USING ITEM RESPONSE THEORY TO ASSESS THE LYNN-FLYNN EFFECT

A Dissertation
presented to
the Faculty of the Graduate School
University of Missouri-Columbia

In Partial Fulfillment
Of the Requirements for the Degree
Doctor of Philosophy

---

by
A. ALEXANDER BEAUJEAN

Dr. Steven J. Osterlind (Educational Psychology)
Dr. Rick J. Short (School Psychology),
Dissertation Supervisors

JULY 2005 (Educational Psychology)
JULY 2006 (School Psychology)

ACKNOWLEDGEMENTS

I would like to thank my parents, William and Lela Beaujean, for their 27 years of support, help, and kindness. I would also like to thank the Educational, School, and Counseling Psychology Department for their unwavering support over the last five years; more specifically, I would like to thank the School Psychology and Educational Psychology programs for allowing me to pursue my academic interests unfettered. In addition, I would like to thank my committee members for their input on the design of this project and their suggestions on how to further the research in this area. I would like to thank Andrew J. Knoop for his comments and help with the proofreading, and Richard Morey for his help in setting up the style sheets in LaTeX.

This dissertation owes much of it content to Craig L. Frisby, who first exposed me to the beauty of differential psychology and data-based intelligence research, and to Steven J. Osterlind, who not only mentored and advised me over my graduate career, but fostered in me a love of quantitatively-oriented psychology.
Soli Deo Gloria.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

USING ITEM RESPONSE THEORY TO ASSESS THE LYNN-FLYNN EFFECT

A. ALEXANDER BEAUJEAN

Dr. Steven J. Osterlind (Educational Psychology)
Dr. Rick J. Short (School Psychology),
Dissertation Supervisors

ABSTRACT

This study examined the utility of using Item Response Theory (IRT) to assess the Lynn-Flynn Effect (LFE), using data from two different studies. The first study used data from a simulation experiment, where samples were generated that mimicked both real increases in cognitive abilities and psychometric artifacts. The results indicated that IRT methods were more effective than methods from Classical Test Theory (CTT) in distinguishing between a real increase in cognitive abilities and pure psychometric artifacts. The second study used data from the Mathematics section of the College Basic Academic Subjects Examination to demonstrate the use of IRT methods in assessing the LFE. This second study showed that from 1996 to 2001 there was a reverse LFE in the examinees' abilities, with ability decreasing approximately .222 standard deviations.

Chapter 1

INTRODUCTION

The Lynn-Flynn Effect (LFE) (i.e., the continued rise of psychometric IQ test scores over time) has been a source of consternation for those doing intelligence research since it was discovered in the early 1980's.[1] The reason is that while psychometric IQ scores have shown a steady increase for (at least) 70 years (since the 1940s) (Flynn, 1984; Lynn & Hampson, 1986), there has not been an appreciable increase in academic achievement (Hunt, 1995), physiological markers of cognitive ability (e.g., reaction time; Nettelbeck & Wilson, 2004), or Spearman's (1904) *g* (Must, Must, & Raudik, 2003; Rushton, 1999). Moreover, when surveyed, teachers in Western countries do not perceive that that there is general rise in student intelligence (Cocodia et al., 2003; Howard, 2001). In addition, while it may seem plausible that the LFE would be greatest for "school-based" knowledge, such as vocabulary tests, the opposite is true. It is the more abstract, nonverbal markers of cognitive ability, such as the constructs measured by Raven's Matrices, that show the largest gains.

The vast majority, if not all, of the studies examining the LFE have used methods derived from classical test theory (CTT), namely statistical comparisons of summed raw scores or factor scores (e.g., the Wechsler FSIQ). This is unfortunate as more modern psychometric techniques, such as those derived from latent trait models (e.g., item response theory [IRT]), are better able to provide data for the questions LFE scholars seek to answer. For example, if intelligence is actually rising, then the latent trait of general intelligence (Spearman, 1904) should show an increase. On the other hand, if the LFE is simply due to the population becoming more test savvy, and not a real increase in intelligence, then cognitive ability test items should show the same items functioning differently across time rather than a change in a latent (cognitive) ability. Analysis using CTT-derived methods are simply unable to answer these questions. (For a much more

elaborate comparison between the CTT and IRT methods and results, see Embretson & Reese, 2002).

Consequently, *the purpose of this manuscript is to demonstrate the use of IRT methods in determining if the increase in mean IQ scores across time (i.e., the LFE) is due to an increase in intelligence, or if it is due to a psychometric artifact.*[2] First, this manuscript will compare the efficacy of IRT and CTT models using purposefully simulated data samples (i.e., data with known distribution parameters) that mimic both a real rise in intelligence as well as artifacts that imitate a psychometric IQ rise. Then, this manuscript will demonstrate the use of IRT in assessing the existence of the LFE using cohort data from the College Basic Academic Subjects Examination (CBASE).

This manuscript's literature review will first give a brief review of the literature concerning the LFE. As this paper is not particulary concerned with possible theories to explain the LFE, *per se*, the review of the LFE will be somewhat cursory, although extensive enough to both explain and validate the LFE. Next, attention will shift to giving a brief overview of CTT, then show how properties of IRT can better answer questions pertinent to the LFE. Last, the review will define various methods for detecting differential item functioning (defined later) as well as define test equating—both of which are necessary precursors before one can delve further into the LFE via IRT methods.

Chapter 2

LITERATURE REVIEW

Lynn-Flynn Effect

The purpose of the section is merely to introduce the Lynn-Flynn Effect (LFE).[3] While representative, it in no way is intended to be exhaustive, as other sources have already provided such a service (e.g., Flynn, 1987; Lynn, 1997; Neisser, 1998). The goal is to familiarize those who do not know much about the LFE as well as expand upon the underlying motivation for the current manuscript.

Defined, the LFE is the continued rise of psychometric IQ test scores (approximately .3 IQ points/year), an effect seen in many parts of the world, although at greatly varying rates.

> ... great care is taken to ensure that the standardization samples of [IQ] tests are representative of the total population. Therefore, if the same group of subjects does better on an old test than a new one, the obvious explanation is that old norms are easier to exceed than more recent ones, which is to say that older standardization samples did not perform as well on IQ tests as more recent samples. (Flynn, 1983, p. 655)

The LFE is named after British differential psychologist Richard Lynn and New Zealand political scientist James R. Flynn, who independently re-discovered the effect in the early 1980's. Richard Lynn (Lynn, 1982; Lynn & Hampson, 1986) published data about the effect in Great Britain and Japan, while James Flynn (1983, 1984, 1999) published data about it in the United States.

Since the original studies were published, the eponymous LFE has been studied in many different populations (even including those with cognitive disorders; Bolen, Aichinger, Hall, & Webster, 1995; Sanborn, Truscott, Phelps, & McDougal, 2003; Truscott & Frank, 2001), both in developed nations and undeveloped countries (Daley, Whaley, Sigman, Espinosa, & Neumann, 2003; Rushton & Jensen, 2003; Sundet, Barlaug, &

Torjussen, 2004) (see Table 1 for a sample of studies). In the 20+ years research has been done in this field, the findings have been enigmatic. While multiple sources have found that psychometric IQ has been rising, general intelligence (g; Spearman, 1904) has not increased (Jensen, 1998; Kane & Oakland, 2000; Must et al., 2003), nor have reaction times (an endophenotype of intelligence; Jensen, 1998) decreased (Nettelbeck & Wilson, 2004), although head size, another endophenotype of intelligence, has increased (Storfer, 1999). In addition, although LFE appears to effect the entire range of the IQ distribution, there does appear to be a definite concentration among those at lower end (Colom, Lluis-Font, & Andres-Pueyo, 2005; Teasdale & Owen, 1989). Another aspect of the LFE that has puzzled researchers is that although there are mean increases in average psychometric IQ scores, ethnic group differences on the same IQ tests have *not* diminished (Murray, 1999; Jensen, 1998; Rushton, 1999, 2003). In fact, the one standard deviation difference between Black and White test takers is as pervasive today as it ever was (Rushton & Jensen, 2003, 2005; but also see Ceci, Rosenblum, & Kumpf, 1998).

Attempts to explain the various findings involved in the LFE has lead scholars down many different avenues of inquiry. Some, such as Lynn (1989, 1990), and Eysenck and Schoenthaler (1997) posit that massive environmental changes, such as changes in available nutrition have, at least in part, been responsible for the IQ increase. Lynn (1990), writing on the LFE in Japan, said that it

> can be easily explained as almost every factor known to influence IQ can be brought into play. Since 1930, Japan has experienced massive urbanization, a cultural revolution from feudal towards western attitudes, the decline of inbreeding and consanguineous marriages, huge advances in nutrition, life expectancy and education. (p. 655)

Others, such as Blair, Gamsonb, Thornec, and Bakerd (2005), explain the mean IQ increase as an artifact of educational curriculum changes, especially with math. Some theorists go the other direction, and posit that the change in IQ scores is not due to an environmental effect *per se*, but rather is a byproduct of an increase in heterosis (outbreeding) (Mingroni, 2004). While others posit that LFE, while extant, is not too much more than a psychometric artifact (Brand, 1996; Burt, 1952; Rodgers, 1999), or that

perhaps the LFE no longer is even operative, at least in certain countries (Sundet et al., 2004; Teasdale & Owen, in press).

The purpose of this paper is not to theorize about the many implications of the LFE, or even why it exists; rather, its purpose is to show how IRT methods can be used to better assess it. Consequently, it will be beneficial to examine the two designs researchers currently use. The first method is to give the same test to two different samples that are the same in (almost) all respects except the year of test administration, and then compare the same derived score from each cohort. An example would be to give a group of 2nd graders the Wechsler Intelligence Scale for Children (WISC) in 1990 and then, in 2000, give another group of 2nd graders the same WISC test. The second way to assess the LFE is to give the same sample two tests that were standardized at different times, and then derive the same (or comparable) scores from them. An example would be if both the Woodcock-Johnson-R and Woodcock-Johnson-III were given to the same group of college students, at (approximately) the same time, and the same two index scores (e.g., General Intellectual Ability) were compared.

While the two designs are appropriate, the current analysis methods are not. The vast majority, if not all, of studies examining the LFE have used methods derived from classical test theory (CTT), namely statistical comparisons (e.g., *t*-test, ANOVA) of summed raw scores or factor scores (e.g., the Wechsler FSIQ). These methods do not extract all the information within the data, a problem that more modern test theory (e.g., IRT) can rectify. For example, if intelligence is actually rising, then an amalgamated full-scale score is not the best variable to assess. The reason is that the increase could be due to a number of possibilities, such as a systemic increase in test "savvyness" either evidenced by allowing the examinee to guess at a correct answer more often or by having the same items lose their level of difficulty. IRT, on the other hand, can specifically assess these properties for all the items on a given test, thus putting the researcher in a position of being able to discern whether the population has become more test savvy without an appreciable increase in intelligence, whether there has been an appreciable rise in intelligence, or a host of other alternatives that analyses using CTT are simply unable to answer.

Models in Measurement

> A measurement theory in psychology must provide a rationale for relating behaviors to the psychological construct. (Embretson & Reese, 2002, p. 41)

Embretson and Reese (2002) write that psychological constructs (e.g., intelligence, personality) are often thought of as latent (i.e., unobservable) variables that underlie observable behavior. Measurements or, more specifically, item responses and test scores, are "indicator[s] of a person's standing on the latent variable, but it does not completely define the latent variable" (p. 40). In other words, the scores on latent variables are inferred, via some model, from some manifest behavior.

In psychology, a model is usually conceptualized in one of two ways. The first is akin to a theory, whereby various variables are purported to impact other variables (e.g., Circumplex model of family systems; Olson, Russell, & Sprenkle, 1989). Not unrelated, another definition of the word model is a mathematical one, whereby independent variables are combined through various formulae to optimally predict dependent variables (Lord & Novick, 1968). Embretson and Reese (2002) delineate three specific features that define the mathematical model: (a) the model defines the scale for the dependent variable(s), (b) the model specifies the independent variables, and (c) the model specifies how the independent variables are combined numerically to predict the dependent variables—these numerical combinations are the model's parameters. In psychological and educational measurement, there are two overlapping mathematical models: those from *classical test theory* (CTT) and those from *item response theory* (IRT). This text will first give an overview of CTT and then show how IRT extends CTT principles and why IRT models are more appropriate for the study of the LFE.

Overview of Classical Test Theory

In a CTT model, the dependent variable is a given person's observed test score, X_i, while the independent variables are the person's "true" score, T_i, and measurement error, ε. The independent variables combine additively and directly (i.e., there are no other coefficients attached) to produce the CTT model:

$$X_i = T_i + \varepsilon \tag{1}$$

where $\varepsilon \sim \mathrm{N}(0, \sigma_\varepsilon^2)$. Consequently,

$$E[X_i] = T_i \tag{2}$$

and

$$E[\varepsilon_i] = 0. \tag{3}$$

In addition, there are some (weak) assumptions about this model:

$$\mathrm{COV}[\varepsilon_i, T_i] = 0, \tag{4a}$$

$$\mathrm{COV}[X_i, X_j] = 0, \underset{i \neq j}{} \tag{4b}$$

and that one can build strictly parallel tests (Lord, 1980). While, (4a) and (4b) can be disproved with certain data structures via path analysis, they generally hold.

From (1), it follows that X_i's variance is made up of two parts:

$$\mathrm{VAR}[X_i] = \mathrm{VAR}[T_i + \varepsilon] = \mathrm{VAR}[T_i] + \mathrm{VAR}[\varepsilon]. \tag{5}$$

Likewise, from (1) and its relationship properties, it follows that

$$COV[X_i, T_i] = VAR[T_i].^4 \tag{6}$$

From (1), (5) & (6), *test reliability*, r_{xx}, can now be defined as:

$$r_{xx} = \rho^2_{X_i T_i} \equiv \frac{\mathrm{COV}^2[X_i, T_I]}{\mathrm{VAR}[T_i]\mathrm{VAR}[X_i]} = \frac{\mathrm{VAR}[T_i]}{\mathrm{VAR}[X_i]} = 1 - \frac{\mathrm{VAR}[\varepsilon]}{\mathrm{VAR}[X_i]} \tag{7}$$

Thus, it is the true score, T_i, not the observed score, that is of real interest in CTT. Unfortunately, T_i cannot be directly observed, but one can make inferences about it.[5]

Lord (1980) writes,

> The true score [T_i] is a mathematical abstraction. A statistician doing an analysis of variance components does not try to define the model parameters

> as if they actually existed in the real world [e.g., can be directly measured]. A statistical model is chosen, expressed in mathematical terms undefined in the world ... It is neither necessary nor appropriate to define a person's true score ... by real world operational procedures. (pp. 6-7)

Fortunately, there is an alternative that allows for an observable result, using the idea of a *parallel test*. By definition, two test forms, X_i and X_i', parallel if and only if,

$$E[X_i] = E[X_i'] = T_i \tag{8a}$$

and

$$\text{VAR}[X_i] = VAR[X_i'] \tag{8b}$$

Consequently, the correlation between the two parallel tests, $\text{COR}[X_i, X_i']$, is simply

$$\text{COR}[X_i, X_i'] = \frac{\text{VAR}[T_i]}{\sqrt{\text{VAR}[X_i] \times \text{VAR}[X_i']}} = \frac{\text{VAR}[T_i]}{X_i} = 1 - \frac{\text{VAR}[\varepsilon_i]}{\text{VAR}[X_i]} = \rho^2_{X_i,T_i} \tag{9}$$

Obviously, the importance of (9) is that it is estimable from actual data. Unfortunately, the time and expense involved in developing parallel forms of a test often precludes it from being done. Instead, a test is usually divided into two sections, and the two sections are considered parallel to each other. Of course, the two halves are not really parallel, so their correlation needs a correction factor. Interestingly, both Brown (1910) and Spearman (1910), independently, came up with this correction. So, for a test, X_i, containing n items, split into 2 parts, V_1 and V_2, each containing $\frac{n}{2}$ items, the reliability coefficient, is

$$\text{COR}[X_i, X_i'] = \frac{2 * \text{COR}[V_1, V_2]}{1 + \text{COR}[V_1, V_2]}.^6 \tag{10}$$

This conception of test reliability then brings up the question of how to split the test. From probability theory, a test with n items, split into two groups of $n/2$ items yields $\frac{n!}{2[(\frac{n}{2}!)]^2}$ different item permutations. One need not even look at the asymptote, because as n approaches, say, 40, there are over 50 billion permutations, which then raises the question of the optimal way to split the test. In studying this issue, Cronbach (1951) derived an inequality that, in effect, set the average of all splits of a given test as the lower

bound for a test's true reliability:

$$\text{COR}(X_i, X_i') \geq \frac{n}{n-1}[1 - \frac{\sum_{1=1}^{n} VAR(V_i)}{VAR(X)}] = \alpha.^{7} \tag{11}$$

If the all the test splits result in the same T_i, the the lower bound sign is replaced with an equal sign.[8]

Since the focus of CTT is the true score and its reliability, item properties are not part of the theory, and must be developed rather *post-hoc*. As tests are usually made up items, Y_i, CTT allows for the following item properties, assuming $X = \sum_{i=1}^{n} Y_i$:

$$\text{VAR}[X] = \sum_i \sum_j \text{VAR}[Y_i]\text{VAR}[Y_j]\text{COR}[Y_i, Y_j] \tag{12}$$

where $\text{COR}[Y_i, Y_j]$ is the correlation between items i and j. From probability theory:

$$\text{E}[X] = \sum_{i=1}^{n} \pi_i \tag{13}$$

where π_i is the probability of correctly responding to item i.

The variance for a given item is

$$\text{VAR}[Y_i] = \pi_i(1 - \pi_i) \tag{14}$$

which eventually leads to the coefficient alpha being:

$$\alpha = \frac{n}{n-1}(1 - \frac{\sum_{i=1}^{n} \text{VAR}[Y_i]}{\sum_i \sum_j \sqrt{\text{VAR}[Y_i]}\sqrt{\text{VAR}[Y_j]}\text{COR}[Y_i, Y_j]}). \tag{15}$$

For more detail on CTT, see chapter 1 of Lord (1980), as well as Crocker and Algina (1986), Nunnally and Bernstein (1994), and Osterlind (2005).

Problems with CTT

Embretson and Reese (2002) cite some shortcomings with the CTT model. First, the observed score (X_i) is test dependent; that is, it applies only to items on a specific test (or a parallel form of the test). Consequently the parameters obtained from this model are

test dependent. Second, even though the CTT model has two independent variables (i.e., true score and error) that are independent of each other for given examinee, they are only separable at the population level. Third, the CTT model does not link the properties of specific items with behaviors. Fourth, since item properties are not specified in the model, they must be justified outside the model, usually by denoting their impact on test information, such as reliability.

Lord (1980), too, had criticisms of CTT, writing that it makes no assumptions about item properties that are beyond the control of the psychometrician. Further, CTT does not allow prediction of item responses unless the items were previously administered to very similar individuals, a serious problem Lord (1980) writes about as follows:

> we need to be able to predict the statistical and psychometric properties of any test that we may build when administered to any target group of examinees. We need to describe the items by item parameters and the examinees by examinee parameters in such a way that we we can predict probabilistically the response of any examinee to any item, *even if similar examinees have never taken similar items before.* (p. 11, emphasis added)

Both Embretson and Reese (2002) and Lord (1980) suggest that IRT can help overcome many of these various problems.

Nandakumar and Ackerman (2004) write that the advantages of IRT are associated with the strong models it uses to characterize examinees' performance on tests, as opposed to CTT, where the theories are "tautologies and not testable" (p. 93). Likewise, Camilli and Shepard (1994) write

> Item response theory has several advantages over classical measurement theory .... First, IRT estimates of item parameters ... are less confounded with sample characteristics than are those of classical measurement theory. Second, the statistical properties of items can be described in a more precise manner, and consequently, when a test item functions differently in two groups, the *differences* can be described more precisely. Third, the statistical properties of items can be more readily graphed with the IRT approach, which speeds and broadens understanding of items showing DIF [Differential Item Functioning]. (p. 47, emphasis in original)

Item Response Theory

Brief History

IRT can trace its roots to Lord and Novick's (1968) classic treatise, where Birnbaum (1968) wrote four chapters on it, although it does have a prehistory before Birnbaum's treatment (see Baker, 1992). Like CTT, IRT is concerned with the measurement of theoretical constructs (e.g., ability) that have no concrete reality (Thissen & Orlando, 2001). Consequently, the theoretical constructs have to be measured by analogy with something that is directly measured.

Binet and Simon (1905) first used age as a concrete indicator of the theoretical concept of cognitive development, and hence birthed "modern" psychometrics by trying to measure children's mental age. Their central idea was that there are a number of tasks that a child at a given chronological age can do, and more mature children (i.e., higher in mental age) can do more of these tasks, and vice versa with less mature children. Put another way, each task has an age associated with it where one can expect proficiency, and by observing whether children are ahead or behind this expected proficiency, one can infer something about mental age.

Although not doing work specifically in IRT, Thurstone (1925) developed scaling techniques to improve the measurement of mental age that would later become a central core of IRT. He wrote,

> We assume the distribution of intelligence of children of any given age group to be approximately normal. Since test-intelligence [e.g., *ability*] is indicated by the correctness of answers to questions, it is legitimate to designate the points on the scale of test-intelligence by means of the questions as landmarks. *Each test question is located at a point on the scale so chosen that the percentage of the distribution to the right of that point is equal to the percentage of right answers to the test question for children of that specified age* . . . If we know the percentage of children of different ages who can answer each question, it is possible to locate the questions on an absolute scale, and it is also possible to locate the means of successive age groups on the same absolute scale. (pp. 436-37, emphasis original)

Obviously, his seminal idea was that test items and examinee ability could be defined on the same scale—in stark contrast to what is done in CTT.

Another idea from Thurstone's (1925) item-ability relationship is that when items are plotted as a function of age, the proportion of a given age group's correct responses for a given item should have a cumulative normal distribution (i.e., normal ogive):

$$F(x_i; \beta_i, \alpha_i, \theta_j) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{\frac{z^2}{2}} dz \tag{16}$$

where $z = \alpha_i(\theta - \beta_i)$, β_i is the difficulty of an item (i.e., the point where 50% of a given group respond correctly), and α_i is the ith item's discrimination (i.e., the rate of change of the ogive as a function of θ). The ogive is sometimes called an *item characteristic curve* (ICC). Because this manuscript deals solely with binary items, ogive and ICC can be used interchangeably. For a geometric example of an normal ogive, see Figure 1.

Each ICC has two major characteristics. The first is the point at which there is a 50-50 chance of responding correctly to the given item. This is called the item's location, and is conceptually the same thing as the difficulty, β, parameter. The second characteristic is the rate of change in the slope, which is conceptually identical to the item's discrimination, α.

Thissen and Orlando (2001) write that although approximating item-response data by normal models was "intuitively satisfying," the early work in IRT was imprecise about what was normally distributed and what was not, nor was it precise about which variables should be fixed and which variables should be stochastic. Lord and Novick (1968) and Birnbaum (1968) helped clarify these issues, as will be explicated in the following section.

In the 1960's, a second "metaphor" appeared using IRT, namely, the work done with dose-response curves in bioassay research. Drug research, especially with non-human subjects, often had binary outcomes to measure: e.g., did the subject die with a given dose of a drug ? (Thissen & Orlando, 2001) Consequently drug researchers often used normal ogives as a way to measure the point as which one expects the subjects to have a 50-50 chance of dying, or the β parameter. What made this line of research so attractive to psychometricians was the high quality of statistical research behind the dose-response curves, which could easily be transported into an IRT context. One major idea IRT transported from the bioassay research was the approximation of the normal ogive by a

logistic response model, which many drug researchers advocated over the normal model. The use of the logistic model came into prominence in IRT when Allen Birnbaum (1968) wrote his seminal chapters. In them, he, too, advocated the use of the logistic model over that of a normal ogive for computational reasons. For more detail of the translation of work in biometrics to IRT, see Chapter 1 of Baker (1992).

Birnbaum (1968) noted that the difference in area between the logistic model (defined below) and the normal cdf [cumulative distribution function, $F(\cdot)$] with mean of zero and standard deviation 1.7 was less than 0.01, and concluded that

> We may view the logistic form for an item characteristic curve as a mathematically convenient, close approximation to the classical normal form, introduced to help solve or to avoid some mathematical or theoretical problems that arise with the normal model. Or we may view it as the form of a test model that is of equal intrinsic interest and of very similar mathematical form. (pp. 399-400)

Consequently, by side-stepping the use of the normal ogive, it means that there is no integration, which can save much computational time. The 1.7 SD that Birnbaum (1968) wrote distinguishes the logistic model from the standard normal can be placed back into the logistic IRT model to give it a shape closer to the normal ogive:

$$f(x_i; \boldsymbol{\kappa}_i, \theta_j) = \frac{1}{1 + e^{-1.7a_i(\theta_j - b_i)}}. \tag{17}$$

This is the extent of IRT history needed for this text. For more elaborate histories, see Baker (1992) and Bock (2003).

Overview

Item response theory is "model-based measurement in which trait level estimates depend on both persons' responses and on the properties of the items that were administered" (Embretson & Reese, 2002 p. 13). More specifically, IRT models specify how an individual's trait level and an item's properties are related to how a person responds to that given item.

In contrast to CTT, IRT begins with (strong) assumptions about the model. First, the ICC is monotonically increasing (although not strictly so), and, second, that there is

local independence between the test items. The (non-strict) monotonic increasing assumption simply holds that for every $k \underset{k \neq k'}{\geq} k'$ in the domain of the IRT function $f(\cdot)$, the range of $f(\cdot)$ is as follows: $f(k) \geq f(k')$. The local independence assumption holds that the items of a given test are independent, *given the model parameters*. More formally, items are locally independent when the probability of response to one item, i, is independent of the outcome of any other item, i', controlling for the examinee's latent ability and the item parameters. It can be written as

$$P(x_i = 1|\theta_j) = P(x_i = 1|\theta_j, X_k, X_l, \ldots) \quad (i \neq k, l, \cdots) \tag{18}$$

where θ_j is person j's ability (defined later) and X_i is the ith item on a test.[9] This definition assumes unidimensionality in θ, but that need not be the case. If the test measured more than one dimension, then θ_j needs to be replaced with a vector of values (i.e., $\underset{j=1,2\ldots,m}{\theta_j} \Rightarrow \boldsymbol{\Theta}_j$).

If the local independence assumption holds, then the number of parameters used for the model is its dimensionality. Nandakumar and Ackerman (2004) write, "local independence and dimensionality assumptions are intertwined. One can only statistically test either of the assumptions assuming the other" (p. 93). Further, they write that many tests in the achievement genre are intended to be unidimensional, but this is something, "given test data, we need to empirically determine" (p. 94).

As a side note, it is often difficult, if not impossible, to develop a strictly unidimensional test that measures academic achievement or cognitive abilities (Carroll, 1993). Nonetheless, decades of work in g theory has shown that tests that measure some construct requiring cognitive ability, also measure general intelligence (Jensen, 1998). Consequently, for the purposes of this manuscript, unidimensionality will follow Nandakumar and Ackerman (2004): "In any test, it is not uncommon to find transient abilities common to one or more items . . . In this sense, unidimensionality refers to the dominant ability measured by the test" (p. 95).

IRT Models

Lord (1980) writes that IRT starts with a mathematical statement describing how a response to a given item depends on the test taker's ability.[10] The IRT model used in this text is based on the cumulative logistic distribution, which is an non-complex, nonlinear way to specify the probability of a given response. More explicitly, if x_i is the response to item i, then the probability that person j correctly responds to item i (assuming no partial credit) is:

$$f(x_i; \boldsymbol{\kappa}_i, \theta_j) = \frac{1}{1 + e^{g(\theta_j, \boldsymbol{\kappa}_i)}} \tag{19}$$

where $g(\theta_j, \boldsymbol{\kappa}_i)$ is a function of the item and examinee parameters (to be defined later) and e is the natural log base (i.e., 2.71...).

IRT is based on the item response function, $f(x_i; \boldsymbol{\kappa}_i, \theta_j)$, which maps the ability of examinee j, as measured by the test containing item i, to the probability of a correct answer of item i (Hambleton & Swaminathan, 1985). The geometric representation of the IRT function is the *item characteristic curve* (ICC; see Figure 1). In the ICC, the abscissa represents various levels of ability, θ, whereas the ordinate represents the probability of getting item i correct, given the item parameters ($\boldsymbol{\kappa}_i$).

The item response function describes the *conditional probability* of the correct response to item i; more specifically, for a given ability level (i.e., $\theta = \theta_j$), $f(x_i; \boldsymbol{\kappa}_i, \theta_j)$ is the probability of a correct response, and its range is $[0, 1]$.

An alternative parametrization uses a normal ogive model. While normal ogive models will contain the same number of parameters as the corresponding logistic model, a different function is used to obtain the ICC.[11] Here, the probability of success is given by the cumulative distribution function (CDF), which means "the normal ogive model gives the ICC as the proportion of cases below a certain standard score $[z_{ij} = \alpha_i(\theta_j - \beta_i)]$" (Embretson & Reese, 2002, p. 76), as follows:

$$P(x_i = 1) = \int_{-\infty}^{z_{ij}} (2\pi)^{-\frac{1}{2}} \exp(\frac{-z^2}{2})\, dz \tag{20}$$

where π is the geometric constant 3.14... In this parametrization, the standard score (z_{ij})

contains the IRT model parameters.

If one plots the ICCs from both the logistic ogive and the normal ogive on the same graph, one will find that they have the same inflection point (β_i) for a given item, but slopes will differ, with the logistic model having more spread. To make the two models the same (or at least very similar), a constant needs to be multiplied to the item response kernel (i.e., $\alpha_i(\theta_j - \beta_i)$). Haley (1952) showed that if the logistic IRT kernel is multiplied by 1.7, the absolute difference between the normal ogive and the logistic ogive is less than .01 across the full range of θ. Thus, for the logistic model, $g(\theta_j, \boldsymbol{\kappa}_i)$ needs to be multiplied by 1.7 for it to be equivalent to the normal ogive model. For simplicity, this text will define $M \equiv 1.7$ in its models.

Baker (1992) writes that the main property of the logistic model that makes it more appealing than the normal ogive model is that its CDF is a closed form and, thus, can be computed directly. In addition to the closed form, the logistic function is related to the logarithm of the odds of getting the item correct. For a given item, for any θ_j, the probability of a correct response is given by $f(x_i; \boldsymbol{\kappa}_i, \theta_j)$, which makes the probability of an incorrect response $1 - f(x_i; \boldsymbol{\kappa}_i, \theta_j)$. Define $P_i(\theta_j) \equiv f(x_i; \boldsymbol{\kappa}_i, \theta_j)$ and $Q_i(\theta_j) \equiv 1 - f(x_i; \boldsymbol{\kappa}_i, \theta_j)$, then

$$\ln\left[\frac{P_i(\theta_j)}{Q_i(\theta_j)}\right] = \alpha_i(\theta_j = \beta_i) = \zeta_i + \lambda_i\theta_j \tag{21}$$

where $\zeta_i + \lambda_i\theta_j$ is a linear regression line relating θ_j to the kernel of the IRT model (for further detail, see Baker, 1992).

One major aspect of IRT models that distinguish them from CTT models is that in IRT, "*the item parameters are not dependent upon the ability level of the examinees responding to the items*" (Baker, 2001, p. 52, emphasis added). In practice, this means that two groups who differ widely in ability, θ, can take the same test and the item parameters, $\boldsymbol{\kappa}_i$, will be the same (for a more in-depth explication of this property, see Chapter 3 of Baker, 2001). For the purposes of this text, the key point about this group invariance property of IRT is that *the item parameters are properties of the item, not the people responding.* Moreover, the number of people at each level of θ_j does not affect the

group invariance property.

The next few subsections will be devoted to defining the various IRT models.

One-Parameter Model (1PL). Rasch (1960/1980) first developed his 1PL IRT model by specifying that a person should be characterized by degree of ability, ξ, and an item by a degree of difficulty, δ, with both being greater than zero. In addition, he believed that the probability of getting an item correct should be "a function of the ratio, ξ/δ, . . . the simplest function I know of, which increases from 0 to 1 as ζ goes from 0 to ∞, is $\frac{\zeta}{1+\zeta}$. If we insert $\zeta = \frac{\xi}{\delta}$ we get $\ldots \frac{\xi}{\xi+\delta}$" (Rasch, 1960/1980, pp. 74-75).

The current one-parameter model (1PL) (sometimes called the Rasch model) has $\boldsymbol{\kappa}_i' = [\alpha, \beta_i]$ which means the kernel becomes $\alpha(\theta_j - \beta_i)$, yielding the 1PL version of (19):

$$f(x_i; \boldsymbol{\kappa}_i, \theta_j) = \frac{1}{1 + e^{-\alpha M(\theta_j - \beta_i)}}, \tag{22}$$

where α has no index because it is held to be the same across items.

Some authors describe the 1PL model as having no α parameter, or, equivalently, having $\alpha = 1$ for all items. Thissen and Orlando (2001) write that this is not technically correct, as the Rasch IRT model simply requires the discrimination parameter to be *the same* across items. It might be that $\underset{\forall\, i}{\alpha_i = 1}$, but it can be another number.

Although not previously explicated, θ_j and β_i are on the same metric, so the difference between the two is a continuous measure that provides an index of how difficult item i should be for person j.

Two-Parameter Model (2PL). The two-parameter model differs from its one-parameter counterpart in that it allows the item discrimination parameter, α_i, to vary across items, yielding:

$$f(x_i; \boldsymbol{\kappa}_i, \theta_j) = \frac{1}{1 + e^{-\alpha_i M(\theta_j - \beta_i)}}, \tag{23}$$

where, obviously, $\boldsymbol{\kappa}_i' = [\alpha_i, \beta_i]$.

An item discrimination parameter is needed when the items of a given test do not equally measure how a person relates to the test's latent trait(s).[12] In more practical terms, it allows a test's ICC's slopes to vary as a function of the ability level θ_j.[13] It will

reach a maximum when $\theta_j = \beta_i$.[14] Consequently, a given item can best distinguish between examinees whose ability, θ_j, is close to β_i.

Three-Parameter Model (3PL). The three parameter model adds an additional parameter to the 2PL model. This extra parameter, γ_i, allows the ICC to have a lower asymptote greater than 0 (i.e., allows the probability of success for the lowest group to be > 0) , hence it is sometimes called a "guessing parameter."[15]

It is modeled as follows:

$$f(x_i; \boldsymbol{\kappa}_i, \theta_j) = \gamma_i + (1 - \gamma_i)\frac{1}{1 + e^{-\alpha_i M(\theta_j - \beta_i)}}.^{16} \tag{24}$$

Following Thissen and Orlando (2001), the derivation of the γ_i parameter is as follows. The probability that a person responds correctly to a test item is influenced by γ_i, the probability that an examinee responds correctly, even if he/she does not know the answer. More formally, if ι_i is some threshold that, if exceeded, results in a correct item response, and Y_i is the response process, then the answer to item i, u_i, is:

$$u_i = \begin{cases} 1, & \text{if } Y_i \geq \iota_i \\ 1, & \text{if } Y_i < \iota_i, \text{ with probability } \gamma_i \\ 0, & \text{if } Y_i < \iota_i, \text{ with probability } 1 - \gamma_i \end{cases}$$

Consequently, the probability of a correct response, $P(X_i = 1)$, can be bifurcated (excluding M):

$$
\begin{aligned}
&\gamma_i P(Y_i < \iota_i|\theta) + P(Y_i \geq \iota_i|\theta) \\
&= \gamma_i[1 - P(Y_i \geq \iota_i|\theta)] + P(Y_i \geq \iota_i|\theta) \\
&= \gamma_i[1 - \frac{\exp[\alpha_i(\theta_j - \beta_i)]}{1 + \exp[\alpha_i(\theta_j - \beta_i)]}] + \frac{\exp[\alpha_i(\theta_s j - \beta_i)]}{1 + \exp[\alpha_i(\theta_j - \beta_i)]} \\
&= \gamma_i - \gamma_i \frac{\exp[\alpha_i(\theta_j - \beta_i)]}{1 + \exp[\alpha_i(\theta_j - \beta_i)]} + \frac{\exp[\alpha_i(\theta_j - \beta_i)]}{1 + \exp[\alpha_i(\theta_j - \beta_i)]} \\
&= \gamma_i + (1 - \gamma_i) \frac{\exp[\alpha_i(\theta_j - \beta_i)]}{1 + \exp[\alpha_i(\theta_j - \beta_i)]} \\
&= \gamma_i + (1 - \gamma_i) \frac{1}{1 + \exp[-\alpha_i(\theta_j - \beta_i)]}.
\end{aligned}
$$

There are problems with including a γ_i parameter, primarily a non-conversion in parameter estimation (Embretson & Reese, 2002). Thus, it is not uncommon to have γ_i be restricted to be the same across q items, where q is upper bounded by the total number of items on a test. Another problem with using a γ_i parameter is that item difficulty, β_i, takes on a different meaning. Specifically, while it still is the point of inflection, it no longer is the trait level where the probability of success is .50 because it is shifted by γ_i . To be more specific, it is now the probability of a correct answer at $(1 + \gamma_i)/2$. Likewise, the discrimination parameter is conceptually the same, but its mathematical form changes so that at $\theta_j = \beta_i$, the discrimination parameter is $\frac{\alpha_i(1-\gamma_i)}{4}$.

It is important to notice that in (24) the value of γ_i *is not* a function of θ_j. Consequently, the guessing parameter acts the same across all ability levels (i.e., the highest and lowest level examinees have the same probability of of getting an item correct by guessing). While in theory $0 \leq \gamma_i \leq 1$, "in practice, values values greater than .35 are not considered acceptable" (Baker, 2001, p. 28).

When choosing an IRT model (or any other mathematical model), one must be cognizant of how close the model fits the actual data. There are multiple indices, each of which have their own (un)desirable characteristics (e.g., Beguin & Glas, 2001; Fox & Glas, 2005; Orlando & Thissen, 2000). As model fit is not the purpose of this manuscript, the reader is directed to Chapters 8 and 9 of Hambleton and Swaminathan (1985) for some

(introductory) walktroughs and Chapter 8 of du Toit (2003) for a more technically detailed presentation.

IRT Parameters.

Estimating Examinee and Item Parameters Estimating model parameters is the most difficult part of nonlinear regression, as it usually involves an iterative procedure that can only stop when some convergent criterion is met (Schabenberger & Pierce, 2002). Fortunately, this manuscript need not concern itself with estimation procedures, *per se*. Except for some Bayesian methods (du Toit, 2003; Mislevy, 1986), parameter estimation is usually done via Maximum Likelihood (ML) estimation or some permutation of ML. Because ML estimates are used, then, the IRT parameters have many desirable properties, such as being unbiased, having minimum variance, being consistent, and being able to compute the parameters' (co)variance from the inverse of Fisher's (1922) Information function. Much more detail can be found in Baker (1992), Hambleton, Swaminathan, and Rogers (1991) and Lord (1980). The reader is especially encouraged to read Baker and Kim (2004) for derivations of much of the calculus involved in IRT parameter estimation.

The Ability Parameter (θ) Hambleton and Swaminathan (1985) write:

> Ability ... is the label that is used to describe what it is that the set of test items measures ... [it] can be broadly defined aptitude or achievement variable ..., a narrowly defined achivement variable ..., or a personality variable. (p. 55)

Obviously, because θ is a latent construct, it cannot be directly measured, thus tests do not measure it in a absolute sense, like a ruler measures length. Instead, what can be determined is relative positions of individual test takers on the θ continuum, which makes θ's scale arbitrary, and only the differences among the θs that have meaning (i.e., they are on an interval scale; Stevens, 1951).

The difficulty parameter (β) The β parameter represents the difficulty of an item, and is on the same scale as θ. Visually, it is the point along the abscissa on Figure 1 where $f(x_i; \boldsymbol{\kappa}_i, \theta_j) = \frac{\gamma_i+1}{2}$. Camilli and Shepard (1994) write:

> 1. The scale of θ is often equivalent to that of z scores–roughly 99.9% of all scores lie between -4 and +4. However, items for a particular test are usually written so that most of the [βs] fall within the range of about -1.5 to +1.5.
> 2. A high positive [β] parameter means that the item is difficult. A large negative [β] parameter means the item is easy.
> 3. The [βs] are measured in the same units as the θs, which are usually considered to approximate an interval scale. Any linear transformation ... of the [β] scale requires an identical transformation of the θ scale.
> 4. When [βs] are estimated separately for two groups of examinees, each of the the two sets of [βs] has an arbitrary scale, and the sets are, therefore, not directly comparable. However, the two sets of [βs] bear a linear relationship if the assumptions of the most widely used IRT models are correct, and consequently a simple linear transformation can be used to convert the [βs] of one group to the scale of the other group for purposes of comparison. (p. 53)

The discrimination parameter (α) Hambleton and Swaminathan (1985) write

> The intercept parameter is directly related to the concept of item difficulty [in CTT]. Also, [it] functions in a similar way to an item discrimination index in classical test theory. The difference between the probabilities of a correct response at any two ability levels increases directly with the value of [α]. ... The discriminating power of the item is considerably better with the higher [α] value. (p. 28-29).

As was shown earlier, item discrimination is best at about β, and, consequently, it deteriorates as θ diverges from β.

Unlike the β parameter, the scale units of α is not the same as those of the θ, so it does not make sense to compare αs with βs. In fact, the scale of the α is the inverse of β. Thus, if β is multiplicatively transformed by a constant, say $\beta^* = \mathrm{k} \times \beta$, then, the transformation to α^* from α is $\alpha^* = \alpha/\mathrm{k}$. (Baker & Kim, 2004).

Invariance and Arbitrary Nature of of Item Parameters

As already stated, the most most important feature of IRT, at least as it relates to the LFE, is the invariance of item parameters across groups. While the concept was mentioned briefly when explicating the properties of IRT, its importance dictates a more comprehensive explication.

A valid interpretation of the item response function, $f(x_i; \boldsymbol{\kappa}_i, \theta_j)$, is as a regression of item score on ability (Lord, 1980). As is common with other regression models, in IRT

the regression function remains unchanged when the ability distribution, $h(\theta_j)$, changes. More explicitly, the probability of answering question i correctly from someone at ability level θ_j depends *only* on θ_j, not the number of people with ability θ_j, nor the number of people at ability $\underset{j \neq j'}{\theta_{j'}}$.

> Since the regression is invariant, its lower asymptote, its point of inflexion, and the slope at this point all stay the same regardless of the distribution of ability in the group tested. *Thus [α_i, β_i], and [γ_i] are invariant item parameters.* (Lord, 1980, p. 34, emphasis added)

This property of IRT is in dire contrast to CTT, where the item parameters differ for each group tested. As will be developed later, this invariance property is extremely valuable for assessing the LFE, as, in theory, item parameters should not differ between samples, even if the samples take the test, say, 10 years apart.

Another property of items in IRT models is that the scale for θ is arbitrary. More specifically, adding the same constant to θ_j and β_i does not change the item response function, thus the origin of θ is arbitrary. Moreover, multiplying the same constant to θ_j and β_i and the inverse of the constant to α_i keeps the kernel of the item response function (i.e., $\alpha_i(\theta_j - \beta_i)$) unchanged, meaning the scale, and hence the choice of unit, for measuring ability is arbitrary as well.[17] While infinitely many options are available, the most common way to scale θ is to give it a mean of 0 and a variance of 1.

Lord (1980) notes that item parameter invariance holds "only as long as the origin and unit of the ability scale is fixed" (p. 36). This is of great importance to the analyses in this manuscript because, if a given IRT parameter, κ_i, is determined for a given set of items from one group, R, and then, independently, from another group, F, there is no reason to expect $\kappa_{i_R} \underset{\forall i}{=} \kappa_{i_F}$, *although they should be related to each other.* This is the issue of *equating*, and is given much more consideration in the so-named section below. The next sections will develop methods to detect differential item functioning and, then, turn to test equating.

Detection of Changing IRT Parameters

To understand the LFE through IRT methods, one must first understand the two fundamental, and interrelated, concepts in IRT: differential item functioning (DIF) and item parameter drift (IPD).

Differential Item Functioning (DIF)

Camilli and Shepard (1994) write that DIF is a specific type of multidimensionality.[18] It occurs when an item measures a dimension in addition to the primary ability *and* when groups significantly differ on the secondary ability. Multidimensionality, itself, is not sufficient for DIF. If the groups do not differ in their distributions for one (or more) secondary abilities, neither group gets a "boost" by the presence of the secondary dimensions, and, hence, DIF doesn't exist.

Hambleton et al. (1991) define DIF in simpler terms, stating

> An item shows DIF if the item response function across different subgroups are not identical. Conversely, an item does not show DIF if the item characteristic functions across different subgroups are identical. (p. 110)

Lord (1980) first observed that the ICC of an IRT model is ideally suited to study DIF. He wrote,

> If each test item in a test had exactly the same item response function in every group, then people at any given level θ of ability or skill would have exactly the same chance of getting the item right, regardless of their group membership. Such a test would be completely unbiased ... If, on the other hand, an item has a different item response function for one group than for another, it is clear that the item is biased ... it seems clear from all this that item response theory is basic to the study of item bias. (pp. 212-213)

As has been discussed, the IRT model can be interpreted as the conditional probability that person j correctly responds to an item, given his/her ability level, θ_j. DIF, using IRT parametrization, then looks to see if the IRT model differs between two groups, usually termed *focal* (F) and *reference* (R) groups.[19] More specifically, Lord (1980) wrote that DIF detection can be approached by computing the item parameter estimates, $\boldsymbol{\kappa}_i$, within each group, and then testing to see if the item parameters differed

between the groups after correcting for the possibility that θ's distribution might differ between the groups.

Figure 2 shows an example of what what DIF might look like between the Focal and Reference groups. In this figure, the ICC for group R is shifted to the left of the ICC for group F, indicating this item is easier for individuals in group F. Mathematically, this means that $\boldsymbol{\kappa}_{i_F} \neq \boldsymbol{\kappa}_{i_R}$, or, more specifically, $\beta_{i_F} < \beta_{i_R}$. DIF is not constrained to be different in only one parameter, though, and it is conceptually possible to have an item be different in all three parameters across both groups.

Camilli and Shepard (1994), write that DIF can fall into two different categories:

> 1. *Uniform or consistent DIF:* ... consistent DIF occurs when the ICCs for two groups are different and do not cross. There is a relative advantage for one group over the entire ability range. This is necessarily the case when two ICCs have the same [α] parameter.
> 2. *Nonuniform or inconsistent DIF:* ... the ICCs for two groups are different but cross at some point on the θ scale. Therefore, the DIF for and against a given group balance or cancel each other out to some degree. Positive and negative DIF may cancel entirely ... depending on the particular pair of ICCs. (p. 59)

For either case, Camilli and Shepard (1994) write that there are two related approaches used to examine DIF using IRT parameters. The first focuses on the measurement of DIF, and an index is used to convey the magnitude of the DIF. In the second, a statistical test is used to measure the significance of DIF, and the central question is "Are the ICCs for two groups the same in the population?" (p. 64). Although they measure different aspects of DIF, both aspects are important, as they provide a different view of the same phenomenon.

Before delving into various methods for detecting DIF, and as previously stated in the text, if parameters for two groups, completing the same item, are analyzed separately, there is no reason to believe that the parameters will be the same because of their arbitrary scaling. Nonetheless, they should be linearly related to each other. (See the section on test equating for more detail.)

Magnitude of DIF.

Area Indices The most visual expression of DIF is represented by the space between the two ICCs, as shown by Figure 3. Obviously, then, items with larger area have larger DIF. An easy way to measure the area is to use elementary (unidimensional) integration. More explicitly, the area for ICCs that do *not* cross, sometimes called the *signed area*, is

$$\int_{-\infty}^{+\infty} f(x_{i_R}; \boldsymbol{\kappa}_{i_R}, \theta) - f(x_{i_F}; \boldsymbol{\kappa}_{i_F}, \theta)\, d\theta. \tag{25}$$

When the focal group outperforms the reference group, the area will be negative. In the instance where the ICCs *do* cross, (25) is slightly altered, and is renamed the *unsigned area*,

$$\int_{-\infty}^{+\infty} \sqrt{[f(x_{i_R}; \boldsymbol{\kappa}_{i_R}, \theta) - f(x_{i_F}; \boldsymbol{\kappa}_{i_F}, \theta)]^2\, d\theta}. \tag{26}$$

When (26) is much larger than (25), this is an indication that the ICCs cross.

Raju (1990) later derived an exact expression for computing the between-ICC area for the 1PL, 2PL, and (invariant γ across groups) 3PL models. The area for the 3PL is

$$\text{Area} = (1 - \widehat{\gamma}_i) \times \text{Abs}[\frac{2(\widehat{\alpha}_{F_i} - \widehat{\alpha}_{R_i})}{(1.7)(\widehat{\alpha}_{F_i})(\widehat{\alpha}_{R_i})} \ln[1 + e^{\frac{(1.7)(\widehat{\alpha}_{F_i})(\widehat{\alpha}_{R_i})(\widehat{\beta}_{F_i} - \widehat{\beta}_{R_i})}{\widehat{\alpha}_{F_i} - \widehat{\alpha}_{R_i}}}] - (\widehat{\beta}_{F_i} - \widehat{\beta}_{R_i})]. \tag{27}$$

To obtain the 2PL model, simple delete the $(1 - \hat{\gamma}_i)$ term. For the 1PL model, it is simply the absolute value of $\hat{\beta}_{F_i} - \hat{\beta}_{R_i}$.

Raju (1988) also developed the standard error for the derived area, and when the area is divided by its standard error, it, approximately, is normally distributed; thus, a standard normal table can give significance thresholds.

Significance of DIF.

Lord (1980) proposed the following test to detect DIF:

$$d_i = \frac{\widehat{\beta}_{i_F} - \widehat{\beta}_{i_R}}{\sqrt{\text{Var}[\widehat{\beta}_{i_F}] + \text{Var}[\widehat{\beta}_{i_R}]}} \tag{28}$$

where $d_i \overset{n \to \infty}{\sim} N(0, 1)$. In addition, he proposed that the same test could be run for α_i.[20] Multivariately, Lord (1980) wrote that a general test of the joint difference between $[\alpha_i, \beta_i]$

for F and R is

$$D_i^2 = \mathbf{v}_i' \Sigma_i^{-1} \mathbf{v}_i{}^{21} \tag{29}$$

where $\mathbf{v}_i = [\widehat{\beta}_{i_F} - \widehat{\beta}_{i_R}, \widehat{\alpha}_{i_F} - \widehat{\alpha}_{i_R}]$, Σ_i^{-1} is the sampling (co)variance matrix of the differences between item parameter estimates, and $D_i^2 \sim \chi^2_{(2)}$.[22]

General IRT-Likelihood Ratio (IRT-LR) Tests of significance using the IRT-LR involve comparisons between the parameters in two nested models: the *compact* ($\boldsymbol{\kappa}_{i_C}$) and the *augmented* ($\boldsymbol{\kappa}_{i_A}$) models. As their names suggest, the augmented has all the parameters of the compact one, in addition to the others that would be needed if DIF existed (i.e., $\boldsymbol{\kappa}_{i_C} \subset \boldsymbol{\kappa}_{i_A}$). The goal, then, is to see if the parameters in $\boldsymbol{\kappa}_{i_A}$ that are not in $\boldsymbol{\kappa}_{i_C}$ are equivalent to $\mathbf{0}$ (which serves as the null hypothesis in this test). The form of the LR test is

$$G^2_{df} = 2 \ln \left[\frac{\text{Likelihood}(f_A)}{\text{Likelihood}(f_C)} \right] \tag{30}$$

where Likelihood $[\cdot]$ is the likelihood of the data, given the ML parameter estimates and *df* is the difference in number of parameters between f_A and f_C (Thissen, Steinberg, & Wainer, 1993). Of course, $G^2_{df} \sim \chi^2_{df}$ under the null hypothesis and some very general assumptions (Rao, 1973), so for any obtained value bigger than $\chi^2_{df(\alpha_0)}$, where α_0 is the probability of the Type I error, reject f_C.

Although Thissen et al. (1993) give two uses for the G^2 statistic, only one need concern this paper. They write,

> to test DIF for item i, we compute the ML (Maximum Likelihood) estimates of the parameters of the compact model (with no DIF for item i) and the likelihood under that model, and the ML estimates and likelihood of the model augmented by some parameters representing differences between the item i parameters for the reference and focal groups. Then the likelihood ratio statistic provides a test of the significance of DIF on [*df*]. (p. 74)

Item Parameter Drift (IPD)

Thissen et al. (1993) write,

> [In IPD studies] there are two (or more) groups of examinees, and the research question is whether the item parameters, and therefore the trace lines [the

> ICCs] differ between or among the groups. The only real difference ... is that the drift question involves time, and ... that the primary concern of the analysis is with the item difficulty parameter. (p. 83)

Basically, the question that IPD studies seek to answer is if a given item has become easier (or harder), more (or less) discriminating, and/or if the probability of a correct answer due to guessing has changed over time. It is important to study because it "threatens the validity of scores by introducing trait irrelevant differences over time," and failing to detect it disadvantages test takers as well as makes statements about trends in a test questionable (Donoghue & Isham, 1998, p. 40).

Obviously, then, IPD is just a special type of DIF. Donoghue and Isham (1998) write,

> The problem of IPD is formally identical to that of DIF. Does the item function the same in two sets of data? Whereas DIF analyses examine whether items function differently in examinee subgroups ..., IPD is particularly relevant to time of testing, but the underlying question is the same. *Thus, DIF procedures may be used to assess IPD.* (p. 33, emphasis added)

Donoghue and Isham (1998) compared various procedures to detect IPD. They found three IRT-based procedures particularly useful in IPD detection: (a) *Lord's* χ^2*, with* γ_i *constrained*, which was developed earlier [see (28)]; (b) *z(H)*, which is essentially the same as a significance test for comparing the area between ICCs that Raju (1988, 1990) developed; and (c) the *Closed Interval Signed Area, constraining* γ, an area measure developed by Kim and Cohen (1991), very similar to the signed area in (25).

The key point to obtain from tests with items containing DIF (or drift) is that those particular items are not measuring the same construct in the two populations; or, more specifically, although they are measuring the same trait (i.e., the unidimensionality assumption), they are not measuring it the same way.

Test Equating

As previously mentioned in the general text, if IRT parameters are estimated for more than one group in separate analyses, they are not directly comparable, because the scale of the θs (and βs) is arbitrary, although linearly related. Consequently, before the

parameters are compared, they need a common metric. Finding a common metric is the process of *test equating*. Hambleton et al. (1991) explain the process as follows:

> Through [equating,] a correspondence between scores on X and Y is established, and the score on test X is converted to the metric of test Y. Thus, an examinee who obtains a score x on test X has a converted score y^* in test Y; this score is comparable to the score y of an examinee taking test Y. (p. 123)[23]

In CTT, there are many assumptions that must be met before test scores can be equated, but, using IRT, many of the problems are overcome as, by the invariance property, "if the item response model fits the data, direct comparison of the ability parameters of two examinees who take different tests is made possible" (Hambleton et al., 1991, p. 125). More importantly for studying the LFE, though, IRT methods allow for *nonequivalent groups equating*, which "has no counterpart in classical test theory."(Zimowski, 2003).

If, as much of the LFE contends, groups of individuals come from different ability groups, then the individuals from the various groups can be placed on the *same* underlying latent distribution via nonequivalent groups equating. This makes comparisons especially easy, as, assuming the underlying latent distribution of intelligence is normal (Burt, 1957; Jensen, 1998; but also see Burt, 1963), one can determine how many standard deviations one group's (average) cognitive ability is from another's. The only caveat, which will not affect LFE studies, is that either (a) there must be a subset of common items in both forms, or (b) the same participants must take both forms of the examination (Yu & Osborn Popp, 2005). With regard to the first, Zimowski (2003) writes that these linking items should "have relatively high discriminating power, middle range difficulty, and should be free of any appreciable DIF [or drift] effect." In other words, they need to measure the same underlying construct(s) the same way in both groups. With regard to the second, more participants are better, preferably those spanning the entire distribution of the latent trait, but especially the portion where most of the area resides.

The purpose of this text is not to explicate various methods for test equating. For the interested reader, there are many sources available: Hambleton et al. (1991), pp. 136-144; Kolen and Brennan (1995), Chapter 6; Hambleton and Swaminathan (1985),

Chapter 10; and Cook and Eignor (1991). For a particulary informative walk-through of *both* item and person equating, see Yu and Osborn Popp (2005).

Chapter 3

METHOD

The purpose of this study is to demonstrate the use of IRT methods in assessing the LFE. More specifically, this text seeks to show how IRT methods can be used to determine whether the increase in mean IQ scores across time (i.e., the LFE) is due to a co-occurring increase in intelligence or if it is due to a psychometric artifact. To that end, this study will assess the LFE in both simulated and real test scores via CTT and IRT methods.

Study 1

Data Generation.

Using the SAS macro IRTGEN (Whittaker, Fitzpatrick, Williams, & Dodd, 2003), dichotomous data was generated according to the 3PL IRT model. An infinite amount of item parameters distributions could be used to generate the items, but for this study the item parameter distributions will be:

$$\alpha = 1$$

$$\beta \sim N(0, 1)$$

$$\gamma = .1$$

$$\text{and examinee distribution :}$$

$$\theta \sim N(0, 1).^{24}$$

The particular parameters for this *original* group were chosen for both their innate simplicity as well as the ease of interpretation when comparing to other item generations.

The sample size for this, and all other simulated samples, is 1000 and the number of items will be 60, which is approximately the number of items analyzed on the *CBASE* in the second data analysis. (For SAS code, see Appendix B).

Comparison Samples As the time span for the two forms of the *CBASE* is approximately six years, the following changes in item and trait levels for the simulated data were used to evidence both a real increase in intelligence and psychometric artifacts that mask themselves as the LFE over approximately six years.

Psychometric artifact: Increase in guessing. The purpose of this data generation will be to produce an artificial increase in intelligence from the the original sample. This will be produced by uniformly increasing the γ parameter to .15. The α, β, and θ values will be the exact same as in the *original* sample.

Psychometric artifact: Decrease in difficulty. The purpose of this data generation will be to produce an artificial increase in intelligence from the original sample. This will be produced by decreasing the location parameter value for the β distribution to: $\beta \sim N(-.1, 1)$. The α, γ, and θ values will be the exact same as in the *original* sample.

Psychometric artifact: Increase in guessing and decrease in difficulty. The purpose of this data generation will be to produce an artificial increase in intelligence from the the original sample by combining both an increase in guessing as well as a decrease in (average) item difficulty. This will be produced by both uniformly increasing the γ parameter value to .15 and decreasing the location parameter value for the β distribution (i.e., $\beta \sim N(-.1, 1)$). The α and θ values will be the exact same as in the *original* sample.

Real rise in intelligence. The purpose of this data generation will be to mimic the *original* sample, but with an actual increase in intelligence. This will be produced by increasing the value of the location parameter for the distribution of θ .14 of a standard deviation (i.e., $\theta \sim N(.14, 1)$). The item parameter values will be the exact same as in the *original* sample. The reason .14 was used is that, assuming intelligence goes up $.\bar{3}$ IQ points a year (Flynn, 1984, 1987, 1999), this is the approximate increase in six years (i.e., $\frac{(.14)(15)}{3} \approx 6$).

Data Analysis.

CTT Analysis The CTT analysis mimicked that normally done with peer-normed tests. More specifically, the mean, μ_0, and standard deviation, σ_0, was obtained from the first 1000 examinees taking the *original* simulated examination. Then, with the subsequent comparison sample generations, each individual's score was placed on the *original* sample's metric. The samples were then averaged to determine the overall mean score increase (i.e., determine the size of the LFE).

IRT Analysis IRT allows for multiple types of analyses, the full scope of which are beyond the purposes of this study. Since it is known, *a priori*, what items contain drift and this study is not a study on methodology of drift analysis, the focus will be on using IRT to obtain more accurate examinee scores. To that end, the situation is akin to that of vertical test equating (See the *Test Equating* section in the literature review).

For vertical test equating there needs to be common items, without DIF/drift, in each item subset. Consequently, the last five items from the *original* test simulation replaced the last five items for the other test simulations, except the real increase in the latent trait group, whose last five items were originally specified to be the same as those in the *original* group. This allows for 60 items to remain for each test as well as ensures that the real parameters for the equating items are (a) the exact same, and (b) a random selection from the "parameter space" of items. All IRT analyses were carried out in BILOG-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). BILOG-MG estimates the item parameters of 2PL and 3PL models using marginal maximum a posteriori estimation, which allows a side-stepping of any possible Heywood cases (du Toit, 2003, pp. 601-602). For examinee's ability scores, BILOG-MG uses maximum likelihood estimation, maximizing the likelihood function via the Newton-Raphson estimation procedure (du Toit, 2003, pp. 606, 833-834).

Study 2

Data Generation.

Data came from the College Basic Academic Subjects Examination (CBASE; Osterlind & Merz, 1990; Osterlind, Sheng, Juve, Beaujean, & Nagel, in preparation). The

CBASE is a criterion-referenced achievement test that assesses an examinee's knowledge and skills in English, mathematics, science, and social studies. It is designed for college-age students; thus, they make up the majority of the examinees in the data set. An extensive review of the CBASE can be found in pp. 102-111 of Juve (2004). For this particular investigation, in an attempt to control for heterogeneity of the data, only respondents who took the examination as part of their teaching credentialing process were used.[25]

The CBASE data for this project came from the Mathematics domain because Osterlind and Beaujean (2005) found that it had both the highest reliability estimates and factor loadings on g of all the subjects (see Tables 2 and 3). In addition to the desirable psychometric properties, to date, there has been no investigation of the LFE with CBASE data, although the test has been administered for over 10 years (Osterlind & Merz, 1990; Osterlind et al., in preparation). Moreover, while more verbal academic subjects such reading and spelling have not shown evidence of the LFE, mathematic achievement tests have (Scott, Bengston, & Gao, 1998). Thus, the mathematics section of the CBASE appears to be a viable instrument to demonstrate the investigation of the LFE using IRT methods.

The CBASE forms used were LK and LO. For the CBASE data available, they have the largest time-span between them, being administered in 1996-1997 and 2000-2001, respectively. The two forms have a relatively large subset of items in common, 16. For the CBASE data available, form LK had 619 respondents, while form LO had over 5500. Consequently, a random sample of 619 was selected from LO to give equal sample sizes for both groups.

CTT Analysis.

The CTT analysis followed the same procedures as with the simulated data. Namely, the mean and standard deviation was obtained for all examinees taking form LK, then were used to create a standardized deviation score for form LM. The standardized deviation scores were then averaged for form LM to assess the magnitude of the LFE.

IRT Analysis.

For the IRT analysis, the item parameters were estimated via BILOG-MG

(Zimowski et al., 1996) for all the items in the Mathematics domain. Then the 16 common items across forms were compared for DIF/drift using two of the methods suggested by Donoghue and Isham (1998): *Lord's* χ^2 and the *z(H)*, which assess the significance of area between group ICCs (Raju, 1988, 1990). Any items not exhibiting DIF/drift were used as anchor items, and the two forms of the test were equated, then the latent trait scores compared.

Chapter 4

RESULTS

The purpose of this study was to demonstrate the use of IRT methods in assessing the LFE. More specifically, this text sought to show how IRT methods can be used to determine whether the increase in mean IQ scores across time (i.e., the LFE) is due to a co-occurring increase in intelligence or if it is due to a psychometric artifact. To that end, this study assessed the LFE in both simulated and real test scores via CTT and IRT methods.

Study 1

The first study involved data simulation. As detailed in the *Method* section, 1000 examinees were generated with an underlying standard normal trait distribution. These examinees took three different simulated tests of 60 items each. The differences between the tests were minor, and are detailed above. Next, 1000 different examines were generated with a $N(.14, 1)$ trait distribution and given the same simulated test as the *original* group of 1000 examinees. This process was repeated 100 times. The item scores were then analyzed via CTT and IRT methods.

CTT Analysis. For the CTT analysis, the answers for the first group of examinees taking the original test were summed and the mean and standard deviation of the summed scores were calculated.[26] Next, for each of the other tests, the mean and standard deviation of the initial sample (who took the *original* test) were used to generate a standardized deviation score, and the average of this score was used as the indicator of the magnitude of the "true score" increase.

The averaged values for each of the 100 iterations can be found in Table 4. At the bottom of the table is the average "true score" increase for each of the four groups. As can be seen from the table, the mean score for the true latent score increase (.135) is practically indistinguishable from the group with only a (mean) decrease in the items' β parameters, as well as the group with only the γ parameter increased .05 units (.117 and

.149, respectively).[27] Thus, using CTT methods, a true increase in the latent trait (intelligence) is indistinguishable from samples with no increase in the latent trait, but with very minor perturbations in the item's properties (i.e., psychometric artifacts).

IRT Analysis.

Since only two of the three psychometric artifact groups were indistinguishable from the true increase in latent ability group, only those groups were compared in the IRT analysis. The averaged values for each of the 100 iterations can be found in Table 5. At the bottom of the table is the average latent variable increase for each of the three groups. As can be seen from the table, the mean score for the true latent score increase (.101) is very different from either of the psychometric artifact groups (.01 and .003, respectively).[28] Thus, using IRT methods, a true increase in the latent trait is distinguishable from (very small) psychometric artifacts.

Of particular note, though, were two iterations that indicated the latent trait score was higher in the psychometric artifact groups than in the true increase group (see iterations marked with a + in Table 5). Moreover, there were some iterations whose average latent trait score for the true increase group was substantially different from the true value of the increase in the latent variable (see iterations marked with a * in Table 5). The cause for these anomalies is unknown, and needs further investigation in future LFE simulation studies.[29]

Study 2

The second study involved data from the College BASE examination (Osterlind et al., in preparation). As detailed in the *Method* section, two forms of the CBASE were used, form LK and form LO. Form LK was administered during 1996 and 1997, while form LO was administered during 2001 and 2002. 619 respondents from both form LK and form LO were used.

CTT Analysis.

Descriptive statistics for the CTT analysis are given in Table 6. As can be seen from the table, there appears to be a reverse LFE, with a standardized increase of -0.178.

As the use of IRT to assess the LFE is unaffected by the direction of the effect, the same procedures outlined in the *Method* section to assess the LFE via IRT were used to see if the (reverse) LFE is a psychometric artifact or a real change in the latent trait.

IRT Analysis.

For the IRT analysis, first the items were rearranged so that of the p total items, the r items in common were first, then the $p - r$ items left over were placed in order of appearance on the CBASE examination. See Table 9 for item labels and orders.

For the 16 items common to both forms, DIF/drift analysis was run using BILOG-MG to obtain initial parameter estimates, separately, for each form using a 3PL model. The forms were then linked via ITERLINK (Stark, Chernyshenko, Chuah, Lee, & Wadington, 2001), a program that also computes Lord's (1980) χ^2 statistic for each item, assessing both the α and β parameters.[30] Using a Bonferroni adjusted α of .00313 (i.e., $\frac{.05}{16}$), none of the items exhibited DIF/drift, although two items (index numbers 8 & 12 in Table 9) did when α was kept at .05.

To obtain the area statistics derived by (Raju, 1988, 1990), the (transformed) α and β estimates obtained from the Lord's (1980) χ^2 analysis were used, constraining all γs to .001.[31]. The values are presented in Table 8, and, as with the analysis using Lord's χ^2, items 8 and 12 are the only ones with significantly different parameter values.

At this point, one could do another DIF/drift item analysis using IRT-LR (Thissen et al., 1993) as it assesses for differences in the γ parameter, but looking at the values in Table 7 there is no need to do so. The only possibly problematic item, with respect to the γ parameter, is number 4, and it can be excluded from the set of linking items, as can the items indexed as number 8 and 12, and still leave 13 items in the linking subset.

The last step is to vertically equate forms LO and LK of the CBASE. As three of the common items of both forms showed slight DIF/drift, they will be excluded from the linking subset and be treated as if they were not common across forms. When the data was run in BILOG-MG, the mean standardized deviation for the θ on form LO was -0.222, with a standard deviation of 1.129. Thus, it appeared that the value obtained from the CTT analysis was an *underestimate* of the true latent score change.

Chapter 5

DISCUSSION

The purpose of this study was to demonstrate the use of IRT methods in assessing the LFE. More specifically, this text demonstrated how IRT methods can be used to determine whether the increase in mean IQ scores across time (i.e., the LFE) is due to a co-occurring increase in intelligence or if it is due to a psychometric artifact. To that end, this study assessed the LFE in both simulated and real test scores via CTT and IRT methods, and, at least for the simulated items under the parameters used for this study, showed IRT's superiority in assessing the LFE.

To reiterate, the simulation study found that CTT methods could not distinguish between a real change in underlying ability and when items (slightly) changed, but ability stayed the same (see Tables 4 and 5). Interestingly, there were two iterations where IRT methods indicated that latent ability in the groups with slight item parameter perturbations was higher than the group with a true increase in latent ability, a phenomenon that merits future inquiry. Still, for the vast majority of the iterations, IRT methods were able to discriminate item perturbations from latent ability increase, a finding without replicate with the CTT methods.

In the study of the CBASE Mathematics section, the study found that there were only a few items that, possibly, exhibited DIF/drift. More importantly, it found that there was a reverse LFE of the magnitude of .222 standard deviations, which, on the College BASE metric (μ: 300, σ: 65), means that the average Mathematics ability decreased from 300 to approximately 286 from 1996 to 2001.[32] This finding could be a product of the sample used, as it was made up of education majors; more likely, though, it is indicative a "newer" trend in cognitive abilities, namely a reverse LFE (Sundet et al., 2004; Teasdale & Owen, in press; for a more historical perspective, see Lynn, 1997). It goes without saying that this is in need of much more systematic investigation.

Significance of Study

This particular study is a significant contribution to the LFE literature in two ways. First, to the author's knowledge, it is the first instance of using methods derived from IRT to assess the LFE. Second, this study showed that under a given set of "true" parameters, methods derived from CTT are impotent in discriminating between psychometric artifacts and a real increase in the latent trait (i.e., cognitive ability). While other authors have surmised that the LFE is merely a psychometric artifact (e.g., Brand, Freshwater, & Dockrell, 1989; Brand, 1996; Rodgers, 1999), they have relied on analyses from CTT. Combining the two major two findings of this dissertation, this then leads to the conclusion that future research looking at the LFE needs to assess whether IRT-based methods could/should be used analyze the data.

Limitations

As with any simulation study, only a small sample of the parameter domain can be sampled, and this limitation holds for this study as well, especially as the α parameter was not varied, and only one increase in the γ was used. Future studies need to directly investigate the influence of the α parameter with respect to the LFE, as well as model how various fluctuations in the γ parameter evidence themselves. In addition, given that the CBASE portion of this study (Study 2) found a reverse LFE (i.e., a dysgenic trend; Burt, 1952; Lynn, 1997), future simulation studies need to specifically model a decrease in cognitive ability; particulary useful would be studies that model a decrease in cognitive abilities, with both mean difficulty decreasing and guessing increasing to see if CTT methods still evidence the LFE.

The simulation portion of this study used a forced choice response format. While it is the format used in many cogntive ability tests (e.g., Raven's Matrices), is not the response format for all given cognitive abilities instruments (e.g., Wechsler scales). Thus, free recall formats need investigation in the future, wherein there is no guessing modelled in the response model. This study simulated its items under an underlying IRT model, as used in IRTGEN (Whittaker et al., 2003). Future studies should address whether the similar findings are obtained using items simulated from CTT models. Last, it would be

beneficial to run similar simulation studies using more iterations, say 500 or 1000.

With regard to the CBASE study, the limitations were mentioned in the text, namely that only a subset of items were used and only a select group of respondents' data was analyzed. Future studies with the CBASE should address whether the (reverse) LFE is extant in the other domains. Likewise, it would be beneficial to examine other samples who took the CBASE over multiple years.

Concerning the manuscript in general, the most significant limitation is that item responses were modelled using the 3PL IRT model. While this model has found much support in the literature, especially in the achievement and cognitive abilities domains, it is not the only model available, and future studies could address simulation (and subsequent analysis) of data using other models.

Implications & Future Research

Although this simulation study gave evidence as to the injudicious practice of using CTT methods to assess the LFE, this finding should not be over interpreted. More specifically, this finding gives no evidence as to whether the LFE (or its reverse) is actually occurring or whether it really is a psychometric artifact. While the second study using the CBASE data provides evidence for the former, more research is needed before more definitive statements can be made, especially with respect to the effect's direction. What the simulation study did show, however, was that CTT methods appear to be unable to distinguish between an actual change in underlying ability and a change in the items that measure said ability, at least under certain conditions. When the effect is due to a true underlying change in abilities, though, the two assessments should be similar, as was found with the CBASE study.

Future studies will need to address the effect of IRT model parametrization on LFE analysis. While this paper consistently used, and simulated data under, the 3PL model, it is by no means the only one available (for examples, see du Toit, 2003; van der Linden & Hambleton, 1997). In addition, the reason for the two anomalous simulation iterations, where the IRT analysis indicated a psychometric artifact group's latent average ability was higher than the group with a true increase, needs further investigation. Moreover,

although the DIF/drift tests used were the ones supported by the literature (Donoghue & Isham, 1998), it might be beneficial to incorporate other measures in future analyses.

With the *en masse* distribution of personal computers and the user-friendly software available, IRT analyses of LFE data is not much more difficult, or least not much more time consuming, the the more traditional CTT analysis. Thus, it is hoped that IRT analysis will begin to make its way into this field. Of course, the one caveat in using IRT methods is that item-level data is needed instead of the more common summed and standardized scores, but with some pre-planning, this should not be too difficult to collect.

References

Baker, F. B. (1992). *Item Response Theory: Parameter estimation techniques*. New York: Mercel Dekker.

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

Baker, F. B., & Kim, S.-H. (2004). *Item Response Theory: Parameter estimation techniques*.

Beguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some fit analysis of multi-dimensional IRT models. *Psychometrika*, *66*, 471–488.

Binet, A., & Simon, T. (1905). Methodes nouvelles pour le diagnostique du niveau intellectuel des anormaux. *L'Annee Psychologique*, *11*, 245–336.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. Lord & R. Novick (Eds.), *Statistical theories of mantal test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

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. (2003). A brief history of Item Response Theory. In M. du Toit (Ed.), *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFAT* (pp. 830–859). Lincolnwood, IL: Scientific Software International.

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.

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. (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.

Brown, W. (1910). Some experimental results in the correlaiton of mental abilities. *British Journal of Psychology*, *3*, 296–322.

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

Burt, C. (1957). The distribution of intelligence. *British Journal of Psychology*, *48*, 161–175.

Burt, C. (1963). Is intelligence distributed normally? *British Journal of Statistical Psychology*, *16*, 175–190.

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

Carroll, J. B. (1993). *Human cognitive abilities*. Cambridge University: New York.

Ceci, S. J., Rosenblum, T. B., & Kumpf, M. (1998). The shrinking gap between high- and low-scoring groups: Current trends and possible causes. In U. Neisser (Ed.), *The rising curve* (pp. 287–302). Washington, DC: American Psychological Association.

Cocodia, E. A., Kim, J.-S., Shin, H.-S., Kim, J.-W., Ee, J., Wee, M. S. W., et al. (2003). Evidence that rising population intelligence is impacting in formal education. *Personality and Individual Differences*, *35*, 797–810.

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.

Cook, L. L., & Eignor, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice*, *10*, 191–199.

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

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*, 297–334.

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.

Donoghue, J. R., & Isham, S. P. (1998). A comparison of procedures to detect item parameter drift. *Applied Psychological Measaurement*, *22*, 33–51.

du Toit, M. (Ed.). (2003). *IRT from SSI: BILOG-MG, MULTILOG, PARSCALE, TESTFAT.* Lincolnwood, IL: Scientific Software International.

Embretson, S., & Reese, S. (2002). *IRT 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.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philisophical Transactions of the Royal Society, Series A*, 222–326.

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. (1990). Massive IQ gains on the Scottish WISC: Evidence against Brand et al.'s hypothesis. *Irish Journal of Psychology*, *11*, 41–51.

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

Fox, J. P., & Glas, C. A. W. (2005). Bayesian modification indices for IRT models. *Statistica Neerlandica*, *59*, 95–106.

Haley, D. C. (1952). *Estimation of the dosage mortality relationship when the does is subject to error* (Technical Report No. 15). Stanford, CA: Stanford University Applied Mathematics and Statistics Laboratory.

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.

Herrnstein, R. J., & Murray, C. A. (1994). *The bell curve: Intelligence and class structure in American life*. New York: Free Press.

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

Howard, R. W. (2001). Searching the real world for signs of rising population intelligence. *Personality and Individual Differences*, *30*, 1039–1058.

Hunt, E. B. (1995). *Will we be smart enough? A cogntive analysis of the coming workforce*. New York: Russell Sage.

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

Juve, J. A. (2004). *Assessing differential item functioning and item parameter drift in the College Basic Academic Subjects Examination*. Unpublished doctoral dissertation, University of Missouri-Columbia.

Kanaya, T. (2004). *Age differences in IQ trends: Unpacking the Flynn Effect*. Unpublished doctoral dissertation, Cornell University.

Kane, H., & Oakland, T. (2000). Secular declines in Spearman's "g": Some evidence from the United States. *Journal of Genetic Psychology*, *161*, 337–345.

Kim, S.-H., & Cohen, A. S. (1991). A comparison of two areas measures for detecting differential item functioning. *Applied Psychological Measurement*, *15*, 269–278.

Kolen, M., & Brennan, R. L. (1995). *Test equating methods and practices*. New York: Springer.

Lord, F. (1968). An analysis of the verbal Sscholastic Achivement Test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement*, *28*, 989–1020.

Lord, F. (1980). *Applications of Item Response Theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Lord, F., & Novick. (1968). *Statistical theories of mental test scores*. London: Addison-Wesley.

Lynn, R. (1982). IQ in Japan and the United States shows a growing disparity. *Nature*, *306*, 291–292.

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.

Lynn, R. (1997). *Dysgenics: Genetic deterioration in modern populations*. Westport, CT: Praeger.

Lynn, R., & Hampson, S. (1986). The rise of national intelligence: Evidence from Britain, Japan, and the USA. *Personality and Individual Differences*, *7*, 23–32.

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

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, *51*, 177–195.

Murray, C. (1999, July). *The secular increase in IQ and longitudinal changes in the magnitutde of the Black-White difference: Evidence from the NLSY.* Paper presented at the meeting of the Behavior Genetics Association, Vancouver, BC.

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.

Nandakumar, R., & Ackerman, T. (2004). Test modeling. In D. Kaplan (Ed.), *The sage handbook of quantitative methodology for the social sciences* (pp. 93–105). Thousand Oaks, CA: Sage.

Neisser, U. (Ed.). (1998). *The rising curve: Long term gains in IQ and related measures.* Washington, DC: American Psychological Association.

Nettelbeck, T., & Wilson, C. (2004). The Flynn Effect: Smarter not faster. *Intelligence*, *32*, 85–93.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Olson, D. H., Russell, C. S., & Sprenkle, D. H. (1989). *Circumplex Model: Systemic assessment and treatment of families.* New York: Haworth Press.

Orlando, M., & Thissen, D. (2000). Likelihood-based item-fit indices for dichotomous Item Response Theory Models. *Applied Psychological Measurement*, *24*(1), 50–64.

Osterlind, S. J. (2005). *Psychometric methods: Development and application of modern mental measures.* New Jersey: Prentice Hall.

Osterlind, S. J., & Beaujean, A. A. (2005, April). *Structural integrity in a Measure of General Education College Achievement.* Paper presentation at the annual meeting of the American Educational Research Association, Montréal, Canada.

Osterlind, S. J., & Merz, W. R. (1990). *College base technical manual.* Columbia, MO: University of Missouri.

Osterlind, S. J., Sheng, Y., Juve, J., Beaujean, A. A., & Nagel, T. (in preparation). *College BASE technical manual* (2nd ed.). Columbia, MO: University of Missouri.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*, 495–502.

Raju, N. S. (1990). Determinig the signifigance of estimated signed and unsigned area between two item response functions. *Applied Psychological Measurement*, *14*, 197–207.

Rao, C. R. (1973). *Linear statistical inference and its applications.* New York: John Wiley.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: University of Chicago. (Original work published 1960)

Rodgers, J. L. (1999). A critique of the Flynn Effect: Massive IQ gains, methodological artifacts, or both? *Intelligence*, *26*, 337–356.

Rushton, J. P. (1997). Review essay. [Review of the books The Bell Curve Debate, The Bell Curve Wars, Final Solutions, The Mismeasure of Man, The Nazi Connection, The Race Gallery, and The Science and Politics of Racial Research]. *Society*, *34*, 78–82.

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.

Rushton, J. P. (2003). Race differences in *g* and the "Jensen effect". In H. Nyborg (Ed.), *The scientific study of general intelligence: Tribute to Arthur R. Jensen* (pp. 147–186). Amsterdam: Pergamon.

Rushton, J. P., & Jensen, A. R. (2003). African-White IQ differences from Zimbabwe on the Wechsler Intelligence Scale for Children-Revised are mainly on the *g* factor. *Personality and Individual Differences*, *34*, 177–183.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, *11*, 235–294.

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.

Schabenberger, O., & Pierce, F. J. (2002). *Contemporary statistical models for the plant and soil sciences*. Boca Raton, FL.

Scott, R., Bengston, H., & Gao, P. (1998). The Flynn Effect: Does it apply to academic achievement? *Mankind Quarterly*, *39*, 109–118.

Smith, S. (1942). Language and nonverbal test performance of racial groups in Honolulu before and after a 14 year interval. *Journal of General Psychology*, *26*, 51–93.

Spearman, C. (1904). "General intelligence": Objectively defined and measured. *American Journal of Psychology*, *15*, 201–292.

Spearman, C. (1910). Correlation calculated with faulty data. *British Journal of Psychology*, *3*, 271–295.

Stark, S., Chernyshenko, S., Chuah, D., Lee, W., & Wadington, P. (2001). *ITERLINK*. [Computer Software]. Urbana-Champaign, IL: University of Illinois.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1–49). New York: John Wiley.

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. (1989). Continuing secular increases in intelligence and a stable prevelance of high intelligence levels. *Intelligence*, *13*, 255–262.

Teasdale, T. W., & Owen, D. R. (in press). A long-term rise and recent decline in intelligence test performance: The Flynn Effect in reverse. *Personality and Individual Differences*.

Thissen, D., & Orlando, M. (2001). Item Response Theory for items scores in two categories. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Mahwah, NJ: Lawrence Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67–122). Mahwah, NJ: Lawrence Erlbaum.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. *Journal of Educational Psychology*, 433–451.

Truscott, S. D., & Frank, A. J. (2001). Does the Flynn Effect affect IQ scores of stuents classified as LD? *Journal of School Psychology*, *59*, 319–334.

Tuddenham, R. D. (1948). Soldier intelligence in World Wars I and II. *American Psychologist*, *3*, 54–56.

van der Linden, W. J., & Hambleton, R. (Eds.). (1997). *Handbook of modern item response theory*. New York: Springer.

Whittaker, T. A., Fitzpatrick, S. J., Williams, N. J., & Dodd, B. G. (2003). IRTGEN: A SAS macro program to generate known trait scores and item responses for commonly used item response theory models. *Applied Psychological Measurement*, *27*, 299–300.

Yu, C. H., & Osborn Popp, S. E. (2005). Test equating by common items and common subjects: Concepts and applications. *Practical Assessment Research & Evaluation*, *10*. 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]. Chicago: Scientific Software International.

Endnotes

[1]For definitions, see Appendix A.

[2]*Note.* Sir Cyril Burt (1952) first hypothesized that the rise in psychometric IQ was due to a psychometric artifact instead of a real increase in intelligence. At the time, work in psychometrics had not developed enough to allow him to fully test the hypothesis, so he had to use ancillary evidence to support his argument. Later, Brand, Freshwater, and Dockrell (1989; cf. Brand, 1990, 1996) analyzed item data and concluded much as Burt did, namely that the effect appears to be an artifact, but the methods are those arising post-hoc from Classical Test Theory (for a critique, see Flynn, 1990)

[3]In other texts, this effect is sometimes referred to as simply the *Flynn Effect*. This is (mainly) due to the fact the Herrnstein and Murray (1994) coined the term in their widely-read book on the importance of IQ in determining life outcomes. In actuality, both Richard Lynn and James Flynn deserve credit for the finding, as (Lynn, 1982) first brought the effect to the world's attention, even though the effect was seen over a half-century earlier (Smith, 1942; Tuddenham, 1948). This text will follow the recommendation made by Rushton (1997) and keep the effect entitled *Lynn-Flynn Effect*.

4

Proof.

$$\begin{aligned}\text{COV}[X_i, T_i] &= \text{COV}[(T_i + \varepsilon), T_i] \\ &= \text{VAR}[T_i] + \text{COV}[\varepsilon, T_i] \\ &= \text{VAR}[T_i]\end{aligned}$$

□

[5]Even though T_i is unobservable, as $\varepsilon \to 0$, $\underset{r_{xx} \in [0,1]}{\uparrow} r_{xx}$ so that, for a perfectly reliable test (i.e., $r_{xx} = 1$), $X_i = T_i$.

[6]One need not stop at just 1 bifurcation. The formula can be generalized to m differnt splits, each having $\frac{n}{m}$ items. Becasue the focus of this paper is not CTT, no more detail will be presented.

[7]In practice, this turns out to just divinding the sum of each item's variance by the test's total varaiance. Then, take one minus this product and multiply by $\frac{n}{n-1}$.

[8]The equivalency is sometimes called having τ-equivalent tests because the Greek letter τ replaces T_i in the population-based formula.

[9]There is also a weaker local independence assumption; for more information, see pp. 94-95 of Nandakumar and Ackerman (2004).

[10]There are multiple other terms used to describe this parameter. While this makes sense when dealing with such things as rating scales, it does not necessarily carry over into

achievement and cognitive ability assessment. Consequently, this text will keep with tradition and refer to it as the *ability* parameter.

[11]While the two functions are very similar, they are not the exact same, especially in the asymptote. Nonetheless, Lord (1980) writes, "The two models $\cdots$ [when corrected] give very similar results for practical work" (p. 14).

[12]For a more philosophical treatise, see pages 90-91 of Thissen and Orlando (2001).

[13]Baker (2001) writes that, technically, α_i is *not* the slope of the ICC, but rather $\frac{\alpha_i}{4}$ is the slope at $\theta_j = \beta_i$. Nonetheless, $\alpha_i \propto \frac{\alpha_i}{4}$, so, conceptually, α_i can be thought of as a close approximation of the ICC slope at β_i. In theory, $-\infty \leq \alpha_i \leq \infty$, but, in practice $\alpha_i \leq |3|$.

As a further footnote, if the items on a test are scored as correct = 1 and incorrect = 0, then for any $\alpha_i \leq 0$ the ICC implies that the probability of correctly answering a given item *decreases* as ability increases. Or, to quote Baker (2001), "This tells you that something is wrong with the item: Either it is poorly written or there is some misinformation prevalent among the high ability students" (p. 33).

14

Proof.

$$
\begin{aligned}
f(x_i; \boldsymbol{\kappa}_i, \theta_j) &= \frac{1}{1 + e^{-\alpha_i M(\theta_j - \beta_i)}}, \text{ and is monotonically increasing from 0 to 1} \\
\frac{\partial f(\cdot)}{\partial \alpha_i} &= \frac{M(\theta_j - \beta_i) e^{-\alpha_i M(\theta_j - \beta_i)}}{(1 + e^{-\alpha_i M(\theta_j - \beta_i)})^2} \\
0 &= \frac{M(\theta_j - \beta_i) e^{-\alpha_i M(\theta_j - \beta_i)}}{(1 + e^{-\alpha_i M(\theta_j - \beta_i)})^2} \\
&= \frac{M(\theta_j - \beta_i)}{2 + e^{-\alpha_i M(\theta_j - \beta_i)} + e^{\alpha_i M(\theta_j - \beta_i)}} \Leftrightarrow 0 = (\theta_j - \beta_i)
\end{aligned}
$$

□

[15]Lord (1968) opines that γ_i should *not* be interpreted as a guessing parameter, but rather the lower bound for the ICC. He is correct, but due to its common usage, γ_i will continue to be entitled the guessing parameter in this manuscript.

[16]This parameterization is the same (after a little algebraic manipulation) as seen in other books (e.g., Lord, 1980), namely:

$$
f(x_i; \boldsymbol{\kappa}_i, \theta_j) = \gamma_i + \frac{1 - \gamma_i}{1 + e^{-\alpha_i M(\theta_j - \beta_i)}} \tag{31}
$$

17

Proof 1.

$$
f(x_i; \boldsymbol{\kappa}_i, \theta_j) \propto h(\gamma_i) + \alpha(\theta - \beta_i)
$$

where $h(\gamma_i)$ is the modification due to a guessing parameter.

Choose any $\delta \in \mathbb{R}$ and add it to both θ and β_i. The new item response function, $f^{(1)}$ is then

$$\begin{aligned} f^{(1)}(x_i; \boldsymbol{\kappa}_i, \theta_j) &\propto h(\gamma_i) + \alpha([\delta + \theta] - [\delta + \beta_i]) \\ &\propto h(\gamma_i) + \alpha(\delta + \theta - \delta - \beta_i) \\ &\propto h(\gamma_i) + \alpha(\delta - \delta + \theta - \beta_i) \end{aligned}$$

$$\begin{aligned} &\propto h(\gamma_i) + \alpha(0 + \theta - \beta_i) \\ &= f(x_i; \boldsymbol{\kappa}_i, \theta_j) \end{aligned}$$

□

Proof 2. Let $f(x_i; \boldsymbol{\kappa}_i, \theta_j)$ be defined as in *Proof 1.* Choose any $\delta \in \mathbb{R}$ and multiply it to both θ and β_i and multiply α_i by δ^{-1}. The new item response function, $f^{(2)}$, is then

$$\begin{aligned} f^{(2)}(x_i; \boldsymbol{\kappa}_i, \theta_j) &\propto h(\gamma_i) + \delta^{-1}\alpha([\delta]\theta - [\delta]\beta_i) \\ &\propto h(\gamma_i) + \delta^{-1}\alpha(\delta[\theta - \beta_i]) \\ &\propto h(\gamma_i) + \delta^{-1}\alpha(\delta)(\theta - \beta_i) \\ &\propto h(\gamma_i) + \delta^{-1}(\delta)\alpha(\theta - \beta_i) \\ &\propto h(\gamma_i) + 1 \times \alpha(\theta - \beta_i) \\ &= f(x_i; \boldsymbol{\kappa}_i, \theta_j) \end{aligned}$$

□

[18]The study of DIF originally developed in response to the plethora of criticisms targeted against standardized assessment, namely that it was biased against certain ethnic groups (Holland & Wainer, 1993). The area of bias research, whether in psychometrics or outside the field, tends to be highly political in nature, and need not concern this text. For the purposes here, it is sufficient to say that DIF is necessary, but not sufficient, for bias to exist (Camilli & Shepard, 1994).

[19]There are CTT parameterizations for DIF studies as well, but they need not concern this manuscript. For more information on them, see Camilli and Shepard (1994) and Holland and Wainer (1993).

[20]Lord (1980) did not consider the γ_i parameter, as he wrote they should be constrained to be equal across groups.

[21]Thissen et al. (1993) cite problems with Lord's procedure, as do Hambleton et al. (1991); both authors, though, keep Lord's (1980) definition of DIF.

[22]More specifically, $\boldsymbol{\Sigma}^{-1}$ is computed as follows. The item information matrix is computed for each group, then inverted. Finally, the (co)variance matrix of each group is added, which yields the (co)variance matrix of the difference between the estimates.

[23]Technically, using an IRT framework means that items, if they measure the same ability, do not need equating, but rather a re-scaling. Nonetheless, due to the popularity of the moniker equating, the terms are used interchangeably.

[24]*Note.* α will be constrained to be a constant, 1, for this study, as when it gets very low or very high, it interacts with the β and γ parameters to produce (random) deviations from the expected effects of altering those same parameters (Baker, 2001). Future studies will need to address the impact of a random α parameter.

[25]The CBASE is used in some states as a proficiency examination for teachers-in-training, whereby passing at a certain level is required before they can obtain a teaching license. The CBASE is given for many other reasons, too; for more information, see Osterlind et al. (in preparation).

[26]As the items were dichotomous, the summed score is simply the number correct score.

[27]A planned contrast between the $\overline{X}_\beta$ and $\overline{X}_\theta$ groups was not significant (t=1.658, df=396, p=.098), nor was a planned contrast between the $\overline{X}_\gamma$ and $\overline{X}_\theta$ group (t=-.801, df=396, p=.424). A planned contrast between the $\overline{X}_{\beta,\gamma}$ group and the $\overline{X}_\theta$ group was significant (t=-7.123, df=396, $p < .001$).

[28]A planned contrast between the $\overline{X}_\beta$ and $\overline{X}_\theta$ groups was significant (t=13.655, df=297, $p <$.000), as was a planned contrast between the $\overline{X}_\gamma$ and $\overline{X}_\theta$ group (t=14.551, df=297, $p <$.000).

[29]*Note.* All of the anomalous cases were checked to make sure the data was both stored and imputed into BILOG-MG correctly. In addition, different models (e.g., 1PL, 2PL) were run for the data. Neither action produced any (significant) changes in the estimates.

[30]The estimated values for the location and scale parameters are -.216487 and 1.043299, respectively. Thus to get form LO's α and β parameters on LK's "metric" use the following relationships:

$$\alpha_{LK} = \frac{\alpha_{LO}}{A}, \tag{32a}$$

$$\beta_{LK} = A\beta_{LO} + K \tag{32b}$$

where K and A are the location and scale values, respectively.

[31]From Table 7 it can be seen that, with the possible exception of the item indexed as number 4, no γ parameter is (much) different from this value.

[32]One could make the argument that since the mathematics tasks loaded so highly on the g factor (see Table 2), that it was general intelligence that decreased over the time period.

Appendix A

DEFINITIONS

Classical Test Theory (CTT). Also known as true-score theory, it is the realm of psychometrics concerned with modelling an examinee's latent ability on an entire test. Item procedures in CTT are *post-hoc* and heavily sample-dependent.

Item Response Theory (IRT). A modern extension of CTT, where the primary concern is modelling item responses. Item procedures are built into the theory, thus are independent of a given sample.

Examinee Ability (θ). The latent examinee variable in IRT that is equivalent to the true score in CTT. Its domain is $-\infty$ to $+\infty$.

Item Difficulty (β_i). The item location parameter in IRT models, expressed in θ units, that indicates the point on the θ scale at which the probability of a correct response is .50. Its domain is $-\infty$ to $+\infty$.

Item Discrimination (α_i). The item scale parameter in IRT models, that indexes the discrimination ability of an item. Its domain is $-\infty$ to $+\infty$.

Guessing parameter (γ_i). The lower asymptote value for an item where the probability of answering an item correctly for low ability examinees does not reach 0. Its domain is 0 to 1, but most values do not exceed .35.

Lynn-Flynn Effect. The observation, made by both Richard Lynn and James Flynn, that psychometric IQ scores have increased over time, at least since the 1940s.

Appendix B

COMPUTER CODE

SAS Code

Code for ith iteration of generating random Item Parameters and IRTGEN for all populations: (a) $\alpha = 1$; $\beta \sim N(0,1)$; $\gamma = .1$; (b) $\alpha = 1$; $\beta \sim N(-.1,1)$; $\gamma = .1$; (c) $\alpha = 1$; $\beta \sim N(0,1)$; $\gamma = .15$; and (d)$\alpha = 1$; $\beta \sim N(-.1,1)$; $\gamma = .15$. For the first four generations, $\theta \sim N(0,1)$, and the for the fifth generation, $\theta \sim N(.14,1)$.

```
title 'ioriginal';
DATA original;
KEEP A B C;
do i=1 to 60;
A = 1;
B = NORMAL(0);
C = .1;
output;
end;
ODS HTML body='C:\ ...\ i OriginalParameters.html';
proc print;
run;
ODS HTML close;
%IRTGEN(MODEL=L3, DATA=original, OUT=oneoriginal, NI=60, NE=1000);
ODS HTML body='C:\ ...\ ioriginal.html';
proc print;
run;
ODS HTML close;
quit;

LIBNAME mylib 'C:\ ...\ ParameterFiles';
DATA mylib.THETA;
set oneoriginal;
keep THETA;
run;
```

```
title 'ibeta';
DATA beta;
KEEP A B C;
do i=1 to 60;
A = 1;
B = NORMAL(0) - .1;
C = .1;
output;
end;
ODS HTML body='C:\ ...\ iBetaParameters.html';
proc print;
run;
ODS HTML close;
%include 'C:\ ...\ THETA';
%IRTGENTheta(MODEL=L3, DATA=beta, OUT=onebeta, NI=60, NE=1000);
ODS HTML body='C:\ ...\ ibeta.html';
proc print;
run;
ODS HTML close;
quit;
```

```
title 'igamma';
DATA gamma;
set original;
keep A B C;
C = .15;
ODS HTML body='C:\ ...\ iGammaParameters.html';
proc print;
run;
ODS HTML close;
%include 'C:\ ...\ THETA';
%IRTGENTheta(MODEL=L3, DATA=gamma, OUT=onegamma, NI=60, NE=1000);
ODS HTML body='C:\ ...\ igamma.html';
proc print;
run;
ODS HTML close;
quit;
```

```
title 'ibetagamma';
DATA betagamma;
set beta;
KEEP A B C;
C = .15;
ODS HTML body='C:\ ...\ iBetaGammaParameters.html';
proc print;
run;
ODS HTML close;
%include 'C:\ ...\ THETA';
%IRTGENTheta(MODEL=L3, DATA=betagamma, OUT=onebetagamma, NI=60, NE=1000);
ODS HTML body='C:\ ...\ ibetagamma.html';
proc print;
run;
ODS HTML close;
quit;
```

```
title i thetaincrease';
data thetaincrease;
set original;
keep A B C;
%include 'C: ... THETA.sas';
%IRTGENTheta(MODEL=L3, DATA=thetaincrease, OUT=origthetain, NI=60,
NE=1000);
ODS HTML body='C:\ ...\ ithetaincrease.html';
proc print;
run;
ODS HTML close;
quit;
```

SAS code for the THETA program; $\theta \sim N(.14, 1)$

```
DATA THETA;
keep THETA;
Do i=1 to 1000;
THETA=Normal(0)+.14;
OUTPUT;
END;
RUN;
```

BILOG-MG Code

```
Lynn-Flynn Effect Program
>COMMENTS
Groups i and i'.
>GLOBAL DFName = 'C: \ ...\ [filename].dat',
NPARM=3, SAVE;
>SAVE SCORE='output.SCO'; >LENGTH NITEMS=115;
>INPUT NTOT=115, NGROUPS=2, NIDCH=4;
>ITEMS INUM=(1(1)115);
>TEST TNAME=SIMUL;
>GROUP1 GNAME='Original', LENGTH=60, INUM=(1(1)60);
>GROUP2 GNAME='comparison', LENGTH=60, INUM=(56(1)115);
(4A1,I1,115A1)
>CALIB NQPT=51, NORMAL, CYCLE=30, TPRIOR, REFERENCE=1;
>SCORE METHOD=2, IDIST=3, NOPRINT, RSCTYPE=3;
```

Program to obtain item parameters to get Lord's χ^2 values (Note. Program needs to be run separately for each CBASE form.

```
>COMMENTS
>GLOBAL DFName = 'C: \ ...\ [filename].dat',
NPArm = 3,
LOGistic,
SAVe;
>SAVE CALib = 'Lord.CAL',
PARm = 'Lord.PAR',
COVariance = 'Lord1.COV';
>LENGTH NITems = (16);
>INPUT NTOtal = 16,
NALt = 1000,
NIDchar = 4;
>ITEMS ;
>TEST1 TNAme = 'TEST0001',
INUmber = (1(1)16);
(4A1, 16A1)
>CALIB ACCel = 1.0000;
>SCORE ;
```

Table 1

Sample of Studies Examining the Lynn-Flynn Effect

Study	Instrument(s)	Dates (or time between) testing	N	Differences
(Bolen et al., 1995) 1	WISC-R (FSIQ)	2.5-3 years	61	5.20
	WISC-III			
(Bolen et al., 1995) 2	WISC-R (PIQ)	2.5-3 years	61	9.21
	WISC-III			
(Bolen et al., 1995) 3	WISC-R (VIQ)	2.5-3 years	61	7.95
	WISC-III			
(Kanaya, 2004) 1	WISC	$\sim$ 3 years	436	3.92
	WISC-R			
(Kanaya, 2004) 2	WISC-R	$\sim$ 3 years	598	6.15
	WISC-III			
(Must et al., 2003) Grade 3	IEA	1994/1999	522	2.4[2]
(Must et al., 2003) Grade 8	IEA	1994/1999	522	4.4
(Sanborn et al., 2003) 1	WJ-R	8 weeks	40	2.03[1]
	WJ-III			
(Sanborn et al., 2003) 2	WISC-R	3 years	169	8.56[1]
	WISC-III			
(Daley et al., 2003)	RCPM	1984/1998	C1: 118	4.5
		" "	C2: 537	
(Daley et al., 2003) 1	VMT	" "	" "	2.57
(Daley et al., 2003) 2	DS	" "	" "	.39

Continued on next page

Sample of Studies Examining the Lynn-Flynn Effect– continued from previous page

Study	Instrument(s)	Dates (or time between) testing	N	Differences
(Colom et al., 2005)	PGT	1970/1999	C1: 459	9.7[3]
		" "	C2: 275	
(Colom et al., 2005)	PGT	" "	" "	6.5[4]
(Colom et al., 2005)	PGT	" "	" "	2.5[5]
(Truscott & Frank, 2001) 1	WISC-R (FSIQ) WISC-III	$\sim$ 3 years	171	6.93[6]
(Truscott & Frank, 2001) 2	WISC-R (VIQ) WISC-III	$\sim$ 3 years	171	6.26[6]
(Truscott & Frank, 2001) 3	WISC-R (PIQ) WISC-III	$\sim$ 3 years	171	7.31[6]

1. In original study, the authors split the analysis by IQ. As the LFE was evidenced across all IQ levels, and there was no IQ $\times$ test interaction, only results for the complete sample are given in the table.

2. In original study, the authors report subscale results as well, which all show the same pattern as the summery scores in table.

3. These are IQ points, units on the PGT

4. Raw PGT points for the bottom 50%

5. Raw PGT points for the top 50%

6. These are the difference in mean IQ from the WISC-R to WISC-III.

Truscott and Frank (2001) uses a different metric in the article, but due to its non-common nature, only IQ differences are presented here.

C1: Cohort 1; C2: Cohort 2;

Continued on next page

Sample of Studies Examining the Lynn-Flynn Effect– continued from previous page

Study	Instrument(s)	Dates (or time between) testing	N	Differences

IEA: International Association for the Evaluation of Educational Achievement literacy evaluation

WJ: Woodcock Johnson Test of Cognitive Ability; WISC: Wechsler Intelligence Scale for Children

RCPM: Raven's Colored Progressive Matrices; VMT Verbal Meaning Test; DS: Digit Span

PGT: Pressey's Graphic Test; ng: not given.

Table 2

Factor Loadings for the Mathematics Skills on form LM of the CBASE

Factor	*g*	SS	S	M	E	h^2
Order	1	2	2	2	2	
Skill:						
Geometrical calculations	**.711**	-.046	-.023	.196	-.024	.547
2- & 3-Dimensional figures	**.687**	-.060	-.031	.259	-.031	.545
Equations & Inequalities	**.658**	-.114	-.058	**.491**	-.059	.694
Evaluating Expressions	**.636**	-.098	-.050	**.419**	-.050	.595
Using Statistics	**.770**	.015	.007	-.063	.008	.597
Properties & Notations	**.673**	-.041	-.021	.176	-.021	.486
Practical Applications	**.736**	.006	.003	.026	.003	.542

Note. Salient loadings of variables on common factors are shown in **bold**.

Factor Names: *g*: General Intelligence; SS: Social Studies; S: Science; M: Math; E: English; h^2: Communality.

Taken from S. Osterlind & A. Beaujean (2005, April) *Structural Integrity in a Measure of General Education College Achievement.* Paper presentation at the annual meeting of the American Educational Research Association, Montréal, Canada.

Table 3

Average Reliability Coefficients of the Math portion of form LM on the CBASE

Topic	Grouping	# items	n	α
Math	Subject	56	29661	0.9002
General Mathematics	Cluster	24	32245	0.7638
Practical Applications	Skill	8	33046	0.6122
Properties & Notations	Skill	8	33046	0.4578
Using Statistics	Skill	8	33018	0.547
Algebra	Cluster	16	31630	0.789
Evaluating Expressions	Skill	8	32582	0.6833
Equations & Inequalities	Skill	8	31983	0.6557
Geometry	Cluster	16	30938	0.7946
2- & 3-Dimensional Figures	Skill	8	31778	0.6773
Geometrical Calculations	Skill	8	31204	0.6828

Taken from S. Osterlind & A. Beaujean (2005, April) *Structural Integrity in a Measure of General Education College Achievement.* Paper presentation at the annual meeting of the American Educational Research Association, Montréal, Canada.

Table 4

Simulated Data Score Increases, using CTT Methods

Simulation	θ_O	θ_I	$\overline{X}_\beta$	σ_β	$\overline{X}_\gamma$	σ_γ	$\overline{X}_{\beta,\gamma}$	$\sigma_{\beta,\gamma}$	$\overline{X}_\theta$	σ_θ
1	-0.023	0.076	-0.034	0.992	0.138	0.956	0.109	0.945	0.084	0.949
2	-0.002	0.161	0.005	1.000	0.164	0.941	0.157	0.968	0.175	0.997
3	0.013	0.104	0.072	0.971	0.151	0.960	0.225	0.936	0.094	0.994
4	-0.028	0.148	-0.009	0.989	0.144	0.922	0.128	0.942	0.175	1.010
5	-0.022	0.127	0.370	0.942	0.155	0.952	0.516	0.890	0.165	0.982
6	0.028	0.154	0.281	0.969	0.138	0.964	0.382	0.936	0.114	1.024
7	-0.044	0.101	-0.070	1.003	0.128	0.954	0.086	0.968	0.112	0.997
8	-0.006	0.139	-0.273	0.985	0.134	0.924	-0.112	0.942	0.128	1.006
9	-0.024	0.160	-0.112	0.992	0.126	0.947	0.035	0.927	0.159	0.964
10	0.060	0.155	0.247	1.016	0.132	0.930	0.378	0.955	0.090	0.960
11	0.035	0.156	0.017	0.993	0.119	0.963	0.135	0.940	0.114	1.010
12	0.013	0.160	-0.068	0.938	0.129	0.951	0.067	0.902	0.125	0.991
13	-0.025	0.155	0.097	0.982	0.120	0.960	0.248	0.955	0.155	0.986
14	-0.020	0.180	-0.201	1.036	0.150	0.938	-0.030	0.987	0.192	1.007
15	0.041	0.099	0.166	0.933	0.142	0.943	0.300	0.879	0.069	1.020
16	0.033	0.135	0.191	1.020	0.169	0.945	0.326	0.951	0.115	0.952
17	-0.002	0.140	0.432	1.026	0.198	0.931	0.531	0.973	0.164	1.016
18	-0.020	0.154	0.133	1.010	0.149	0.952	0.255	0.971	0.177	1.000
19	-0.019	0.128	0.180	1.026	0.180	0.949	0.319	0.985	0.146	1.108
20	-0.019	0.149	0.168	1.019	0.125	0.975	0.297	0.977	0.164	1.012
21	0.036	0.153	-0.110	1.006	0.145	0.954	0.047	0.952	0.116	1.007
22	0.001	0.186	0.118	0.968	0.135	0.967	0.249	0.926	0.175	1.024

Continued on next page

Simulated Data Score Increases, using CTT Methods- continued from previous page

Simulation	θ_O	θ_I	$\overline{X}_\beta$	σ_β	$\overline{X}_\gamma$	σ_γ	$\overline{X}_{\beta,\gamma}$	$\sigma_{\beta,\gamma}$	$\overline{X}_\theta$	σ_θ
23	-0.056	0.115	0.030	1.000	0.132	0.976	0.151	0.965	0.136	1.011
24	0.034	0.135	0.068	1.002	0.133	0.966	0.180	0.962	0.079	0.993
25	0.014	0.219	0.291	0.963	0.125	0.968	0.442	0.925	0.213	1.033
26	-0.058	0.072	0.250	1.044	0.180	0.959	0.379	0.999	0.135	1.001
27	0.004	0.161	0.029	1.031	0.143	0.941	0.132	0.993	0.149	0.969
28	0.038	0.146	0.078	1.007	0.139	0.953	0.211	0.954	0.105	0.988
29	-0.033	0.108	-0.046	0.982	0.125	0.967	0.082	0.943	0.133	0.982
30	0.010	0.101	0.253	1.015	0.160	0.956	0.398	0.958	0.079	1.016
31	0.011	0.086	0.031	1.036	0.146	0.958	0.181	0.977	0.080	1.017
32	-0.008	0.087	-0.091	1.021	0.140	0.956	0.051	0.964	0.098	1.009
33	0.030	0.168	0.080	0.987	0.163	0.949	0.206	0.930	0.133	0.989
34	0.004	0.148	0.115	0.949	0.166	0.966	0.248	0.908	0.136	1.036
35	0.014	0.126	-0.205	1.008	0.148	0.947	-0.106	0.984	0.116	0.984
36	0.008	0.162	0.202	1.021	0.135	0.952	0.347	0.947	0.155	1.012
37	0.040	0.139	0.070	0.999	0.155	0.949	0.221	0.973	0.084	0.988
38	0.000	0.152	-0.097	1.029	0.158	0.943	0.088	0.969	0.157	1.010
39	-0.030	0.140	0.006	1.069	0.148	0.964	0.135	1.022	0.138	1.015
40	0.025	0.143	0.008	0.973	0.130	0.943	0.157	0.932	0.112	1.009
41	-0.028	0.166	0.379	1.015	0.131	0.953	0.536	0.949	0.153	1.029
42	-0.019	0.065	0.201	0.993	0.148	0.961	0.342	0.958	0.078	0.941
43	0.014	0.149	0.182	0.982	0.151	0.962	0.318	0.928	0.130	1.060
44	0.025	0.148	0.158	0.957	0.154	0.946	0.272	0.919	0.118	0.974
45	-0.034	0.115	0.236	0.969	0.178	0.951	0.368	0.941	0.167	1.021
46	0.027	0.094	0.180	0.981	0.145	0.950	0.321	0.933	0.082	1.007

Continued on next page

Simulated Data Score Increases, using CTT Methods- continued from previous page

Simulation	θ_O	θ_I	$\overline{X}_\beta$	σ_β	$\overline{X}_\gamma$	σ_γ	$\overline{X}_{\beta,\gamma}$	$\sigma_{\beta,\gamma}$	$\overline{X}_\theta$	σ_θ
47	-0.012	0.200	-0.029	0.996	0.155	0.963	0.126	0.954	0.182	1.028
48	0.028	0.190	0.135	0.975	0.149	0.957	0.272	0.930	0.164	1.038
49	0.010	0.148	0.010	0.988	0.134	0.971	0.182	0.951	0.154	1.059
50	-0.029	0.093	-0.157	0.949	0.143	0.954	0.021	0.895	0.098	1.010
51	0.009	0.121	0.206	0.971	0.149	0.948	0.335	0.929	0.132	1.023
52	-0.027	0.151	0.282	1.048	0.201	0.964	0.438	0.986	0.193	1.002
53	0.067	0.141	-0.024	1.003	0.138	0.944	0.117	0.960	0.061	0.953
54	-0.019	0.166	0.428	0.991	0.155	0.953	0.551	0.945	0.175	0.995
55	-0.017	0.058	-0.013	0.993	0.136	0.954	0.144	0.945	0.043	0.997
56	-0.006	0.113	0.128	1.056	0.148	0.965	0.281	0.995	0.111	1.028
57	-0.042	0.143	0.033	1.039	0.160	0.938	0.193	0.981	0.156	0.957
58	-0.021	0.125	0.379	1.021	0.186	0.968	0.512	0.974	0.170	1.029
59	0.029	0.121	-0.177	1.056	0.150	0.959	-0.043	1.000	0.096	0.983
60	-0.069	0.141	0.195	0.982	0.147	0.992	0.333	0.939	0.214	1.005
61	0.077	0.144	0.061	0.984	0.122	0.955	0.206	0.940	0.066	0.967
62	0.030	0.172	0.308	1.019	0.144	0.969	0.449	0.962	0.137	1.004
63	-0.017	0.185	0.465	1.061	0.178	0.973	0.595	1.012	0.199	0.999
64	0.005	0.116	-0.026	1.022	0.115	0.959	0.106	0.963	0.102	0.992
65	-0.008	0.177	0.234	0.978	0.166	0.950	0.406	0.918	0.198	1.003
66	-0.045	0.130	0.275	1.029	0.151	0.972	0.400	0.970	0.139	0.994
67	-0.027	0.113	0.338	0.999	0.165	0.960	0.467	0.953	0.140	1.011
68	-0.028	0.116	0.150	0.996	0.166	0.951	0.291	0.946	0.148	0.995
69	0.021	0.106	0.106	0.964	0.151	0.939	0.257	0.931	0.103	1.032
70	-0.070	0.134	0.087	1.029	0.169	0.958	0.253	0.975	0.197	0.990

Continued on next page

Simulated Data Score Increases, using CTT Methods- continued from previous page

Simulation	θ_O	θ_I	$\overline{X}_\beta$	σ_β	$\overline{X}_\gamma$	σ_γ	$\overline{X}_{\beta,\gamma}$	$\sigma_{\beta,\gamma}$	$\overline{X}_\theta$	σ_θ
71	0.054	0.215	-0.218	1.003	0.156	0.969	-0.066	0.954	0.177	1.029
72	-0.015	0.185	0.127	1.026	0.174	0.941	0.266	1.001	0.217	0.989
73	0.016	0.068	0.176	0.995	0.125	0.956	0.289	0.927	0.047	1.024
74	-0.007	0.130	0.283	0.969	0.132	0.936	0.405	0.924	0.114	0.935
75	0.007	0.170	0.021	0.975	0.139	0.945	0.152	0.942	0.161	0.987
76	0.033	0.149	0.214	1.011	0.176	0.949	0.347	0.954	0.123	0.965
77	-0.063	0.123	0.027	0.955	0.142	0.929	0.142	0.904	0.173	0.956
78	0.015	0.139	0.164	0.974	0.131	0.969	0.288	0.920	0.127	1.002
79	-0.026	0.126	0.233	0.973	0.143	0.965	0.369	0.955	0.140	0.983
80	0.006	0.169	-0.010	1.053	0.141	0.950	0.157	0.980	0.138	0.973
81	0.047	0.143	0.040	1.022	0.148	0.972	0.199	0.977	0.089	0.975
82	-0.033	0.165	0.111	0.987	0.161	0.940	0.276	0.899	0.189	0.989
83	-0.028	0.130	0.060	0.985	0.137	0.953	0.198	0.941	0.163	1.000
84	0.048	0.209	0.241	0.917	0.127	0.971	0.351	0.882	0.143	1.049
85	-0.023	0.132	0.315	0.998	0.152	0.944	0.459	0.946	0.140	1.002
86	-0.021	0.170	0.085	0.987	0.145	0.958	0.240	0.943	0.186	0.992
87	0.008	0.140	0.388	0.999	0.155	0.943	0.514	0.945	0.130	1.014
88	0.025	0.117	0.059	0.965	0.160	0.951	0.212	0.890	0.096	1.016
89	0.008	0.117	0.050	1.014	0.151	0.940	0.184	0.979	0.114	0.980
90	-0.025	0.153	-0.042	0.994	0.115	0.945	0.129	0.940	0.158	1.002
91	0.012	0.105	-0.062	0.993	0.138	0.939	0.079	0.947	0.081	0.989
92	-0.024	0.113	0.006	0.992	0.166	0.924	0.163	0.944	0.136	1.028
93	0.014	0.152	0.322	1.005	0.171	0.940	0.454	0.952	0.144	1.030
94	0.002	0.213	-0.025	1.010	0.127	0.948	0.126	0.986	0.196	0.957

Continued on next page

Simulated Data Score Increases, using CTT Methods- continued from previous page

Simulation	θ_O	θ_I	$\overline{X}_\beta$	σ_β	$\overline{X}_\gamma$	σ_γ	$\overline{X}_{\beta,\gamma}$	$\sigma_{\beta,\gamma}$	$\overline{X}_\theta$	σ_θ
95	-0.004	0.140	-0.057	0.991	0.142	0.948	0.085	0.957	0.138	1.034
96	0.017	0.134	0.452	0.981	0.150	0.967	0.595	0.950	0.117	1.002
97	-0.006	0.087	0.006	1.042	0.143	0.934	0.166	0.989	0.094	1.007
98	-0.009	0.142	0.239	0.974	0.174	0.936	0.376	0.931	0.124	0.946
99	0.019	0.159	0.105	1.016	0.146	0.939	0.254	0.940	0.122	0.968
100	-0.044	0.148	0.147	1.044	0.141	0.957	0.298	0.986	0.183	1.011
mean	<0.000	0.140	0.117	1.000	0.149	0.953	0.257	0.952	0.135	1.001
SD	0.030	0.033	0.160	0.029	0.017	0.013	0.155	0.028	0.039	0.028

θ_O: Average latent trait of initial sample of simulated examinees, $E[\theta] = 0$;

θ_I Average latent trait of sample of simulated examinees with true latent trait increase, $E[\theta] = .14$;

$\overline{X}_\beta$ & σ_β: Average and variation of standard score increase when original simulated sample took test with (mean) decreased β values, respectively.

$\overline{X}_\gamma$ & σ_γ: Average and variation of standard score increase when original simulated sample took test with increased γ values, respectively.

$\overline{X}_{\beta,\gamma}$ & $\sigma_{\beta,\gamma}$: Average and variation of standard score increase when original simulated sample took test with both (mean) decreased β values and increased γ values, respectively.

$\overline{X}_\theta$ & σ_θ: Average and variation of standard score increase when simulated sample with increased latent ability took original test, respectively.

Table 5

Simulated Data Score Increases, using IRT Methods

Simulation	θ_O	θ_I	$\overline{X}_\beta$	σ_β	$\overline{X}_\gamma$	σ_γ	$\overline{X}_\theta$	σ_θ	$(\theta_O - \theta_I)$ $-\overline{X}_\theta$
1	-0.023	0.076	-0.003	1.004	0.005	0.968	0.073	0.949	0.026
2	-0.002	0.161	-0.002	0.993	0.004	0.998	0.178	0.985	-0.015
3	0.013	0.104	-0.001	0.988	<0.00	0.975	0.05	0.94	0.041
4	-0.028	0.148	-0.001	0.999	<0.00	0.994	0.106	1.043	0.069
5	-0.022	0.127	0.013	0.986	0.006	0.984	0.145	0.929	0.003
6*	0.028	0.154	0.003	0.983	0.001	0.985	0.046	1.057	0.079
7	-0.044	0.101	-0.001	0.994	<0.00	0.987	0.121	1.015	0.024
8	-0.006	0.139	-0.011	1.01	0.002	0.989	0.167	1.041	-0.022
9	-0.024	0.16	-0.004	0.991	0.003	0.971	0.158	0.992	0.026
10	0.06	0.155	0.003	0.986	0.003	0.991	0.096	1.001	-0.001
11	0.035	0.156	0.001	1.004	0.001	0.995	0.09	1.058	0.031
12	0.013	0.16	-0.004	0.999	<0.00	0.967	0.118	0.979	0.029
13	-0.025	0.155	0.005	1.005	-0.002	0.987	0.12	0.972	0.059
14	-0.02	0.18	-0.006	0.996	-0.003	0.972	0.115	1.005	0.085
15	0.041	0.099	0.006	0.992	-0.002	0.968	0.041	1.029	0.016
16	0.033	0.135	0.001	0.995	0.007	0.959	0.17	0.929	-0.069
17	-0.002	0.14	0.005	1.005	-0.001	0.963	0.171	1.001	-0.029
18	-0.02	0.154	0.001	0.979	0.004	0.974	0.131	0.974	0.043
19*	-0.019	0.128	0.006	1.011	0.008	0.99	0.067	1.12	0.08
20	-0.019	0.149	<0.00	1.003	<0.00	0.998	0.138	1.055	0.031
21	0.036	0.153	0.002	1.016	0.002	0.984	0.097	0.985	0.019

Continued on next page

Simulated Data Score Increases, using IRT Methods- continued from previous page

Simulation	θ_O	θ_I	$\overline{X}_\beta$	σ_β	$\overline{X}_\gamma$	σ_γ	$\overline{X}_\theta$	σ_θ	$(\theta_O - \theta_I)$ $-\overline{X}_\theta$
22	0.001	0.186	0.005	1.016	0.002	0.998	0.162	1.056	0.023
23	-0.056	0.115	-0.005	0.98	-0.002	0.985	0.059	1.056	0.112
24*	0.034	0.135	0.001	0.998	0.002	0.993	0.033	0.981	0.067
25	0.014	0.219	0.009	0.987	0.002	0.98	0.152	1.017	0.053
26	-0.058	0.072	0.008	0.967	0.003	0.972	0.053	1.02	0.077
27	0.004	0.161	-0.002	0.996	-0.001	0.987	0.100	0.972	0.057
28*	0.038	0.146	0.007	1.016	0.007	0.986	0.036	0.976	0.072
29	-0.033	0.108	0.001	1.002	0.004	0.993	0.103	0.997	0.038
30	0.01	0.101	0.006	1.002	-0.004	0.99	0.063	1.014	0.029
31	0.011	0.086	0.001	1.005	0.003	0.994	0.062	1.013	0.012
32	-0.008	0.087	-0.003	1.016	0.003	0.982	0.024	0.968	0.071
33*	0.03	0.168	0.006	1.008	0.005	0.988	0.087	0.98	0.052
34	0.004	0.148	0.006	1.005	0.005	0.99	0.145	1.008	-0.001
35*	0.014	0.126	-0.008	0.993	<0.00	0.991	0.044	0.947	0.068
36	0.008	0.162	0.005	0.994	0.007	0.984	0.131	1.051	0.023
37*	0.04	0.139	0.001	1.02	0.004	0.99	0.049	1.014	0.049
38	<0.00	0.152	-0.008	1.01	0.001	0.981	0.138	0.997	0.014
39	-0.03	0.14	<0.00	1.015	-0.002	0.994	0.182	1.000	-0.012
40	0.025	0.143	-0.001	0.985	0.001	0.973	0.092	0.983	0.026
41	-0.028	0.166	0.01	1.007	0.004	0.97	0.13	1.073	0.064
42	-0.019	0.165	0.007	0.993	0.005	0.988	0.145	0.916	-0.061
43*	0.014	0.149	0.001	0.983	0.004	0.97	0.037	1.042	0.098
44	0.025	0.148	0.003	0.978	-0.001	0.976	0.067	0.982	0.056

Continued on next page

Simulated Data Score Increases, using IRT Methods- continued from previous page

Simulation	θ_O	θ_I	$\overline{X}_\beta$	σ_β	$\overline{X}_\gamma$	σ_γ	$\overline{X}_\theta$	σ_θ	$(\theta_O - \theta_I)$ $-\overline{X}_\theta$
45	-0.034	0.115	0.006	1.021	0.001	0.984	0.112	1.081	0.037
46	0.027	0.094	<0.00	0.999	-0.001	0.974	0.035	0.967	0.031
47	-0.012	0.2	-0.006	1.002	0.001	0.981	0.171	1.068	0.041
48	0.028	0.19	-0.002	1.011	<0.00	0.971	0.157	1.063	0.005
49	0.01	0.148	0.006	1.009	0.003	1.002	0.107	1.164	0.031
50	-0.029	0.093	-0.002	1.004	<0.00	0.976	0.077	0.991	0.044
51	0.009	0.121	0.004	0.004	0.005	0.981	0.086	0.994	0.026
52	-0.027	0.151	-0.001	0.974	0.001	0.973	0.171	0.924	0.006
53+	0.067	0.141	0.006	1.015	0.006	1.004	0.003	0.934	0.072
54	-0.019	0.166	0.015	0.984	0.004	0.962	0.151	0.966	0.033
55	-0.017	0.058	0.003	0.991	-0.002	0.988	0.052	1.005	0.024
56*	-0.006	0.113	-0.006	0.986	0.001	0.971	0.029	1.009	0.091
57	-0.042	0.143	-0.001	1.024	0.002	0.989	0.141	0.944	0.045
58	-0.021	0.125	0.011	0.985	0.006	0.959	0.106	0.963	0.04
59	0.029	0.121	-0.005	1.026	0.003	1.008	0.049	1.039	0.043
60	-0.069	0.141	0.007	1.026	0.002	0.992	0.143	1.012	0.067
61*	0.077	0.144	0.004	1.012	0.001	1.006	0.018	0.965	0.049
62*	0.03	0.172	0.012	1.004	0.004	0.986	0.085	1.003	0.057
63	-0.017	0.185	0.006	0.978	-0.002	0.987	0.175	0.991	0.028
64*	0.005	0.116	<0.00	1.013	<0.00	0.976	0.053	0.978	0.058
65	-0.008	0.177	0.005	0.978	0.001	0.97	0.138	0.981	0.048
66	-0.045	0.13	<0.00	0.991	<0.00	1.003	0.116	0.955	0.06
67	-0.027	0.113	<0.00	0.972	-0.001	0.975	0.091	0.993	0.048

Continued on next page

Simulated Data Score Increases, using IRT Methods- continued from previous page

Simulation	θ_O	θ_I	$\overline{X}_\beta$	σ_β	$\overline{X}_\gamma$	σ_γ	$\overline{X}_\theta$	σ_θ	$(\theta_O - \theta_I)$ $-\overline{X}_\theta$
68*	-0.028	0.116	0.004	1.007	-0.003	0.996	0.049	1.034	0.095
69	0.021	0.106	0.002	0.976	0.005	0.982	0.078	1.042	0.007
70	-0.07	0.134	0.006	0.999	0.001	0.971	0.153	0.987	0.051
71*	0.054	0.215	-0.01	1.008	0.002	0.984	0.086	1.053	0.076
72	-0.015	0.185	<0.00	1.003	-0.001	0.981	0.145	1.016	0.055
73	0.016	0.068	0.003	0.985	0.004	0.983	-0.013	1.013	0.066
74	-0.007	0.13	0.004	0.99	0.002	0.983	0.159	0.922	-0.022
75	0.007	0.17	-0.004	0.998	0.001	0.994	0.145	0.973	0.018
76	0.033	0.149	<0.00	0.995	-0.001	1.005	0.124	0.914	-0.008
77	-0.063	0.123	0.003	0.999	0.004	0.993	0.153	0.934	0.032
78	0.015	0.139	0.012	0.988	0.002	0.979	0.137	0.998	-0.014
79	-0.026	0.126	0.002	0.98	0.002	0.997	0.094	0.996	0.059
80*	0.006	0.169	-0.006	0.98	-0.004	0.971	0.102	0.962	0.062
81*	0.047	0.143	0.002	0.992	0.004	0.992	0.082	0.934	0.014
82	-0.033	0.165	0.005	1.019	0.007	0.987	0.123	0.995	0.076
83	-0.028	0.13	<0.00	0.994	0.001	0.988	0.129	1.021	0.029
84*	0.048	0.209	0.005	1.003	-0.001	0.977	0.089	0.967	0.071
85	-0.023	0.132	0.007	1.004	<0.00	0.982	0.083	0.95	0.073
86	-0.021	0.17	0.005	1.011	0.004	0.99	0.119	1.001	0.072
87	0.008	0.14	0.008	0.994	0.004	0.966	0.142	0.988	-0.01
88*	0.025	0.117	-0.002	0.985	0.004	0.994	0.036	0.972	0.055
89	0.008	0.117	-0.001	1.01	-0.009	0.982	0.121	0.921	-0.012
90	-0.025	0.153	0.001	1.001	<0.00	0.987	0.139	1.017	0.038

Continued on next page

Simulated Data Score Increases, using IRT Methods- continued from previous page

Simulation	θ_O	θ_I	$\overline{X}_\beta$	σ_β	$\overline{X}_\gamma$	σ_γ	$\overline{X}_\theta$	σ_θ	$(\theta_O - \theta_I)$ $-\overline{X}_\theta$
91	0.012	0.105	-0.005	0.997	-0.003	0.981	0.083	0.967	0.01
92	-0.024	0.113	-0.002	0.974	0.004	0.981	0.086	1.013	0.051
93+	0.014	0.152	0.705	1.132	0.094	0.993	0.107	1.039	0.031
94*	0.002	0.213	-0.002	0.994	0.001	0.98	0.143	0.942	0.068
95*	-0.004	0.14	-0.003	1.013	0.002	0.995	0.064	1.029	0.08
96	0.017	0.134	0.01	0.984	0.008	0.98	0.119	1.001	-0.002
97	-0.006	0.087	-0.008	0.967	-0.001	0.963	0.137	0.981	-0.044
98*	-0.009	0.142	0.01	0.98	<0.00	0.968	0.07	0.958	0.081
99	0.019	0.159	0.001	0.989	-0.001	0.957	0.098	0.949	0.041
100	-0.044	0.148	0.003	1.005	0.001	0.99	0.136	1.005	0.056
Average	<0.00	0.140	0.010	0.988	0.003	0.983	0.101	0.996	0.039
SD	0.030	0.033	0.070	0.101	0.010	0.011	0.045	0.0443	0.034

θ_O: Average latent trait of initial sample of simulated examinees, $E[\theta] = 0$;

θ_I Average latent trait of sample of simulated examinees with true latent trait increase, $E[\theta] = .14$;

$\overline{X}_\beta$ & σ_β: Average and variation of standard score increase when original simulated sample took test with (mean) decreased β values, respectively.

$\overline{X}_\gamma$ & σ_γ: Average and variation of standard score increase when original simulated sample took test with increased γ values, respectively.

$\overline{X}_\theta$ & σ_θ: Average and variation of standard score increase when simulated sample with increased latent ability took original test, respectively.

*: large discrepancy between θ_I and $\overline{X}_\theta$

Note. In all instances, $\overline{X}_\theta > max(\overline{X}_\beta, \overline{X}_\gamma)$, except where indicated by a (+).

Table 6
Statistics for the CTT Analysis of the CBASE Data (n=619)

Form	$\overline{X}$	σ	$\overline{X}_S$	σ_S
LK	36.214	10.219		
LO	34.393	10.929	-0.178	1.069

Note. $\overline{X}_S$ & σ_S: Average and variation of standard score increase, respectively (LK is reference group).

Table 7

Item Parameter Estimates for CBASE forms LK, LO, and Transformed LO

	LK			LO			LO (Transformed)		
index	α	β	γ	α	β	γ	α	β	γ
1	1.078	-0.462	0.001	0.985	-0.557	0.001	0.944	-0.798	0.001
2	0.692	-0.986	0.001	0.803	-0.490	0.001	0.770	-0.727	0.001
3	0.892	-1.011	0.001	0.963	-0.693	0.001	0.923	-0.940	0.001
4	0.990	-0.262	0.001	1.509	0.174	0.029	1.446	-0.035	0.029
5	0.746	-1.203	0.001	1.106	-0.858	0.001	1.060	-1.111	0.001
6	1.035	-1.558	0.001	0.860	-1.460	0.001	0.824	-1.739	0.001
7	0.752	-0.375	0.001	0.743	-0.161	0.001	0.712	-0.385	0.001
8*	1.265	-0.859	0.001	1.212	-0.258	0.001	1.161	-0.485	0.001
9	0.771	-0.562	0.001	0.879	-0.177	0.002	0.842	-0.401	0.002
10	1.335	-0.709	0.001	1.422	-0.465	0.001	1.363	-0.701	0.001
11	1.346	-1.169	0.001	1.485	-1.084	0.001	1.423	-1.348	0.001
12*	1.161	-0.443	0.001	1.096	-0.613	0.002	1.051	-0.856	0.002
13	1.216	-1.109	0.001	1.244	-0.860	0.001	1.193	-1.114	0.001
14	1.281	-0.740	0.001	1.501	-0.440	0.002	1.438	-0.675	0.002
15	1.318	-0.847	0.001	1.223	-0.560	0.002	1.172	-0.800	0.002
16	1.140	0.136	0.001	0.928	0.301	0.001	0.890	0.098	0.001

Note. Items with an * exhibited DIF/drift at (non-adjusted) $\alpha = .05$, using Lord's (1980) χ^2.

Scaling Parameters: A (Scale): 1.043; K (Location): -0.216

Table 8

Item Parameter Statistics and Area Indices

	LK			LO					
	SE			SE					
index	α	β	cov(α,β)	α	β	cov(α,β)	SA	UA	H
1	0.141	0.106	0.005	0.129	0.116	0.007	-0.335	0.335	-0.337
2	0.114	0.193	0.015	0.116	0.134	0.006	0.258	0.258	0.267
3	0.134	0.164	0.015	0.13	0.128	0.009	0.071	0.071	0.073
4	0.128	0.106	0.003	0.288	0.146	0.025	0.227	0.227	0.325
5	0.118	0.209	0.019	0.144	0.125	0.011	0.091	0.091	0.333
6	0.149	0.198	0.024	0.137	0.218	0.024	-0.181	0.181	-0.254
7	0.111	0.136	0.005	0.11	0.129	0.002	-0.010	0.010	-0.061
8*	0.171	0.115	0.012	0.146	0.091	0.003	0.373	0.373	-0.373
9	0.118	0.143	0.008	0.122	0.120	0.002	0.161	0.161	0.171
10	0.166	0.104	0.009	0.165	0.089	0.005	0.008	0.008	0.014
11	0.184	0.136	0.018	0.197	0.118	0.015	-0.178	0.178	0.178
12*	0.143	0.1	0.005	0.138	0.112	0.007	-0.413	0.413	-0.413
13	0.164	0.137	0.016	0.158	0.115	0.012	-0.004	0.004	-0.013
14	0.16	0.107	0.009	0.188	0.087	0.005	0.065	0.065	0.090
15	0.178	0.11	0.011	0.154	0.102	0.007	0.046	0.046	-0.086
16	0.143	0.095	-0.001	0.125	0.115	-0.004	-0.038	0.038	-0.204

SE: Standard Error; SA: Signed Area; UA: Unsigned Area ($\alpha_{LK} = \alpha_{LO}$);

H: Unsigned Area ($\alpha_{LK} \neq \alpha_{LO}$);

Note. Indexes with an * indicate items with areas whose standardized statistics are greater than 3.

Table 9

CBASE Items, in Order Used in Analysis

	LK		LO	
Index	Question	Name	Question	Name
1	q42	202B-MA-08-I	q42	202B-MA-08-I
2	q44	202C-MA-05-S	q44	202C-MA-05-S
3	q52	201A-MA-03-S	q52	201A-MA-03-S
4	q56	201C-MA-74-I	q55	201C-MA-74-I
5	q58	203B-MA-70-S	q58	203B-MA-70-S
6	q60	203B-MA-08-S	q60	203B-MA-08-S
7	q62	203A-MA-07-A	q64	203A-MA-07-A
8	q70	204B-AL-07-I	q69	204B-AL-07-I
9	q80	205C-AL-01-I	q80	205C-AL-01-I
10	q83	206A-GE-18-I	q83	206A-GE-18-I
11	q85	206C-GE-11-I	q85	206C-GE-11-I
12	q89	206C-GE-71-A	q88	206C-GE-71-A
13	q91	207A-GE-02-S	q91	207A-GE-02-S
14	q92	207A-GE-03-S	q93	207A-GE-03-S
15	q96	207A-GE-12-I	q96	207A-GE-12-I
16	q97	207C-GE-02-S	q97	207C-GE-02-S
17	q43	202E-MA-02-S	q43	NG
18	q45	202B-MA-01-I	q45	NG
19	q46	202D-MA-03-S	q46	202C-MA-08-I
20	q47	202A-MA-05-I	q47	202B-MA-07-I
21	q48	202A-MA-02-S	q48	202E-MA-03-I

Continued on next page

	LK		LO	
Index	Question	Name	Question	Name
22	q49	202C-MA-03-S	q49	202C-MA-09-I
23	q50	201C-MA-04-S	q50	NG
24	q51	201A-MA-73-I	q51	NG
25	q53	201A-MA-07-S	q53	NG
26	q54	201B-MA-06-S	q54	201C-MA-06-S
27	q55	201C-MA-03-S	q56	201B-MA-09-I
28	q57	201B-MA-04-S	q57	201B-MA-11-I
29	q59	203A-MA-02-S	q59	203C-MA-02-I
30	q61	203B-MA-01-A	q61	203A-MA-11-I
31	q63	203B-MA-04-A	q63	NG
32	q64	203A-MA-04-A	q62	203B-MA-13-A
33	q65	203A-MA-03-S	q65	203A-MA-10-I
34	q66	204B-AL-10-I	q66	204B-AL-06-I
35	q67	204A-AL-02-I	q67	204A-AL-70-I
36	q68	204A-AL-04-I	q68	204A-AL-06-I
37	q69	204A-AL-05-I	q70	204B-AL-72-I
38	q71	204B-AL-09-I	q71	204B-AL-02-I
39	q72	204A-AL-03-I	q72	NG
40	q73	204B-AL-04-I	q73	204B-AL-73-I
41	q74	205A-AL-14-I	q74	NG
42	q75	205A-AL-03-I	q75	205A-AL-73-I
43	q76	205A-AL-08-I	q76	205A-AL-07-I
44	q77	205B-AL-03-I	q77	205B-AL-01-I
45	q78	205B-AL-02-I	q78	205B-AL-09-I

Continued on next page

	LK		LO	
Index	Question	Name	Question	Name
46	q79	205A-AL-04-I	q79	205A-AL-06-I
47	q81	205B-AL-05-I	q81	205B-AL-10-I
48	q82	206B-GE-10-I	q82	NG
49	q84	206C-GE-06-I	q84	NG
50	q86	206A-GE-05-I	q86	206A-GE-70-I
51	q87	206A-GE-07-I	q87	206A-GE-01-I
52	q88	206C-GE-04-S	q89	206A-GE-06-S
53	q90	207B-GE-71-S	q90	NG
54	q93	207A-GE-01-S	q92	207A-GE-14-S
55	q94	207C-GE-04-S	q94	207C-GE-03-I
56	q95	207B-GE-05-I	q95	207B-GE-01-I

NG: Item name not given in raw data.

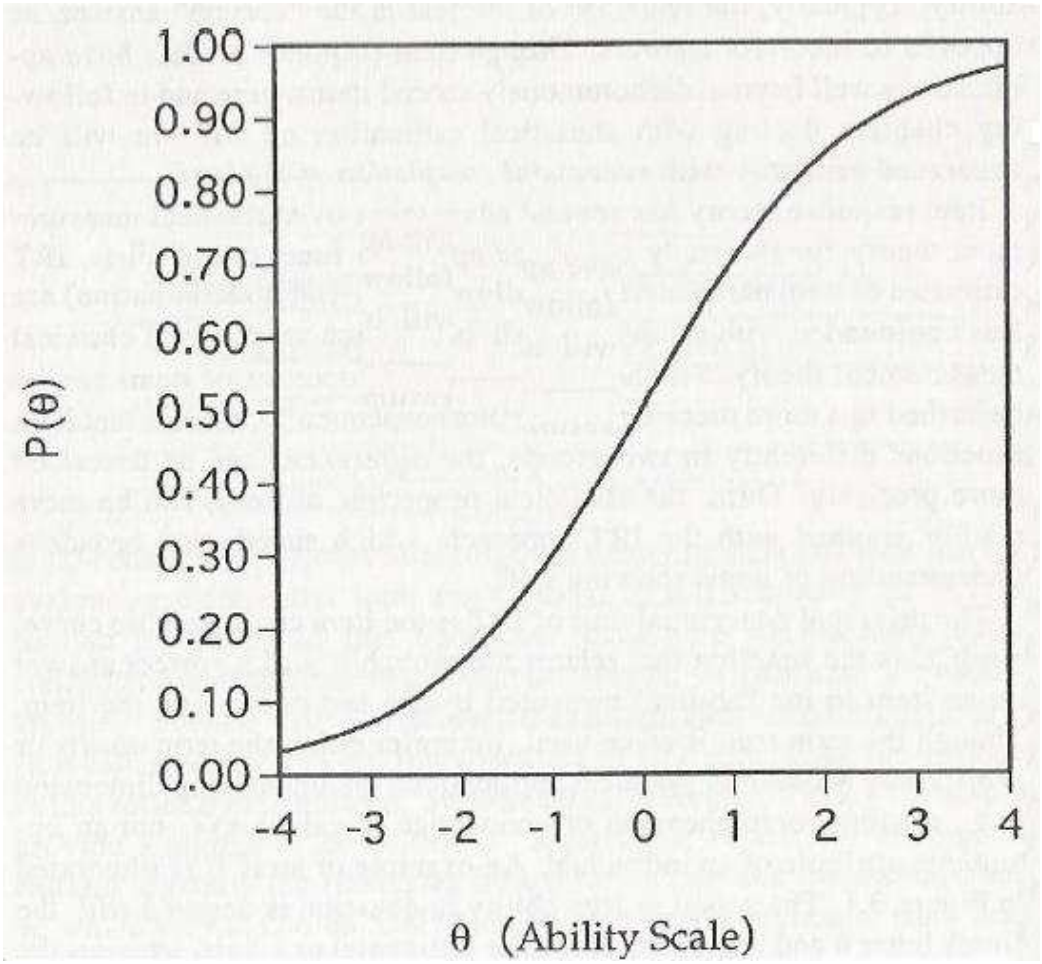
Note. Question is the question number that that given item was given for that form of the CBASE examination.

Figure Captions

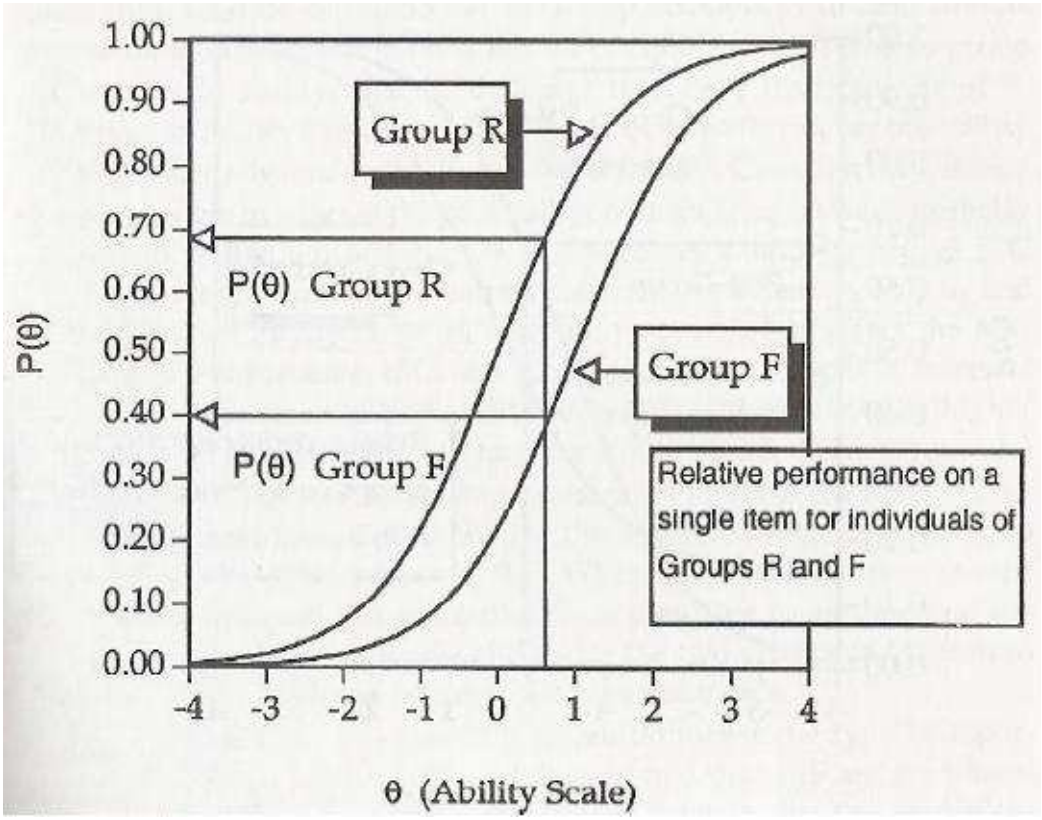
Figure 1. Item Characteristic Curve, $f(x_i|\boldsymbol{\kappa}_i)$

Figure 2. DIF between 2 groups

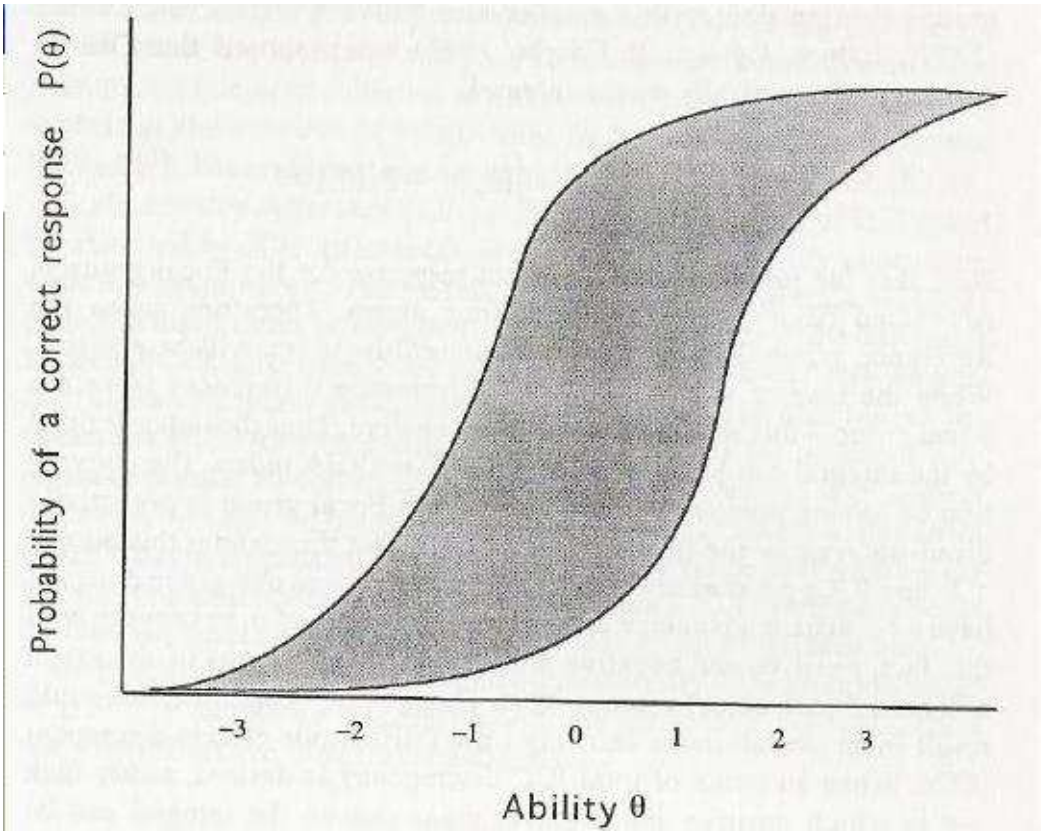
Figure 3. Area between ICCs exhibiting DIF



Note. Adapted from G. Camilli & L. A. Shepard (1994). *Methods for identifying biased test items*, p. 48.



Note. From G. Camilli & L. A. Shepard (1994). *Methods for identifying biased test items*, p. 59.



Note. From G. Camilli & L. A. Shepard (1994). *Methods for identifying biased test items*, p. 65.

VITA

A. Alexander Beaujean was born May 25, 1978 in Springfield, Ohio. He has received the following degrees: BA in Psychology and History (*high honors*) from Cedarville University in Cedarville, OH (1999); MA in School Psychology (2003), MA in Educational Psychology, with a minor in Statistics (2004), PhD in Educational Psychology (2005), and PhD in School Psychology (2006) from the University of Missouri-Columbia. He will complete his doctoral internship at the Applewood Centers, Inc. in Cleveland, Ohio during the 2005-2006 academic year.