intelligence: Rodgers' hypothesis yes, Brand's hypoth..., *Intelligence* (2013), http://dx.doi.org/10.1016/j.intell.2013.06.005