Personality and Individual Differences 48 (2010) 294–298

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Examining the Flynn Effect in the General Social Survey *Vocabulary* test using item response theory

A. Alexander Beaujean [a,*], Yanyan Sheng [b]

[a] Baylor Psychometric Laboratory, Baylor University, Department of Educational Psychology, One Bear Place #97301, Waco, TX 76798-7301, USA
[b] Southern Illinois University, Department of Educational Psychology & Special Education, Carbondale, IL 62901, USA

ARTICLE INFO

Article history:
Received 11 March 2009
Received in revised form 5 September 2009
Accepted 12 October 2009
Available online 11 November 2009

Keywords:
Flynn Effect
Item response theory
General Social Survey

ABSTRACT

Most studies of the Flynn Effect (FE) use classical test theory (CTT)-derived scores, such as summed raw scores. In doing so, they cannot test competing hypotheses about FE, such as it is caused by a real change in cognitive ability versus it is a change in the tests that measure cognitive ability. An alternative to CTT-derived scores is to use latent variable scores, such as those from item response theory (IRT). This study examined the FE on the *Vocabulary* test in the General Social Survey using IRT. The results indicate that while there has been a decrease–increase trend since the 1970s, the IRT-based scores never differed from the 1970s comparison point more than would be expected from random fluctuation. In contrast, while the CTT-derived summed scores showed the same decrease–increase pattern, all comparisons among the time points and the 1980s group were outside a 95% confidence interval. Multiple reasons for these results are discussed, with the conclusion being there is a need for more multiple-time point studies of the FE using IRT.

1. Introduction

The Flynn Effect (FE) (i.e., rise in IQ scores in the 20th century; Flynn, 1984, 1987) has been an active area of inquiry over the past three decades (Daley, Whaley, Sigman, Espinosa, & Neumann, 2003; Kanaya, Scullin, & Ceci, 2003; Sanborn, Truscott, Phelps, & McDougal, 2003; Sundet, Barlaug, & Torjussen, 2004). Those who think the FE represents real change in cognitive ability have made multiple attempts to explain this rise, ranging from nutritional changes (Lynn, 2009), to curricular changes (Blair, Gamsonb, Thornec, & Bakerd, 2005), to heterosis (outbreeding; Mingroni, 2004). However, others argue that the FE does not represent a real change in cognitive ability. Instead, the FE is the result of various psychometric artifacts (i.e., the tests' properties change over time, not the respondents; Brand, 1996; Wicherts et al., 2004). In actuality, the FE is likely a combination of multiple factors working concurrently converging (Jensen, 1998).

One common thread in most FE research is the reliance on scores derived from classical test theory (CTT) (for exceptions, see Beaujean & Osterlind, 2008; Flieller, 1988; Wicherts et al., 2004). CTT is concerned with the estimation of a "true score" and the resulting statistical analysis uses a function of the summed raw scores to estimate this true score (Crocker & Algina, 1986). Analyzing CTT-derived scores to study the FE is unfortunate for multiple reasons (Borsboom, 2005), the most cogent being they cannot differentiate between the two very distinct and important hypotheses (Chan, 1998): the FE is the result of an increase in cognitive ability versus the FE is the result of the change of cognitive ability tests over time.

In contrast to analyzing CTT-derived scores, latent variable analysis allows the investigator to differentiate between the manifest test scores and the trait(s) they are designed to measure. When the variables under investigation are individual test items (instead of summed scores), the latent variable model is called an item response theory (IRT) model. An IRT model specifies how an individual's (latent) trait level and a specific test item relate, as well as the item set where the individual item resides (Baker & Kim, 2004). Whereas CTT focuses on examinees' total test score, IRT focuses on both individual items and the examinees' (latent) trait score. This crucial difference allows for two very useful properties when examining the FE. First, IRT methods allow for non-equivalent groups equating (Zimowski, 2003). Consequently, even though groups may significantly differ on the trait a test is measuring, using an IRT model allows for the groups' scores to be equated onto the same scale. Second, in IRT models the item parameters are not dependent on the ability of the examinees responding to the items and the examinee's scores are not dependent on the specific test items. Thus, groups can differ widely on the trait a test is measuring, but the item parameters should be the same (within a linear transformation). So, if two groups of examinees take the same test at different time points and there is a significant change in the

* Corresponding author. Tel.: +1 254 710 1548; fax: + 254 710 3265.
E-mail address: Alex_Beaujean@Baylor.edu (A. Alexander Beaujean).

0191-8869/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.paid.2009.10.019

trait the items measure, the item parameters should not differ between samples after transforming them onto the same scale.

Sometimes, items work differently in one group than they do in another irrespective of the groups' trait distributions. When this occurs, it is called item non-invariance (Meredith, 1993) or differential item functioning (Holland & Wainer, 1993). However, even if a test has a number of items that exhibit non-invariance, IRT models can still estimate the examinee's trait(s) as long as enough of the test's items are invariant (Byrne, Shavelson, & Muthén, 1989).

The objective of the present study is to examine the FE with an IRT model. Using a large sample of adult respondents from the late 20th and early 21st century who all took the same test, we examined the average score over time using both CTT-derived scores and IRT-based scores.

2. Method

2.1. Item response theory models

IRT provides a fundamental framework in modeling the person–item interaction. Conventional IRT models assume that the probability of the *i*th respondent's dichotomous response (0/1) to the *j*th item (y_{ij}) takes the form[1]

$$P(y_{ij} = 1|\theta_i, \alpha_i, \delta_j) = \int_{-\infty}^{\alpha_j(\theta_i - \delta_j)} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-z^2}{2}\right] dz, \quad (1)$$

where δ_j denotes item difficulty, α_j is a positive scalar parameter describing item discrimination, and θ_i is a scalar latent trait parameter. Given the number of item parameters, this model is called the two-parameter normal ogive (2PNO) IRT model. If $\alpha = 1$, Eq. (1) becomes a one-parameter normal ogive (1PNO) IRT model. The model assumes one latent trait (θ) for each person (i.e., unidimensional), signifying that each test item measures some facet of the unified latent trait. These models share much in common with the one-factor model in the factor analytic (FA) framework (Kamata & Bauer, 2008). Specifically, the latent trait and two item parameters, α and δ, in IRT carry the same meaning as the terms common factor, factor *loading* and item *threshold*, respectively, in FA models (McDonald, 1985). Hence, to coalesce with other invariance literature, we discuss loading and threshold instead of the more traditional IRT terminologies discrimination and difficulty. For conversion formulae, see Kamata and Bauer (2008). It is noted that IRT is not limited to parametric models, such as that in (1), and that it includes non-parametric models as well (Mokken, 1971; Sijtsma & Molenaar, 2002). The latter, however, are not the focus of the study and thus were not considered.

2.2. Instrument

Data for this study came from the General Social Survey (GSS; Davis, Smith, & Marsden, 2007a). Part of the GSS is a *Vocabulary* test comprised of ten multiple choice items, which are labeled as items A through J in this article. The respondent can choose from five words (meanings) as his/her response, and the administrator scores each answer as correct or incorrect. The ten items were selected from the twenty-item Gallup-Thorndike Verbal Intelligence Test, Form A (Thorndike, 1942). These words are notable for their variety, ranging from verbs to nouns and covering topics ranging from psychiatry to musicology. The original test was administered to 538 students in grades 7, 8, and 9; 456 students in grades 10 and 11; and 268 entering college freshmen. Based on the correlations between the two parallel forms for a "cross-section" of the evaluated adult group, Thorndike (1942) "estimated that for such a group the correlations between two forms of the test would be .83, and the correlation of the test with a perfect criterion would be .90" (p. 132). For the sample in the current study, the test's internal consistency (Cronbach's α) was .68.

While vocabulary knowledge is not synonymous with intelligence, the relationship between the two variables is very strong (Jensen, 2001). For example, Carroll (2003, pp. 11, 16) reports factor loadings of .60 and .75 for Picture Vocabulary and Oral Vocabulary, respectively, on the general intelligence factor using the Woodcock–Johnson Psychoeducational Battery-Revised. Sattler (2008) reports similar findings for the Vocabulary subtests of the Wechsler Adult Intelligence Scale-Third Edition and the Stanford-Binet-Fifth Edition.

Table 1
Descriptive statistics for Vocabulary respondents.

	All years	1970s	1980s	1990s	2000s
Female (%)	56.64	55.64	57.28	57.36	55.56
Caucasian (%)	83.07	89.10	85.30	82.17	76.63
Black (%)	12.49	10.19	11.56	13.14	14.57
Average age[1]	45.52	44.63	44.98	45.67	46.70
Highest education	12.76	11.81	12.39	13.15	13.43

[1] All persons over 89 years of age were coded as 89 years old.

2.3. Data

The GSS has been regularly administered to American adult household members of all ages since the early 1970s (Davis, Smith, & Marsden, 2007b). The years of the GSS used for this study were 1972 through 2008, but the *Vocabulary* test has not been administered every year and in some years was only administered to a random subset of the respondents (Malhotra, Krosnick, & Haertel, 2007). There were 25,555 participants in those years given the opportunity to answer the Vocabulary items. Items were coded as 1 if it was answered correctly and 0 if incorrect or the respondent chose not to answer the question. The groups were then combined by decade, giving four decade groups: 1970s ($n = 4, 515$), 1980s ($n = 7, 146$), 1990s ($n = 8, 356$), and 2000s ($n = 5, 538$).[2] Table 1 gives descriptive statistics for the sample.

2.4. Assessing measurement invariance

Assessing measurement invariance is a multistep procedure (Bontempo & Hofer, 2007). We assessed measurement invariance using the following procedures:

1. Assess dimensionality of GSS *Vocabulary* test.
2. Assess fit of IRT model.
3. Assess for invariance loadings and thresholds. If full invariance is not tenable, assess for partial invariance.
4. If full or partial invariance exists, compare average latent scores among groups.

[1] The model can be made more complicated by incorporating a pseudo-chance-level parameter (Lord, 1980). However, difficulty arises in fitting such models (Embretson & Reise, 2000), especially when examining invariance. When such models were attempted in the current study (under the fully Bayesian framework), there were difficulties with model convergence. Nonetheless, the obtained θ estimates between the 2 and 3 parameter models were practically identical ($r > .99$). Thus, a pseudo-chance-level parameterization was not considered in the study.

[2] African-Americans were over-sampled in some years in the 1980s. In addition, after 2004 the GSS adopted a new design to account for non-respondents and subsampling. Consequently, respondents were weighted by the *Oversamp* and *Wtssnr* variables, and *Sampcode* was used as the clustering variable.

3. Results

Except where indicated, all analyses were executed in Mplus (version 5) using its complex survey features and traditional IRT constraints (Kamata & Bauer, 2008, p. 139). The exploratory factor analysis (EFA) used item-level information with a robust maximum likelihood extraction (Muthén & Muthén, 1998–2007, p. 484). The confirmatory factor analysis (and subsequent multigroup analysis) used the WLSMV estimator (Flora & Curran, 2004) and the Delta parameterization as Muthén and Muthén (1998–2007, p. 485) recommend. The results are similar to fitting a normal ogive model with traditional IRT software.

3.1. Step 1: Assess dimensionality of Vocabulary test

To assess the *Vocabulary* test's dimensionality, the items were factor analyzed separately for each decade group, extracting 1–3 factors. The item loadings from the three-factor solution did not form an interpretable pattern.[3] For the two-factor solution, the only discernible pattern was item difficulty: one factor consisted of easier words (C, G, H, and J) and the other consisted of more difficult words (A, B, D, E, F, and I). However, it is not uncommon to have difficulty related factors when analyzing item-level data (Bernstein & Teng, 1989) and this pattern has been found in other analyses of the *Vocabulary* test (Malhotra et al., 2007). Moreover, the factor correlation was positive and large ($.60 < r < .70$) in each decade group. Consequently, we treated the test as being unidimensional, which justifies the use of an IRT model with one latent trait for the subsequent analyses.

3.2. Step 2: Assess fit of item response theory model

For each decade group, we assessed the fit of a 2PNO model versus a 1PNO model using the *ltm* package (Rizopoulos, 2006). Using unweighted and unclustered data, the resulting item χ^2 fit statistics (du Toit, 2003, pp. 29–30) for the two-parameter model were small for all items (<1.1) in all groups, which indicates this model fits the data well. When a one-parameter model was subsequently fit, there was a statistically significant ($p < .05$) increase in the log likelihood values, indicating the two-parameter model fit the data better.

3.3. Step 3: Assessing invariance

We assessed item invariance following the steps outlined in Glöckner-Rist and Hoijtink (2003). For all analyses, all groups' item scale factors were constrained to unity and the 1970s group latent trait mean was constrained to zero to make it the reference group. The results from all invariance analyses are given in Table 2.

There is not a well developed literature indicating what fit indices work best when fitting or testing invariance in IRT models. Consequently, we used the traditional (a) χ^2, (b) Root Mean Square Error of Approximation (RMSEA), (c) Comparative Fit Index (CFI), and (d) Tucker-Lewis Index Values (TLI). For the former two, lower values indicate better fit, while higher values indicate better fit for the latter two. As we are using the WLSMV estimator, the traditional change in χ^2 index cannot be used; however, Mplus can compare nested models and outputs a χ^2 test for testing model differences.

For the baseline model (model 1), the item thresholds and loadings were constrained to be equal and the latent variances were constrained to unity across all groups; the latent means were estimated in all but the reference group. The fit indices' values indicated the model fit relatively well; moreover, the modification indices values did not indicate that freeing any item thresholds or loadings would help model fit. The latent variances were then freed (model 2), but the resulting fit indices indicated this model fit the data worse than the original model, so was not kept. The parameter estimates for the final model are given in Table 3.

Table 2
Model statistics.

Model	$\chi 2$	*df*	CFI	TLI	RMSEA
1	798.75	43	0.978	0.978	0.052
2	1703.54	81	0.952	0.975	0.056

Note: Model 1 constrained all items' thresholds and loadings to be equal. All variances were set at one.
Model 2 freed the variances for all models. CFI, Comparative Fit Index; TLI, Tucker-Lewis Index; RMSEA, Root Mean Square Error of Approximation. Three numbers after decimal place were used to indicate where difference in values was located.

Table 3
Item statistics for final invariance model.

	Classical		Latent variable				IRT	
Item	δ	α	λ	(s.e.)	τ	(s.e.)	δ	α
A	0.78	0.57	0.67	(0.01)	−0.78	(0.02)	−1.17	0.89
B	0.87	0.64	0.93	(0.01)	−1.18	(0.03)	−1.27	2.50
C	0.21	0.50	0.63	(0.01)	0.79	(0.02)	1.26	0.81
D	0.88	0.63	0.92	(0.01)	−1.21	(0.03)	−1.31	2.42
E	0.70	0.68	0.81	(0.01)	−0.54	(0.03)	−0.67	1.37
F	0.74	0.68	0.82	(0.01)	−0.68	(0.03)	−0.83	1.45
G	0.32	0.55	0.60	(0.01)	0.45	(0.02)	0.74	0.76
H	0.28	0.57	0.67	(0.01)	0.55	(0.02)	0.82	0.91
I	0.73	0.58	0.67	(0.01)	−0.62	(0.02)	−0.92	0.91
J	0.23	0.55	0.70	(0.01)	0.73	(0.02)	1.04	0.97

Note: Latent variables *t* constraining all scale factors to unity and using Mplus' delta parameterization.
λ, item loading; τ, threshold; δ, difficulty; α, discrimination; s.e., standard error. Classical item difficulties and discriminations were percent correct and point-biserial correlations between individual item responses and summed scores, respectively.

3.4. Step 4: Compare latent means

Since all of the *Vocabulary* items showed invariance across time groups, it is possible to compare the average latent trait scores among the four decade groups. As another point of comparison, we also compared the average summed *Vocabulary* scores across groups. The results are given numerically in Table 4 and graphically in Fig. 1. To facilitate the comparison between the IRT and CTT scores, we subtracted the average summed scores of the 1970s group from each of those of other decades. This makes the 1970s group the reference group for both scales, and the other groups' scores measure how far they deviate from the 1970s group average. In addition, we computed Cohen's (1988) *d* for each group comparison.

For both types of scores the overall trend is the same. The average score decreases from the 1970s group to the 1980s group, then begins to rise for the 1990s and 2000s groups, but never above the level of the 1970s group. The magnitude of the mean differences between the summed and IRT scores somewhat differs, which is highlighted by Fig. 1. For the summed scores, the 95% confidence interval (CI) for the 1980s group does not encompass the CIs for the other three groups. This does not appear to be the case for the latent variable scores. While it is impossible to estimate all groups' means and standard errors simultaneously, the standard error for the latent variable means is approximately .04, irrespective of what group is used as the reference group. This is the same

[3] The National Opinion Research Center allowed the first author to examine the item stems for the purpose of this study.

Table 4
Average Vocabulary score comparison by decade group.

Summed Vocabulary score				
	1970s	1980s	1990s	2000s
$\bar{x}$	0.00	−0.22	−0.09	−0.02
sd	2.43	2.54	2.49	2.39
$d(t-1)$		−0.09	−0.04	−0.01
$d(t-2)$			0.05	0.08
$d(t-3)$				−0.03
Latent Variable score				
	1970s	1980s	1990s	2000s
$\bar{x}$	0.00	−0.09	−0.03	−0.01
sd	1.00	1.00	1.00	1.00
$d(t-1)$		−0.09	−0.03	−0.01
$d(t-2)$			0.06	0.08
$d(t-3)$				−0.02

Note: $d(t-i)$: Cohen's (1988) *d* for time points of distance *i* decades. For example, for the summed score, under the 2000s column, $d(t-3) = -.03$ indicates there was a decrease in the measured trait of 0.03 standard deviations from the 1970s group. Summed score mean of 1970s group was 5.82.

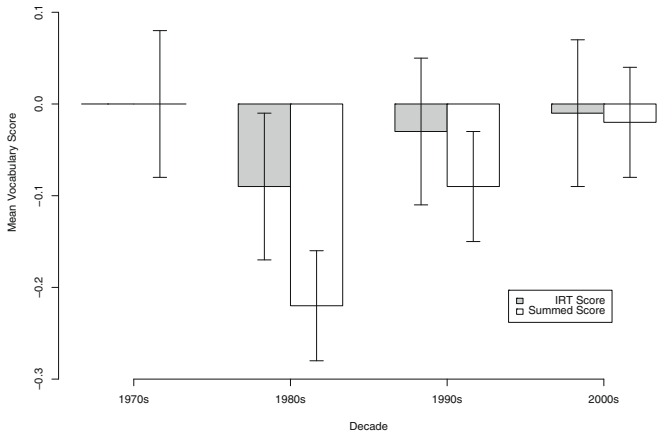

Fig. 1. Average *Vocabulary* score (with 95% confidence interval) across decade groups.

as the summed score standard error for the 1970s group, so the error bars for the CTT score can be used to approximate a 95% CI for the IRT score. Thus, the 95% IRT CIs for the 1980s, 1990s, and 2000s groups all encompass that of the 1970s group.

When examining the results in Table 4, the absolute group differences are smaller for the IRT scores, as would be expected given the dispersion of means in Fig. 1. However, the effect sizes are somewhat larger in the IRT metric. The seemingly contradiction in results is due to constraining the latent variable's variance to be one across all groups. Consequently, the standard deviation of the summed score, for all groups, is approximately twice as large as that for the latent variable, which shrinks the *d* statistic.

4. Discussion

The study examined the Flynn Effect (FE) in the General Social Survey's (GSS) *Vocabulary* test using respondents from 1972 to 2008. While other FE research has found a consistent increase in IQs throughout most of the 20th century (Flynn, 2007), the current data showed a decline in the 1980s and, although it was followed by a steady increase throughout the 1990s and 2000s, it never rose above the 1970s average. Using the IRT scores, all the years' confidence intervals (CI) overlapped, indicating "non-significant" change in the latent vocabulary trait. As a point of comparison, the *Vocabulary* items were summed across the items to create a total test score, as is typically done in FE studies. The results from the summed score showed the same pattern as the IRT scores, but the magnitude of the score differences was larger. Specifically, none of the CIs for 1970s, 1980s, or 2000s overlapped the CI for the 1980s.

The results from the current study could be explained in multiple ways. First, vocabulary knowledge and general intelligence are not synonymous, and much FE research has shown the magnitude of change is smaller for more crystallized tasks (e.g., vocabulary) than for more fluid tasks (e.g., deductive reasoning) (Flynn, 2007). However, when examining normative data on the Wechsler Adult Intelligence Scale (WAIS) scales by subtest, Flynn (2009, p. 102) reported that the Vocabulary subtest showed the second largest increases over time (3.4 points). Nonetheless, the comparison between the WAIS-Revised and WAIS-Third Edition (1978 and 1995, respectively) showed the smallest increase (0.6 points), which mirrors the time period of the decrease in the GSS *Vocabulary* scores. Moreover, Herrnstein and Murray (1996) show that the SAT Verbal scores displayed a 50-point decline between the mid-1960s and 1980. As the average person who takes the SAT is 16-18 years old and the average person in the GSS sample is approximately 45, it is likely that many of the people who took the SAT between 1965 and 1980 were included in the GSS after the 1970s. Thus, perhaps there was something unique in this particular American cohort.

An alternative explanation is that the FE is just part of the *Vocabulary* test's random fluctuation in scores. That is, in the current study the difference in latent variable means were all within two standard errors of each other. So, while there may appear to be a FE in vocabulary knowledge between the 1980s and 2000s of 1.20 IQ units, when taken into a larger perspective, the gains in the 1990s and 2000s do not counter balance the decrease in the 1980s. Thus, the average in the 2000s group is still lower than that of the 1970s.

A third alternative is that the *Vocabulary* score trends are due to the race and/or sex composition of the sample. Both hypotheses are unlikely though. First, there is less than a 2% difference in the male–female ratio across decades (see Table 1) and the pattern of these small male–female differences do not directly follow the pattern of change in mean scores. Second, while there has been a monotonic increase in African-Americans sampled and decrease in Caucasians sampled in the GSS, the score differences are not consistent, first going down and then up.

5. Implications

The first implication is that gathering two points when studying the FE may not give enough information about the effect. Rather, perhaps multiple-time points are needed in order to be able to examine if there are both increases and decreases in the measured trait across time. The second implication of this study is that IRT scores will not necessarily give the same results as an analysis of CTT-derived scores. Although the items used in the current study showed invariance across all groups, the IRT and CTT results still differed: all CIs for the average IRT scores overlapped, but CI for 1980s group's average summed score was "significantly" different from those of other decade groups. Consequently, while the pattern was the same for both types of scores, the inference made from one would not necessarily be the same as for the other. As much has been written on the general benefits of using latent variables (Borsboom, 2005), it might be the case that the IRT scores could offer more accurate appraisal of the FE than most CTT-based scores.

6. Limitations

The study's main limitation was the instrument used. Future studies need to look at gathering item-level data from instruments

that measure a wider array of cognitive skills than just knowledge of vocabulary. However, if one wants to use item-level data, there are a limited number of data sets available. Most large assessment companies deny access to item scores, and most small research projects do not have the sample size and national representation typical of a test standardization or survey questionnaire. Consequently, until more item-level data become available, most investigators are stuck between having a strong sampling frame or having a strong data set.

References

Baker, F. B., & Kim, S.-H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York: Mercel Dekker.

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 Young Adults data. *Intelligence, 36*, 455–463.

Bernstein, I. H., & Teng, G. (1989). Factoring items and factoring scales are different: Evidence for multidimensionality due to item categorization. *Psychological Bulletin, 105*, 465–477.

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.

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.

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.

Byrne, B. M., Shavelson, R. J., & Muthén, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*, 456–466.

Carroll, J. B. (2003). The higher-stratum structure of cognitive abilities: Current evidence supports g and about ten broad factors. In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 5–21). New York: Pergamon.

Chan, D. (1998). The conceptualization and analysis of change over time: An integrative approach incorporating longitudinal mean and covariance structures analysis (LMACS) and multiple indicator latent growth modeling (MLGM). *Organizational Research Methods, 1*, 421–483.

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

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.

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.

Davis, J. A., Smith, T. W., & Marsden, P. V. (2007a). *General social surveys, 1972–2008*. [machinereadable data file] Chicago: National Opinion Research Center [producer]. Storrs, CT: The Roper Center for Public Opinion Research, University of Connecticut [distributor].

Davis, J. A., Smith, T. W., & Marsden, P. V. (2007b). *General social surveys, 1972–2006: Cumulative codebook*. Chicago: National Opinion Research Center (National Data Program for the Social Sciences Series, no. 18).

du Toit, M. (2003). *IRT from SSI*. Lincolnwood, IL: Scientific Software International.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

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.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466–491.

Flynn, J. R. (1984). The mean IQ of Americans: Massive gains 1932–1978. *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. (2007). *What is intelligence?: Beyond the Flynn Effect*. New York: Cambridge University.

Flynn, J. R. (2009). The WAIS-III and WAIS-IV: Daubert motions favor the certainly false over the approximately true. *Applied Neuropsychology, 16*, 98–104.

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.

Herrnstein, R. J., & Murray, C. (1996). *Bell curve: Intelligence and class structure in American life*. New York: Free Press.

Holland, P., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

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

Jensen, A. R. (2001). Vocabulary and general intelligence. *Behavioral and Brain Science, 24*, 1109–1110.

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.

Kanaya, T., Scullin, M. H., & Ceci, S. J. (2003). The Flynn Effect and US 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.

Lynn, R. (2009). What has caused the Flynn Effect? Secular increases in the Development Quotients of infants. *Intelligence, 37*, 16–24.

Malhotra, N., Krosnick, J. A., & Haertel, E. (2007). *The psychometric properties of the GSS wordsum vocabulary test (methodological report no. 111)*. Chicago: National Opinion Research Center.

McDonald, R. P. (1985). *Factor analysis and related methods*. Hillside, NJ: Erlbaum.

Meredith, W. (1993). Measurement invariance, factor analysis and factorial invariance. *Psychometrika, 58*, 525–543.

Mingroni, M. A. (2004). The secular rise in IQ: Giving heterosis a closer look. *Intelligence, 32*, 65–83.

Mokken, R. J. (1971). *A theory and procedure of scale analysis*. Berlin: De Gruyter.

Muthén, L. K., & Muthén, B. O. (1998–2007). *Mplus user's guide*. Los Angeles, CA: Author.

Rizopoulos, D. (2006). Ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software, 17*, 1–25.

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 Psycheducational Assessment, 21*, 145–159.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). La Mesa, CA: Author.

Sijtsma, K., & Molenaar, I. W. (2002). *Introduction to nonparametric item response theory*. Thousand 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.

Thorndike, R. L. (1942). Two screening tests of verbal intelligence. *Journal of Applied Psychology, 26*, 128–135.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., 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.

Zimowski, M. F. (2003). Multiple-group analyses. In M. du Toit (Ed.), *IRT from SSI* (pp. 531–537). Lincolnwood, IL: Lincoln.