Available online at www.sciencedirect.com
ScienceDirect
Intelligence 36 (2008) 455–463
INTELLIGENCE

Using Item Response Theory to assess the Flynn Effect in the National Longitudinal Study of Youth 79 Children and Young Adults data ☆

A. Alexander Beaujean [a,*], Steven J. Osterlind [b]

[a] *Baylor University, United States*
[b] *University of Missouri-Columbia, United States*

Received 24 February 2006; received in revised form 4 October 2007; accepted 24 October 2007
Available online 20 February 2008

Abstract

The purpose of this manuscript is to assess the magnitude of the Flynn Effect (i.e., increase in mean IQ scores across time) using Item Response Theory (IRT). Unlike using methods derived from Classical Test Theory, IRT has the capability to determine if the Flynn Effect is due to a genuine increase in intelligence, if it is due to a psychometric artifact (i.e., items changing properties over time), or a combination of the two. Using the Peabody Picture Vocabulary Test—Revised and Peabody Individual Achievement Test—Math from the *National Longitudinal Study of Youth 79 Children and Young Adults*, the results of this study indicate that while using raw and standardized scores, the Flynn Effect is evident in a predicted magnitude, but when using scores based from IRT analysis, the magnitude Flynn Effect substantially decreases, and, at least for the Peabody Picture Vocabulary Test—Revised, goes away. Thus, for the data used in this study, the Flynn Effect appears to be largely the result of changing item properties instead of changes in cognitive ability.

Defined, the Flynn Effect (FE) is the rise of psychometric IQ test scores over time. It is named after New Zealand political scientist James R. Flynn (1983, 1984, 1987, 1999). Since Flynn's original work, the FE has been studied in many different populations across the globe (Bolen, Aichinger, Hall, & Webster, 1995; Daley, Whaley, Sigman, Espinosa, & Neumann, 2003; Sanborn, Truscott, Phelps, & McDougal, 2003; Sundet, Barlaug, & Torjussen, 2004; Truscott & Frank, 2001). While multiple sources have found that psychometric IQ has been rising, general intelligence (*g*) has not shown a parallel increase (Rushton, 1999; Must, Must, & Raudik, 2003, but see Colom, Juan-Espinosa, & Garcia, 2001). Further, reaction time–a reliable correlate with *g*–has not decreased over the same time periods (Nettelbeck & Wilson, 2004). To add to the difficulty in understanding the effect, while the FE appears to impact scores throughout the entire range of the IQ distribution, it is particularly concentrated for lower ability examinees (Colom, Lluis-Font, & Andres-Pueyo, 2005; Teasdale & Owen, 1989; but see Flynn, 2006b).

Attempts to explain the various findings involved in the FE have led scholars down a wide range of avenues of inquiry into its nature and putative causes. Some

☆ This manuscript was based on the first author's doctoral dissertation.

* Corresponding author. Department of Educational Psychology, Baylor University, P.O. Box 97301, Waco, TX 76798-7301.
E-mail address: abeaujean@gmail.com (A.A. Beaujean).

0160-2896/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2007.10.004

researchers, such as Lynn (1989, 1990), and Eysenck and Schoenthaler (1997) posit that massive environmental changes, such as dramatic changes in available nutrition, have, at least in part, been responsible for the IQ increase. Others, like Blair, Gamsonb, Thornec, and Bakerd (2005) explain the mean IQ increase as an artifact of educational curriculum developments, especially in mathematics instruction. Some theorists go the other direction, holding that the FE is not due to an environmental effect at all, but rather is a byproduct of an increase in heterosis (outbreeding; Mingroni, 2004). There are still others that opine the FE is not much more than a psychometric artifact (Brand, 1996; Brand, Freshwater, & Dockrell, 1989; Burt, 1952; Rodgers, 1999; Wicherts et al., 2004), or that perhaps the FE no longer is even operative, at least in certain countries (Sundet et al., 2004; Teasdale & Owen, 2005).

Methodologically, most FE studies rely on procedures based on statistical and measurement models derived from Classical Test Theory (CTT). [There are two notable exceptions: work done by Flieller (1988) and by Wicherts (2005; Wicherts et al., 2004).] CTT is concerned with the estimation of a "true score" and statistical comparisons with true score models are of summed raw scores or factor scores (e.g., the Wechsler FSIQ) (Lord & Novick, 1968). The near-universal reliance on the CTT model in FE work is unfortunate because the models do not make use of all the information available in a given test performance, and, consequently, cannot differentiate between the latent constructs they are designed to measure, and the test scores that purport to measure them. In contrast, modern measurement methods, such as Item Response Theory (IRT), allow for more appropriate estimation of a change in underlying latent ability, and can help rectify the problems in using CTT (Borsboom, 2005).

More concretely, raw scores, or amalgamated full-scale scores, might not be the best types of variables to use in FE studies because an increase in scores could be due to a number of possibilities, such as an increase in cognitive ability, a systemic decline in item-level difficulty, or an interaction of these possibilities. Unless it can be shown that the measures are invariant between the groups, CTT models are not sensitive to such sources of variance and can show a difference in true scores, even if there is no change in the underlying latent variable (Beaujean, 2005; Meade, Lautenschlager, & Hecht, 2005). IRT, on the other hand, can assess these properties for items on a given test, which, theoretically, allows the researcher to discern a difference between a true rise in intelligence (measured via a latent construct), changing item properties, or an interaction of the two.

1. Item Response Theory

Item response models specify how an individual's (latent) trait level and an item's properties are related to how an examinee responds to that item, as well as a set of items (Hambleton & Swaminathan, 1985; Lord, 1980). Unlike classical test models, it can concurrently model both a examinee's ability, θ, and item properties: difficulty (b), discrimination (a), and a guessing correction (c) (for alternative IRT models, see Sijtsma & Molenaar, 2002; van der Linden & Hambleton, 1997). Modeling all three item properties results in a three-parameter model, while constraining c to zero results in a two-parameter model. If c is held to zero and a is held constant across all items, then this results in a one-parameter model. To determine what model to use for a set of items requires both theory and data (i.e., testing the item-model fit).

Much has been written on IRT (Embretson & Prenovost, 1999; Embretson & Reese, 2002; Hambleton & Swaminathan, 1985; Hambleton, Swaminathan, & Rogers, 1991; Lord, 1980) and the interested reader can find more in-depth explication of the theory and models there. For the purposes of this paper, only a few details need recognized. First, when a and b are estimated separately for two groups of examinees, each of the two sets of variables has an arbitrary scale that are not directly comparable. However, the two sets are linearly related, so a simple transformation can be used to convert the parameter estimates of one group to the scale of the other group for purposes of comparison. Most current IRT computer programs can do this calibration, so results from the different groups are on the same scale.

Second, while test equating using CTT requires that groups have similar ability distributions, this is not so with IRT models. Instead, IRT methods allow for *non-equivalent groups equating* (Zimowski, 2003), which allows test equating for groups who significantly differ in distribution of θ. Concerning the FE, if intelligence is actually rising, then the individuals who took the test at different time points can be placed on the *same* underlying θ (ability) distribution, which makes ability comparisons especially easy, as one can determine how many standard deviations one group's (average) cognitive ability is from another's (Yu & Osborn Popp, 2005).

Third, and for the purposes of studying the FE, the most salient aspect of IRT, is that *"the item parameters are not dependent upon the ability level of the examinees responding to the items"* (Baker, 2001, p. 52, emphasis added). Thus, two groups who differ widely in θ can take the same test and the item parameters should be the same (within a linear transformation). This is starkly different

from CTT models, where item properties are directly related to the sample of test takers (Crocker & Algina, 1986). Regarding the FE, invariance means that item parameters should not differ between samples, even if the samples take the test some number of years apart and the underlying ability (θ) of the groups has had a mean increase. Of course, this claim needs to be empirically assessed for all common items across the instruments, a topic that falls under the rubric of *differential item functioning*.

Defined, differential item functioning (DIF) occurs when item parameters differ across subgroups (e.g., race, age, time; Osterlind, 1983). To determine if DIF exists using IRT models, one just examines if the (linearly transformed) parameters for the IRT models differ between two (time) groups. This is done by computing the item parameter estimates within each group, calibrating the item parameters to put them on a common scale (after correcting for the possibility that θ's distribution might differ between the groups), and then testing to see if the item parameters differ between the groups (Lord, 1980). If a given item is shown to exhibit DIF, however, it can still be used in the estimate of the θ parameter for the different groups. It is just assumed that the item is measuring differently between the groups (i.e., is not invariant; Raju, Laffitte, & Byrne, 2002) and is not used when equating the groups' scores (Reise, Widaman, & Pugh, 1993).

Perhaps a more concrete example would be useful. In large-scale educational assessments, it is common to put scores from different grades on the same scale (i.e., θ) to make comparisons easier. However, it would not be prudent to, say, give 5th graders and 3rd graders the same items; difficult items for 5th graders would not discriminate well for 3rd graders, and easy items for 3rd graders would not discriminate well for 5th graders. However, if there is a sufficient number of common items between the two tests that do not exhibit DIF, the two tests can be placed onto the same θ scale, which allows for a direct comparison between 5th grade and 3rd grade scores. The items that were not used for equating, however, are still useful as they can help determine θ within a grade.

In studying the FE, the same concept applies. As long as there is a sufficient number of items not exhibiting DIF across the groups, the underlying θ for both groups can be placed on the same scale. The items not used for equating are still useful as they give information about θ within a group. However, they should not be used in the direct assessment of change in ability over time. Consequently, before determining if ability has actually changed over time, FE studies need to first determine if the items are measuring the same construct, the same way, at different points in time. This is conceptually similar to using multigroup confirmatory factor analytic methods, and testing for measurement invariance (Raju et al., 2002). Although the theory behind the two types of analysis is distinct (Meade & Lautenschlager, 2004; Reise et al., 1993), the two methods have much to contribute to the FE literature (see the discussion in Wicherts et al., 2004).

2. Method

Data for our Flynn Effect (FE) examination came from the Children and Young Adults of 1979 edition of the *National Longitudinal Survey of Youth* (CNLSY79) (Bureau of Labor Statistics, U.S. Department of Labor, & National Institute for Child Health and Human Development, 2002).[1] More specifically, the 1990 and 2004 administrations of the Peabody Picture Vocabulary Test—Revised (PPVT-R; Dunn & Dunn, 1981) and the 1996 and 2004 administrations of the Math section of the Peabody Individual Achievement Test (PIAT-Math;Dunn & Markwardt, 1970) were used. Although the tests were administered to children/adolescents in the 1980s, there were significant changes in the administration and scoring procedures to the instruments in the early 1990s, making the comparison of standard and/or raw scores tenuous (Center for Human Resource Research, 2004).

The PPVT-R and PIAT-Math tests were chosen for multiple reasons: (a) item-level data were available; (b) the tests' content is very homogenous; (c) while the PIAT-Math is an achievement test, Scott, Bengston, and Gao (1998) found some evidence of a FE in math achievement; (d) the PPVT-R is often used as a proxy for IQ and correlates highly with various IQ instruments (Dunn & Dunn, 1997, p. 62), and (e) the sample from which the data was drawn (all the children born to *National Longitudinal Study of Youth 79* mothers) is comprised of a wide variety of US national subpopulations.

For a respondent to be included in the current study, he or she (a) had to have his/her raw, standardized, and percentile scores reported in the CNLSY79 database; (b) have answered enough questions correctly to reach the basal level for the respective test (8 for the PPVT-R; 5 for the PIAT-Math); (c) be between ages 4 and 12 for the PPVT-R administration or between ages 5 and 15 in the PIAT-Math administration to make sure the age groups were consistent across cohorts; and (d) not have had the test administration ended prematurely (i.e., the test administration had to end because the responded reached a ceiling or answered all the test items). Selecting on these criteria produced sample sizes that were different across the age cohorts for both tests, so, of the eligible respondents, a random sample was chosen such that the n was the same

[1] The CNSLY79 survey is sponsored and directed by the U.S. Bureau of Labor Statistics and the National Institute for Child Health and Human Development. The survey is managed by the Center for Human Resource Research at The Ohio State University and interviews are conducted by the National Opinion Research Center at the University of Chicago.

within an age group. The sample size was determined by using the minimum n from either administration sample. For example, if the number of four year-old respondents for the 1990 PPVT-R administration was 100 and the number of four year-old respondents for the 2004 PPVT-R administration was 150, then all respondents in the 1990 group were used and a random sample of 100 from the 2004 sample was used. The sample sizes are reported in Tables 1 and 2.

To assess the unidimensionality assumption (i.e., that the tests are only measuring one construct), a full information factor analysis was run (Bock, Gibbons, & Muraki, 1988) for both years of both tests using the TESTFACT program (Wood et al., 2004). The results indicated that the unidimensionality assumption is warranted across all tests and administration years.

For the IRT analysis, the item parameters were estimated in the BILOG-MG program (Zimowski, Muraki, Mislevy, & Bock, 1996). Initially, each administration of each test was analyzed separately to determine if a one-, two-, or three-parameter IRT model fit the data best. Using results from BILOG-MG's item χ^2 analysis, as well as a change in −2 Log Likelihood values (for more information on the tests, see Embretson & Reese, 2002; Zimowski et al., 1996; Embretson & Reese, 2002), a two-parameter model appeared to fit the item data best for all administrations of all tests. This means the "guessing" parameter (c) was constrained to 0, but the discrimination and difficulty parameters (a and b, respectively) were freely estimated.

The tests were then assessed for DIF in the b parameter using the DIF procedure in BILOG-MG (BILOG-MG does not allow for detection of DIF in the a parameter). Using an $\alpha = .01$, the ratio of the difference of (adjusted) b parameters between groups to the difference's standard error was compared for each item, as the ratio is asymptotically distributed as a standard normal (Zimowski et al., 1996). The remaining items were assessed for DIF in the a parameter using the

Table 1
Peabody Individual Achievement Test—Math data

Age		1996 PIAT-Math					2004 PIAT-Math			
		Raw	Standard	Percentile	IRT		Raw	Standard	Percentile	IRT
All	n	2006	2006	2006	2006	n	2006	2006	2006	2006
	$\bar{X}$	42.04	100.45	50.97	0.00	$\bar{X}$	44.68	104.39	58.55	0.13
	s	15.10	13.66	27.11	1.00	s	16.03	15.25	28.86	1.00
5	n	44	44	44	44	n	44	44	44	44
	$\bar{X}$	10.86	96.32	41.11	−2.21	$\bar{X}$	12.89	106.05	61.50	−1.97
	s	4.04	15.07	28.16	0.43	s	4.43	15.62	29.84	0.46
6	n	120	120	120	120	n	120	120	120	120
	$\bar{X}$	15.56	102.36	55.02	−1.78	$\bar{X}$	16.18	102.71	54.94	−1.68
	s	6.05	15.09	28.70	0.52	s	5.58	13.57	27.54	0.45
7	n	161	161	161	161	n	161	161	161	161
	$\bar{X}$	23.16	103.01	56.50	−1.22	$\bar{X}$	24.75	105.03	60.01	−1.11
	s	8.17	12.09	24.72	0.55	s	9.53	12.37	24.96	0.63
8	n	156	156	156	156	n	156	156	156	156
	$\bar{X}$	30.01	101.43	52.66	−0.79	$\bar{X}$	33.81	106.25	61.89	−0.54
	s	9.44	13.16	26.34	0.59	s	11.42	15.56	29.34	0.74
9	n	167	167	167	167	n	167	167	167	167
	$\bar{X}$	37.66	102.05	53.75	−0.29	$\bar{X}$	41.34	107.51	64.59	−0.06
	s	9.98	14.88	29.26	0.64	s	11.13	16.59	30.11	0.70
10	n	205	205	205	205	n	205	205	205	205
	$\bar{X}$	42.24	100.45	51.04	0.02	$\bar{X}$	47.20	108.59	66.88	0.31
	s	8.77	14.07	28.00	0.59	s	9.51	14.63	27.16	0.61
11	n	254	254	254	254	n	254	254	254	254
	$\bar{X}$	47.62	102.37	55.09	0.39	$\bar{X}$	49.39	104.85	59.61	0.45
	s	8.91	14.68	28.39	0.61	s	10.09	16.13	29.80	0.63
12	n	296	296	296	296	n	296	296	296	296
	$\bar{X}$	50.54	101.64	53.28	0.57	$\bar{X}$	52.19	104.11	58.28	0.61
	s	7.92	12.33	25.51	0.51	s	9.92	14.93	28.77	0.59
13	n	293	293	293	293	n	293	293	293	293
	$\bar{X}$	51.04	98.13	46.99	0.60	$\bar{X}$	54.42	102.68	55.28	0.73
	s	9.69	13.73	26.29	0.61	s	10.20	15.01	28.15	0.59
14	n	295	295	295	295	n	295	295	295	295
	$\bar{X}$	53.66	97.26	44.02	0.75	$\bar{X}$	56.06	100.83	51.39	0.80
	s	9.02	12.31	25.30	0.54	s	10.66	15.36	28.96	0.58
15	n	15	15	15	15	n	15	15	15	15
	$\bar{X}$	52.80	94.60	37.93	0.66	$\bar{X}$	53.20	95.07	39.67	0.65
	s	8.16	10.47	24.31	0.45	s	10.39	16.28	29.71	0.56

Table 2
Peabody Picture Vocabulary Test—Revised data

Age	1990 PPVT-R					2004 PPVT-R				
		Raw	Standard	Percentile	IRT		Raw	Standard	Percentile	IRT
All	n	717	717	717	717	n	717	717	717	717
	$\bar{X}$	89.07	87.03	30.71	0.00	$\bar{X}$	92.99	95.71	45.40	0.06
	s	30.39	19.42	27.64	1.00	s	32.18	22.88	32.97	1.10
4	n	43	43	43	43	n	43	43	43	43
	$\bar{X}$	36.79	85.91	29.09	−1.67	$\bar{X}$	41.21	95.42	43.09	−1.67
	s	13.64	18.79	27.24	0.43	s	13.83	17.14	31.40	0.40
5	n	94	94	94	94	n	94	94	94	94
	$\bar{X}$	47.12	86.67	31.65	−1.38	$\bar{X}$	49.52	92.96	41.81	−1.43
	s	18.58	23.46	33.21	0.56	s	17.88	23.73	32.27	0.52
6	n	37	37	37	37	n	37	37	37	37
	$\bar{X}$	56.05	85.14	29.27	−1.10	$\bar{X}$	60.30	92.84	42.84	−1.08
	s	19.01	20.94	28.70	0.58	s	19.69	23.46	33.92	0.59
7	n	17	17	17	17	n	17	17	17	17
	$\bar{X}$	70.35	85.94	29.71	−0.64	$\bar{X}$	81.53	103.35	57.65	−0.42
	s	19.10	20.38	34.33	0.64	s	14.48	17.73	33.23	0.50
8	n	11	11	11	11	n	11	11	11	11
	$\bar{X}$	72.09	78.27	16.27	−0.62	$\bar{X}$	81.73	90.45	52.27	−0.35
	s	10.18	17.07	17.45	0.32	s	31.48	35.92	32.13	1.05
9	n	9	9	9	9	n	9	9	9	9
	$\bar{X}$	71.11	64.56	24.89	−0.61	$\bar{X}$	93.22	93.00	42.00	0.06
	s	27.16	30.13	28.06	0.90	s	18.07	20.61	32.66	0.70
10	n	130	130	130	130	n	130	130	130	130
	$\bar{X}$	101.49	88.46	32.60	0.41	$\bar{X}$	102.96	96.62	48.45	0.42
	s	15.13	17.62	27.34	0.54	s	19.74	22.91	32.75	0.71
11	n	251	251	251	251	n	251	251	251	251
	$\bar{X}$	105.66	89.22	32.41	0.55	$\bar{X}$	108.88	96.26	45.35	0.61
	s	14.67	17.14	26.44	0.51	s	21.06	23.52	33.86	0.74
12	n	125	125	125	125	n	125	125	125	125
	$\bar{X}$	107.46	84.89	27.45	0.60	$\bar{X}$	113.45	96.29	44.54	0.75
	s	19.16	20.16	25.22	0.63	s	19.70	22.08	32.48	0.66

Likelihood ratio test (Camilli & Shepard, 1994) as implemented in the program IRTLRDIF (Thissen, 2001), using the same α of .01, but using a one df χ^2 distribution instead of the standard normal distribution.

As the ns for the PIAT-M and PPVT-R are large, the standard errors will necessarily be small (assuming the items are not extremely easy or difficult), which could result in too many false positives. As the number of common items was plentiful for both instruments, we felt it better to err on the side of removing too many items from the equating set than to leave in items that could possibly exhibit DIF. Any items not exhibiting DIF were used as equating (anchor) items, and the two forms of the test (using all items) were equated, again, in BILOG-MG. This allows the estimated latent trait (θ) scores to be directly comparable across groups.

3. Results

3.1. Differential item functioning

Using the procedure for investigating differential item functioning (DIF) as outlined above, 87 items on the PPVT-R showed DIF in the b parameter, and none showed DIF in the a parameter. For the PIAT-Math, 33 items showed DIF in the b parameter, and 1 item showed DIF in the a parameter. The items exhibiting DIF are listed in Appendix A. The items found not exhibiting DIF were all used as anchoring items to equate the two forms of the tests.

3.2. Peabody Individual Achievement Test—Math

The results from the PIAT-Math analysis are shown in Table 1, with the columns labeled IRT being the derived IRT latent trait scores. The table includes data from raw, standard, and percentile scores as well. To facilitate the across-year comparison, Cohen's (1988) d (with a pooled standard deviation) was calculated for the raw scores, standard scores, percentile scores, and the IRT latent trait scores (see Table 3). For all calculations, the 1996 scores were subtracted from the 2004 scores, meaning that a positive number indicates an increase over time, whereas a negative number indicates a decrease over time.

Looking at the results for all ages, there was an increase in raw, standardized, and percentile scores of the magnitude of .17, .27, and .27 standard deviations, respectively. For the IRT scores, though, there was a smaller increase over time of the

Table 3
Effect sizes for Peabody Picture Vocabulary Test—Revised and Peabody Individual Achievement Test—Math

Age	Raw	Standard	Percentile	IRT
PIAT-Math				
All	0.17	0.27	0.27	0.13
5	0.48	0.63	0.70	0.56
6	0.11	0.02	0.00	0.21
7	0.18	0.17	0.14	0.20
8	0.36	0.33	0.33	0.36
9	0.35	0.35	0.36	0.34
10	0.54	0.57	0.57	0.50
11	0.19	0.16	0.16	0.10
12	0.18	0.18	0.18	0.07
13	0.34	0.32	0.30	0.22
14	0.24	0.26	0.27	0.10
15	0.04	0.03	0.06	−0.02
PPVT-R				
All	0.13	0.41	0.48	0.06
4	0.32	0.53	0.48	0.00
5	0.13	0.27	0.31	−0.09
6	0.22	0.35	0.43	0.04
7	0.66	0.91	0.83	0.38
8	0.41	0.43	1.39	0.35
9	0.96	1.10	0.56	0.84
10	0.08	0.40	0.53	0.02
11	0.18	0.34	0.43	0.10
12	0.31	0.54	0.59	0.24

magnitude of .13 standard deviations, less than half the increase of the standardized and percentile scores. While this pattern does not hold over all groups when the data are broken down into the various age groups, it does for the age groups with appreciable sample sizes (i.e., $n > 200$).

3.3. Peabody Picture Vocabulary Test—Revised

The results from the PPVT-R analysis are shown in Table 2, with the columns labeled IRT being the derived IRT latent trait scores. As with the PIAT-Math scores, Cohen's (1988) *d* (with a pooled standard deviation) was calculated for all score types to facilitate comparison (see Table 3). Like the PIAT-Math, the raw, standardized, and percentile scores show an increase over time of the magnitude of .13, .41, and .48 standard deviations, but the IRT scores show a negligible increase over time of the magnitude of .06. This pattern is generally repeated when the data are grouped by age, when the *n* is of appreciable size.

4. Discussion

The purpose of this study was to assess the Flynn Effect (FE) by Item Response Theory (IRT) methods using scores on the Peabody Picture Vocabulary Test—Revised (PPVT-R) and scores on the Peabody Individual Achievement Test—Math (PIAT-Math) from the data provided by the Children and Young Adults of 1979 edition of the *National Longitudinal Survey of Youth* (CNLSY79). When all age groups were averaged, this study found that on both the PPVT-R and the PIAT-Math, the raw scores, standard scores, and percentile scores showed an increase in score performance over time. This result is confirmed by Rodgers and Wanstrom (2007), who found a FE in the same data (using regression methods) before adjusting for mother's IQ score. When using IRT-based scores, however, an increase in scores was also found, but was much smaller and, at least for the PPVT-R, a very negligible increase. When the age groups were analyzed separately, in general, similar results were found, although there were some exceptions. However, as the *n* for many age groups was very small, any effect (or lack thereof) evidenced should be interpreted with caution.

The results from this study fit well into some of the more recent literature concerning the FE. First, as Sundet et al. (2004) and Teasdale and Owen (2005) have recently shown, there appears to be a stop in the FE starting (at least) sometime in the 1990s. It is very hard to speculate as to the reason for this, as the reasoning behind the initial FE was not well understood. As the FE (or lack thereof) can have very potent "real world" effects (Beaujean & Guiling, 2006; Flynn, 2006a; Kanaya, Scullin, & Ceci, 2003), there needs to be much more research devoted to this issue both by research scholars and clinical practitioners. Of particular note in this research is a need for better understanding of the psychological processes involved in changing item parameters versus changing levels of cognitive ability. Put another way, why do some item parameters change over time, and are these reasons of any practical significance, above and beyond the significance of general cognitive ability? As the CNLSY79 data does not provide the actual item stems for the PPVT-R or PIAT-Math, the current study cannot answer such questions.

Second, this study aligns well with some recent scholarship (Beaujean, 2005; Wicherts, 2005, December; Wicherts et al., 2004) that has shown the difficulty in unfolding the FE without using more sophisticated psychometric models than just differences between raw or full-scale scores. As Classical Test Theory-derived scores are unable to partial out the influence of factors other than the trait being measured (Borsboom, 2005), comparing them without controlling for, or at least investigating, other related issues, such as measurement invariance, is a potentially risky research endeavor where the results could be very misleading (Meade et al., 2005). For example, in looking at the PPVT-R in the current study, if one compared standard scores across time, one would have concluded that there has been a rise of 6.15 points (i.e., .41 * 15) over 14 years, which turns out to be approximately .44 points a

year—a figure close to Flynn's (1987) estimates. However, looking at the IRT-based scores, one gets a very different picture because the results show a miniscule change over time (a .06 point increase). Of course, these findings do not necessarily negate the use of Classical Test Theory models in Flynn Effect investigations, but do call for more research in this area, especially controlled experiments and simulation studies (e.g., Beaujean, 2005) designed to compare Classical Test Theory and IRT-derived scores.

Two separate issues that arise from this study and its findings are how it coalesces with other research that has shown that the rise in IQs over time might be real increases (Eysenck & Schoenthaler, 1997; Lynn, 1989, 1990), or even a rise in *g* (Colom et al., 2001; te Nijenhuis & van der Flier, 2006). To the latter, a possible answer is that the analysis was done with CTT-derived scores and exploratory factor analysis. These methods cannot assess measurement invariance or item invariance (Raju et al., 2002; Reise et al., 1993; Wicherts et al., 2004), and so cannot adequately separate out variance due to *g* or change in measurement instrumentation. A second argument is that there are other studies (e.g., Rushton, 1999) that show different findings, and more research needs to be accumulated before any definitive statement can be made. To the former, a possible answer is that nutrition had an influence in the mid-twentieth century, but its effects have leveled off. In Lynn's (1990) thorough review, a large portion of the studies cited were conducted before the 1980s, so it is possible that the effects of nutrition had an effect mid-century, but have exhausted their ability to further increase cognitive abilities by the time the current data was collected post 1990. All of these surmises, of course, are merely hypotheses that bare further research.

A further issue that arises from this, and most other FE studies, is the validity of IQ scores. We feel that, at least within a generation, the data is replete enough to say that scores are valid measures of individual differences in cognitive ability, given the instrument was administered in a language in which the test taker was fluent (Deary, 2000; Jensen, 1998). Moreover, given their powerful predictive capabilities (Deary, Strand, Smith, & Fernandes, 2007; Schmidt & Hunter, 2004), we feel that IQ scores do have, and will continue to have, a place in educational and occupational settings. What this study has called into question is the direct comparison of CTT-based scores across generations without assessing measurement invariance.

Last, with the *en masse* distribution of personal computers and the user-friendly software available (e.g., BILOG-MG, Mplus, R), IRT-based analysis of FE data is not much more difficult, or least not much more time consuming, than the more traditional CTT analysis. Thus, it is hoped that IRT analysis will begin to make its way into this field. In addition, we hope that researchers doing FE data collection will allow for secondary analysis of their data at the item level and, perhaps more importantly, standardized IQ test publishers will begin to make their item-level data available to qualified researchers so that IRT analysis can start to be applied to norming data across instrument editions.

4.1. Limitations

As an anonymous reviewer of the paper indicated, this study should be viewed in the context of an exploratory study. That is, other studies, using similar latent variable methods but different measuring instruments, should be conducted before anything definitive is stated about the FE. This would be particularly informative if it were done with data gathered in the mid-to-late twentieth century, when the FE appears to be most evident. Moreover, future studies should look to gather data specifically from tests more akin to typical IQ tests, such as that from the Wechsler instruments or the various Raven's matrices tests.

Acknowledgements

The authors would like to thank James Flynn and two anonymous reviewers for their comments and suggestions regarding this manuscript.

Appendix A

Items exhibiting differential functioning

PPVT-R		PIAT-Math	
b	*a*	*b*	*a*
3	None	12	7
7		14	
20		16	
22		18	
27		27	
31		32	
33		37	
36		42	
40–42		47	
44–48		51	
50–52		53–59	
55–56		61–68	
58		70–72	
60		74–76	
64–68		78	
70		80	
72–74			
76–77			
81–82			
85–87			
90			

(continued on next page)

Appendix A (*continued*)

PPVT-R		PIAT-Math	
b	*a*	*b*	*a*
92–93			
95–96			
98–100			
106			
108			
111			
113–117			
119–121			
125			
130			
133–136			
139–144			
147–151			
153			
155–156			
159			
161–162			
164			
166–167			
172–174			

b: threshold/difficulty.

References

Baker, F. B. (2001). The basics of Item Response Theory Retrieved February 1, 2003, from ERIC Clearinghouse on Assessment and Evaluation Web site: http://edres.org/irt/baker

Beaujean, A. A. (2005). Using item response theory to assess the Lynn–Flynn effect. Unpublished doctoral dissertation, University of Missouri-Columbia.

Beaujean, A. A., & Guiling, S. F. (2006). The Lynn–Flynn Effect and school psychology: A call for research. *The School Psychologist*, *60*, 17−20.

Blair, C., Gamsonb, D., Thornec, S., & Bakerd, D. (2005). Rising mean IQ: Cognitive demand of mathematics education for young children, population exposure to formal schooling, and the neurobiology of the prefrontal cortex. *Intelligence*, *33*, 93−106.

Bock, R. D., Gibbons, R., & Muraki, E. (1988). Full-information factor analysis. *Applied Psychological Measurement*, *12*, 261−280.

Bolen, L. M., Aichinger, K. S., Hall, C. W., & Webster, R. E. (1995). A comparison of the performance of cognitively disabled children on the WISC-R and WISC-III. *Journal of Clinical Psychiatry*, *51*, 89−94.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge: Cambridge University.

Brand, C. R. (1996). *The g factor: General intelligence and its implications.* Chichester: Wiley.

Brand, C. R., Freshwater, S., & Dockrell, W. B. (1989). Has there been a "massive" rise in IQ levels in the west? *Irish Journal of Psychology*, *10*, 388−394.

Bureau of Labor Statistics, U.S. Department of Labor, & National Institute for Child Health and Human Development. (2002). Children of the NLSY79, 1979–2004. [computer file] Columbus, OH: Center for Human Resource Research.

Burt, C. (1952). *Intelligence and fertility*, (2nd ed.). London: Eugenics Society.

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items Thousand Oaks, CA: Sage.

Center for Human Resource Research. (2004). *NLSY79 child & young adult data users guide.* Columbus,OH: Author.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, (2nd ed.) Hillsdale, NJ: Erlbaum.

Colom, R., Juan-Espinosa, M., & Garcia, L. F. (2001). The secular increase in test scores is a "Jensen effect". *Personalilty and Individual Differences*, *30*, 553−559.

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart & Winston.

Daley, T. C., Whaley, S. E., 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.

Deary, I. J. (2000). *Looking down on human intelligence: From psychometrics to the brain.* Oxford: Oxford University Press.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence*, *35*, 13−21.

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test—Revised.* Bloomington, MN: Pearson Assessments.

Dunn, L. M., & Dunn, L. M. (1997). *Examiner's manual for the Peabody Picture Vocabulary Test*, (3rd ed.). Circle Pines, MN: American Guidance Service.

Dunn, L. M., & Markwardt, F. C. J. (1970). *Peabody Individual Achievement Test.* Circle Pines, MN: American Guidance Service.

Embretson, S., & Prenovost, K. (1999). Item Response Theory in assessment research. In P. C. Kendall, J. N. Butcher, & G. N. Holmbeck (Eds.), *Handbook of research methods in clinical psychology* (pp. 276−294). New York: JohnWiley & Sons.

Embretson, S., & Reese, S. (2002). Item Response Theory for psychologists Mahwah, NJ: Lawrence Erlbaum.

Eysenck, H. J., & Schoenthaler, S. J. (1997). Raising IQ level by vitamin and mineral supplementation. In R. J. Sternberg, & E. L. Grigorenko (Eds.), *Intelligence, heredity, and environment* (pp. 363−392). New York: Cambridge University Press.

Flieller, A. (1988). Application du modele de Rasch a un probleme de comparaison de generations [Applications of the Rasch model to a problem of intergenerational comparison]. *Bulletin de Psychologie*, *42*, 86−91.

Flynn, J. R. (1983). Now the great augmentation of the American IQ. *Nature*, *301*, 655.

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. (1999). Searching for justice: The discovery of IQ gains over time. *American Psychologist*, *54*, 5−20.

Flynn, J. R. (2006a). Tethering the elephant: Capital cases, IQ, and the Flynn Effect. *Psychology, Public Policy, and Law*, *12*, 170−189.

Flynn, J. R. (2006b). Efeito Flynn: Repensando a inteligencia e seus efeitos [The Flynn Effect: Rethinking intelligence and what affects it]. In C. Flores-Mendoza, & R. Colom (Eds.), *Introducao a psicologia das diferencas individuais [Introduction to the psychology of individual differences]* (pp. 387−411). Porto Alegre: Artmed.

Hambleton, R. K., & Swaminathan, H. (1985). *Item Response Theory: Principles and applications.* Boston: Kluwer.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of Item Response Theory Newbury Park, CA: Sage.
Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.
Kanaya, T., Scullin, M. H., & Ceci, S. J. (2003). The Flynn Effect and U. S. policies: The impact of rising IQ scores on American society via mental retardation diagnoses. *American Psychologist*, *58*, 778−790.
Lord, F. M. (1980). *Applications of Item Response Theory to practical testing problems.* Hillsdale, NJ: Lawrence Erlbaum.
Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental tests.* Reading, MA: Addison-Wesley.
Lynn, R. (1989). A nutrition theory of the secular increases in intelligence: Positive correlations between height, head size, and IQ. *British Journal of Educational Psychology*, *59*, 372−377.
Lynn, R. (1990). The role of nutrition in secular increases in intelligence. *Personality and Individual Differences*, *11*, 273−285.
Meade, A. W., & Lautenschlager, G. J. (2004). A comparison of Item Response Theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*, 361−388.
Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2005). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing*, *5*, 279−300.
Mingroni, M. A. (2004). The secular rise in IQ: Giving heterosis a closer look. *Intelligence*, *32*, 65−83.
Must, O., Must, A., & Raudik, V. (2003). The secular rise in IQs: In Estonia, the Flynn effect is not a Jensen effect. *Intelligence*, *31*, 461−471.
Nettelbeck, T., & Wilson, C. (2004). The Flynn Effect: Smarter not faster. *Intelligence*, *32*, 85−93.
Osterlind, S. J. (1983). Test item bias Newbury Park, CA: Sage.
Raju, N. S., Laffitte, L. J., & Byrne, B. M. (2002). Measurement equivalence: A comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*, 517−529.
Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and Item Response Theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552−566.
Rodgers, J. L. (1999). A critique of the Flynn Effect: Massive IQ gains, methodological artifacts, or both? *Intelligence*, *26*, 337−356.
Rodgers, J. L., & Wanstrom, L. (2007). Identification of a Flynn Effect in the NLSY: Moving from the center to the boundaries. *Intelligence*, *35*, 187−196.
Rushton, J. P. (1999). Secular gains in IQ not related to the *g* factor and inbreeding depression—Unlike Black–White differences: A reply to Flynn. *Personality and Individual Differences*, *26*, 381−389.
Sanborn, K. J., Truscott, S. D., Phelps, L., & McDougal, J. L. (2003). Does the Flynn Effect differ by IQ level in samples of students classified as learning disabled? *Journal of Psychoeducational Assessment*, *21*, 145−159.
Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology*, *86*, 162−173.
Scott, R., Bengston, H., & Gao, P. (1998). The Flynn Effect: Does it apply to academic achievement? *Mankind Quarterly*, *39*, 109−118.
Sijtsma, K., & Molenaar, I. (2002). *Introduction to Nonparametric Item Response Theory.* Thousaad Oaks, CA: Sage.
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.
te Nijenhuis, J., & van der Flier, H. (2006, December). Gains in g and empty gains. Paper presented at the annual meeting of the International Society for Intelligence Research, San Francisco, CA.
Teasdale, T. W., & Owen, D. R. (1989). Continuing secular increases in intelligence and a stable prevalence of high intelligence levels. *Intelligence*, *13*, 255−262.
Teasdale, T. W., & Owen, D. R. (2005). A long-term rise and recent decline in intelligence test performance: The Flynn Effect in reverse. *Personality and Individual Differences*, *39*, 837−843.
Thissen, D. (2001). IRTLRDIF v.2.0b: Software for the computation of the statistics involved in Item Response Theory likelihood-ratio tests for differential item functioning Chapel Hill, NC: LL. Thurstone Psychometric Laboratory Retrieved: http://www.unc.edu/~dthissen/dl.html[Computer software].
Truscott, S. D., & Frank, A. J. (2001). Does the Flynn Effect affect IQ scores of students classified as LD? *Journal of School Psychology*, *59*, 319−334.
van der Linden, W. J., & Hambleton, R. K. (Eds.). (1997). *Handbook of modern Item Response Theory.* New York: Springer.
Wicherts, J. M. (2005, December). Flynn effect in the Woodcock–Johnson Cognitive Ability and Achievement Tests 1976–1999. Paper presentation at the annual meeting of the International Society of Intelligence Research, Albuquerque, NM.
Wicherts, J. M., Dolan, C. V., Hessen, D., Oosterveld, P., Baal, G. C. M., van Boomsma, D. I., et al. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn Effect. *Intelligence*, *32*, 509−537.
Wood, R., Wilson, D., Gibbons, R., Schilling, S., Muraki, E., & Bock, D. (2004). TESTFACT. [Computer program] Lincolnwood, IL: Scientific Software International.
Yu, C. H., & Osborn Popp, S. E. (2005). Test equating by common items and common subjects: Concepts and applications.*Practical Assessment Research & Evaluation*, *10*(4) Retrieved May 30, 2005, from http://pareonline.net/getvn.asp?v=10&n=4.
Zimowski, M. F. (2003). Multiple-group analyses. In M. du Toit (Ed.), IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFAT (pp. 531–537) Lincolnwood, IL: Scientific Software International.
Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: Multiple-group IRT analysis and test maintenance for binary items. [Computer program] Lincolnwood, IL: Scientific Software International.