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.