PSYCHOMETRIC IMPACTS OF ABOVE-LEVEL TESTING

A Dissertation

by

RUSSELL THOMAS WARNE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2011

Major Subject: Educational Psychology

Psychometric Impacts of Above-Level Testing

Copyright 2011 Russell Thomas Warne

PSYCHOMETRIC IMPACTS OF ABOVE-LEVEL TESTING

A Dissertation

by

RUSSELL THOMAS WARNE

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Ernest Goetz
	Myeongsun Yoon
Committee Members,	Joyce Juntune
	Aaron Taylor
Head of Department,	Victor Willson

May 2011

Major Subject: Educational Psychology

ABSTRACT

Psychometric Impacts of Above-Level Testing. (May 2011)

Russell Thomas Warne, B.S., Brigham Young University

Co-Chairs of Advisory Committee: Dr. Myeongsun Yoon
Dr. Ernest Goetz

Above-level testing is the practice of administering a test level—of usually an academic achievement or aptitude test—to a gifted or high achieving child. This procedure is widely accepted in gifted education circles, on the basis of theoretical claims that above-level testing raises the test ceiling, increases variability among gifted students' scores, improves reliability of data, reduces regression toward the mean, and improves interpretation of data from gifted students. However, above-level testing has not been subject to careful psychometric scrutiny.

In this study, I examine reliability data, growth trajectories, distributions, and group differences of above-level test scores obtained from the Iowa Tests of Basic Skills (ITBS) and Iowa Tests of Educational Development (ITED). Subjects in this study were 224 students who were tested a total of 435 times while enrolled in a gifted magnet program for middle schoolers. Longitudinal analyses performed with hierarchical linear modeling indicate that substantial differences exist between students from overrepresented ethnicities (White and Asian Americans) and those from underrepresented ethnicities (Hispanic and African Americans) in both initial scores and the rate of score gains. Gender differences existed only for the rate of score increases for

above-level reading scores. Socioeconomic differences existed, but did not have a unique impact beyond that of the ethnicity variable.

A discussion of these results within the wider gifted education research context and suggestions for further research are included. An appendix to the study gives information about item difficulty indexes for every item in the ITBS/ITED core battery for the eighth, ninth, and tenth grade levels of Form C.

ACKNOWLEDGMENTS

My name is the only one on the cover of this dissertation, but that should not imply that others did not have a hand in this work. The strongest influences came from my committee members. Dr. Myeongsun Yoon was indispensible for guiding me through the technical aspects of the study—especially in reporting standards for hierarchical linear modeling results. She deserves an extra big *kam sam ham nee dah*. Dr. Ernest Goetz had an intense eye for detail as he reviewed the design for the study, providing useful suggestions for handling the practical elements of the study. His careful examination of the drafts of every chapter also saved me from many embarrassing mistakes. Dr. Joyce Juntune made this study possible through her connections with gifted personnel in Texas school districts. Indeed, she is an essential resource for anyone who wishes to conduct a study on gifted students in the state. Finally, Dr. Aaron Taylor put to good use his knack for asking deep, probing questions, which often helped me see shortcomings in my work. I am humbled and honored that these four extraordinary scholars and teachers would agree to take me as a student.

Other professors in the Department of Educational Psychology at Texas A&M University made important indirect contributions to this study. Dr. Oimon Kwok taught me everything I know about hierarchical linear modeling and structural equation modeling and clarified many misconceptions I had about statistical methods. Dr. Bruce Thompson provided a firm foundation for my statistical knowledge through his classes and several theoretical conversations about the general linear model and other topics.

Dr. Victor Willson provided encouragement, good humor, and an excellent training on longitudinal data analysis.

Several classmates also provided substantive support, encouragement, and friendship during my years at Texas A&M. Fatih and Sumeyye Kaya performed (and checked) the data entry from the two followups—thereby saving me *a lot* of time. Dr. Beth Barnes and Diana Hood helped me gain access to the data and helped me understand the local context in which the study took place. Dr. Susan Skidmore was always willing to take the time to listen to my ramblings about obscure psychometric topics. Dr. Ross Larsen was a willing co-author on two manuscripts and (hopefully) a lifelong collaborator. Dr. Jiun-Yu Wu and Dr. Karen Lee were constantly supportive and believed in my study when problems arose while it was being carried out. Other classmates were supportive or helpful in too many ways to list individually: Eun Sook Kim, Dr. Hsien-Yuan Hsu, Eunju Jung, Minjung Kim, Yan Li, Alma Contreras, Dr. Leena Landmark, Jackie Pacha, Maria Lazo, and Tammy Ramos. The order of this list should not imply anything about the relative importance of these classmates' support or influence; any omissions are unintentional.

I also appreciate the friendship and encouragement of my Texas friends whom I met outside of a graduate student context: Dr. Donald Adams, Tony Brown, Rosalma Arcelay, Jared Porter, and Will Hausman all listened to *way* too many rants about how awful it is to be a graduate student. Rich and Caroline White graciously opened up their home to me many times over the course of four years—especially during college football season. Other friends who live further away were also supportive of my dissertation and

graduate school endeavors. Dave Mortensen, Justin and Katie Ferrell, Dale and Judy Rex, Dr. Ed Gantt, and Fáiver González all remember when I was applying to graduate school years ago and now get to watch from afar as I finish up.

I also appreciate the folks at the Research and Evaluation Network of the National Association for Gifted Children, especially Dr. D. Betsy McCoach. The network named this study as the best doctoral-level research in progress at their 2009 and 2010 national conferences, which was both an appreciated ego boost and a strong signal that what I was doing in this study was important to other researchers.

Finally, I'm thankful for my family through this entire process. Even when they didn't know what psychometrics was, they supported my decision to study it. Unfortunately, that decision has burdened my parents with the task of explaining to their friends how their son could be a psychologist and yet (a) never have any clients, and (b) care more about test items than about people. It has also forced my brothers to one day explain to their children that their uncle actually *likes* standardized tests. I also appreciate my family's sincere efforts to understand what exactly it is I study so they can stop asking the question, "So what is it you do, again?" at every family gathering.

The dissertation committee gave me a significant amount of freedom as I selected my topic and methodology, which I greatly appreciate. But with that freedom comes a high level of responsibility for any shortcomings in this study. I take full responsibility for any such problems, but I liberally share credit for the study's positive aspects with all those mentioned in these pages.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

Gifted education experts have long recognized that regular standardized achievement and aptitude tests are not suitable for testing the abilities of gifted children. Grade-level tests are designed to measure the middle levels of ability—where the majority of students' abilities lie—as effectively as possible (Lohman, 2005; Minnema, Thurlow, Bielinski, & Scott, 2000; Stanley, 1977). The emphasis that typical standardized tests place on average students often makes the tests unsuitable for obtaining accurate data on gifted children. This has led researchers in gifted education to look for different methods of objective assessment of gifted students. One method that gifted education researchers have used to test high ability children is called *above-level testing* (Stanley & Benbow, 1981-1982). Above-level testing is the procedure of administering a test to a gifted child who is younger or in a lower grade than the group for which the test was originally designed.

Above-level testing is a widespread and accepted practice in gifted education, where it is used to screen students for Talent Search participation (Swiatek, 2007) and full-grade acceleration (Assouline, Colangelo, Lupkowki-Shoplik, Lipscomb & Forstadt, 2009; Rogers, 2002). Although there are isolated cases of above-level testing throughout most of the 20th century (e.g., Almack & Almack, 1921; Hollingworth, 1926, 1942; Stanley, 1951; Stedman, 1924; Terman, 1926; Terman & Fenton, 1921; Witty & Jenkins, 1935), it was not a regular and widely accepted practice until the 1970's. In that

---

This dissertation follows the style of *Gifted Child Quarterly*.

decade, Stanley began the first Talent Search and screened seventh- and eighth-grade gifted children for admission by administering the SAT to them. Through the efforts and research of Stanley, above-level testing has become a widespread practice in gifted education, mostly in a Talent Search context (S.-Y. Lee, Matthews, & Olszewski-Kubilius, 2008, for a review of the present state of Talent Search programs). Above-level testing is also advocated by gifted education researchers in academic acceleration (Assouline et al., 2009; Rogers, 2002) and other gifted education practices (Gross, 1999; Rogers, 2002).

Advocates of above-level testing give four main reasons for conducting above-level testing: (a) above-level testing raises the test ceiling which also makes high-ability examinees' scores more variables and discriminating, (b) improves score reliability when scores are obtained from above-level tests, (c) makes gifted students' scores more comparable to the scores of the older pupils for whom the test was designed for, and (d) reduces regression toward the mean.

Problem Statement

Above-level testing has rarely been the subject of psychometric study. Indeed, most proponents of the practice cite a mixture of personal experience and theoretical considerations to justify the practice (e.g., Assouline et al., 2009; Olszewski-Kubilius, 1998a; Swiatek, 2007). Rarely have researchers attempted to examine the psychometric properties of above-level test scores such as reliability or the validity of using an academic test to screen younger students for gifted education programs and interventions—a population and purpose for which the test was not designed for.

Because current standards in education testing mandates that changes in the mode of administration of a test or change the population that takes a test be validated by the test user (American Educational Research Association [AERA], American Psychological Association, & National Council on Measurement in Education, 1999), it is important that research be conducted on above-level testing.

In this dissertation, I will focus on two heretofore uninvestigated aspects of above-level test scores: reliability and predictors of above-level test score and score change. Reliability is an important property of test scores, because it measures how stable those scores are across different conditions (Kaplan & Saccuzzo, 2005). Although advocates of above-level testing claim that above-level testing raises the reliability of gifted students' test scores (e.g., Keating, 1975), reporting reliability of above-level test scores is rare (i.e., Loyd, 1980; Stanley, 1951). Therefore, in this dissertation I will examine the reliability of a set of above-level test scores to see if the coefficients meet the recognized standards of reliability in research.

Despite the widespread opinion among gifted education experts that above-level testing is a suitable method of measuring gifted children's abilities, no research has attempt to examine the trajectory of score change for a sample of gifted students. I will do this type of growth modeling through use of hierarchical linear modeling (HLM). HLM is a technique that also permits an investigation of how student characteristics (i.e., demographic factors) are related to both individual student scores and changes in student scores over time.

In addition to the reliability and change of above-level test scores, I will also examine the influence of demographic variables on observed above-level test scores. The impact of demographic variables is important to examine, because above-level test scores consistently show differences in the performance of different ethnic or racial groups (e.g., Ebmeier & Schmulbach, 1989; S.-Y. Lee & Olszewski-Kubilius, 2006). Moreover, gender group differences in above-level test scores are frequently observed, although not as consistently and with much smaller gaps than are observed between ethnic or racial groups (e.g., Barnett & Gilheany, 1996; Benbow, 1992; S.-Y. Lee & Olszewski-Kubilius, 2006). Examining the strength the influence of demographic variables on above-level test scores could give clues into the barriers of entry that underrepresented groups—mostly African Americans, Hispanics, and low income students—must overcome before participating in programs like Talent Search.

Another area of interest within gifted education is the investigation of the rate of student learning gains. Because most academic achievement tests have ceilings that are too low to measure gifted students' learning gains, very little is known about the rate of long-term learning of gifted students. Although theory dictates that gifted students should learn faster than their peers (Gagné, 2005), few studies have been done to determine specific information and influences on learning rate. Through this study, I seek to examine the rate of score gains on an above-level achievement test and examine demographic influences on score gains.

An examination of the psychometric and predictors of score and growth of above-level test scores may also provide researchers with important information that

could have implications for program evaluation and individual educational planning for gifted children. Program evaluation is an area that gifted education practitioners and researchers have performed poorly (Borland, 2003; Gallagher, 2006; VanTassel-Baska, 2006). If above-level scores can be used to track gifted students' progress through an education program, then the practice may potentially be incorporated into program evaluation procedures.

Research Questions

The topics of above-level psychometric score characteristics and growth modeling will be addressed through the five research questions below:

1. What is the internal consistency reliability of the global battery, reading/language arts, and mathematics scores drawn from an above-level administration of an achievement test?
2. Do gifted children make larger achievement gains in overall, reading/language arts, and mathematics scores than average students in a more advanced grade?
3. Do demographic variables (gender, ethnicity, and SES) influence the initial scores or rate of overall, reading/language arts, and mathematics score growth of gifted students?
4. What is the relationship between initial above-level overall, reading/language arts, and mathematics scores and rate of score growth?
5. What percentage of overall, reading/language arts, and mathematics score variance is explainable through time, demographic variables, and cohort membership?

These questions were answered through a longitudinal study in which all students in a local middle school gifted magnet program were administered above-level versions of the Iowa Tests of Basic Skills (ITBS) and the Iowa Tests of Educational Development (ITED). Through this study I hoped to increase substantive and psychometric researchers' understanding of the properties and uses of above-level test scores.

CHAPTER II

LITERATURE REVIEW[1]

Gifted education experts have long recognized that regular standardized achievement and aptitude tests are not suitable for testing the abilities of gifted children. Grade-level tests are usually designed to measure the middle levels of ability—where the majority of students' abilities lie—as effectively as possible (Lohman, 2005; Minnema, Thurlow, Bielinski, & Scott, 2000; Stanley, 1977). The emphasis that typical standardized tests place on average students has led researchers in gifted education to look for different methods of objective assessment in order to obtain accurate data on gifted children. One method that gifted education researchers have used to test high ability children is called *above-level testing* (Stanley & Benbow, 1981-1982). Above-level testing is the procedure of administering a test to a gifted child who is younger or in a lower grade than the group for which the test was originally designed.

The purpose of this chapter is to provide a comprehensive literature review that traces the genesis, development, and present status of above-level testing in gifted education. In this chapter, I will place a special emphasis on the psychometric logic behind above-level testing as I describe the justifications that gifted education researchers have used in support of above-level testing. I will also critically evaluate the current state of the literature supporting above-level testing, give recommendations for further research on the practice, and describe the research goals that I have for my study.

---

[1] A version of this literature review has been submitted to *Roeper Review*.

Terminology and Search Procedures

Above-level testing can be contrasted with *below-level testing*, which is the administration of a test form to a child who is older or in a higher grade than the group for which the test was designed, such as in a special education situation (Minnema, Thurlow, Bielinski, & Scott, 2001). Both above- and below-level testing are included in the term *out-of-level testing*, although some researchers use "out-of-level testing" to exclusively refer to either above-level or below-level testing. For the sake of clarity, this article will use the term "above-level testing" because there is less ambiguity with the term than with "out-of-level testing." It should be noted that above-level testing is also called *off-grade testing* (e.g., S.-Y. Lee et al., 2008) and *off-level testing* (e.g., Gross, 2004), but these terms could be applied to below-level testing as well.

Several procedures were used in the attempt to gather all relevant scholarly literature on above-level testing. First, a search was performed for all of the above terms in the PsycINFO, ERIC, and Google Scholar databases and all relevant articles were read and analyzed. Second, the reference lists of articles from the database searches were examined to find articles, papers, and other literature that did not appear in the database searches. Third, the early case studies of high ability children were examined in order to find early (pre-1970's) examples of above-level testing. Finally, a few miscellaneous searches on specific tests (such as the Army Alpha and the Terman Group Test) were also performed in order to see how those tests were used in above-level testing. This final search procedure was performed in an effort to find additional early case studies of above-level testing. It should be noted that the various terms defined in

this section also appear in the literature unhyphenated, which was taken into account during the literature search.

Development of Above-Level Testing

Above-level testing is almost as old as standardized testing itself. During the process of the creation and norming of the Army Alpha and Army Beta tests, elementary and high school students were administered both tests (Yoakum & Yerkes, 1920). Shortly after World War I, the Army Alpha was also administered to students as young as 11 years old in studies that would today be viewed as primitive validity studies (Almack & Almack, 1921; Madsen, 1920; Madsen & Sylvester, 1919).

Like many milestones in the history of gifted education, the first case of true above-level testing in the literature was conducted by Lewis M. Terman. Along with his colleague, Jessie C. Fenton, Terman administered the Army Alpha and the Terman Group Test to a 7-year-old girl in November 1919. The child scored 71 on the Army Alpha—approximately equal to the average score of a fourteen-year-old native-born White American male—and 151 on the Terman Group Test, which was the median score for grade 12 (Terman & Fenton, 1921, pp. 164-165). Unfortunately, Terman and Fenton did not explain why they gave these above-level tests to the seven-year-old examinee. However, at the time, Terman was preparing for his landmark longitudinal study of gifted children and the test administrations may have served as a pilot test for the suitability of using the Army Alpha and the Terman Group Test in his later research. Indeed, the girl was later a member of the gifted sample in Terman's study (Burks, Jensen, & Terman, 1930; Terman, 1926).

Terman would later administer the eighth-grade level of the Stanford Achievement Test in an above-level fashion to 100 high IQ students with an average age of 9.86 years in order to compare them to a group of 96 regular eighth-graders from a previous study performed by Kelley (1923). Terman explained the logic of his choice of using above-level testing by saying, "A group of gifted eighth grade children would not be satisfactory because their scores would too often be close to or actually at the maximum possible with the Stanford Achievement Test" (Terman, 1926, p. 310). In other words, the ceiling for the Stanford Achievement Test was too low for gifted eighth graders, so Terman had to choose a younger group of gifted children for the test in order to measure the gifted children's ability. This desire to overcome the limited range of a grade-level test is a long-running theme in the literature on above-level testing.

Other instances of above-level testing are scattered throughout the early gifted education literature. Under Terman's influence, Stedman (1924) administered the Terman Group Test and the Army Alpha to children as young as 11- and 9-years-old, respectively. Similarly, Witty and Jenkins (1935) drew upon Terman's work when administering adult-level tests (the Otis S. A., Army Alpha, and McCall Multi-Mental tests) to a 9-year-old African American girl. Outside of Terman's sphere of influence, Almack and Almack (1921) administered the Army Alpha to a convenience sample of gifted high school students, which included two 11-year-olds who had been accelerated in their school progress. Similarly, Hollingworth (1926, 1942) seems to have independently thought of above-level testing when she gave the Army Alpha to children aged 7 to 13.

None of these other early researchers explained clearly why they were administering above-level tests. Perhaps the problems of the low test ceiling of grade- and age-level tests were so obvious to these researchers that they didn't bother explaining the rationale behind their above-level testing. For example, the pattern of Hollingworth's (1942) records could indicate that she administered the Army Alpha in the early 1920's when her students were scoring at or near the ceiling of the 1916 Stanford-Binet IQ test, but she did not explicitly say this.

Also, none of the early above-level testing practitioners—including Terman—indicated whether or how the above-level test scores were used in educational practice or planning for the gifted children. The only exception to this is Hollingworth (1942), who stated that Army Alpha scores from two of her high IQ case studies (labeled Child C and Child F) influenced their placement in the special schools that she ran in New York City, but the details on the decision making process and the magnitude of the role of above-level test scores in decision making are unclear.

Of all the early incidents of above-level testing, Hollingworth's (1926; 1942) work had the greatest future impact. In 1969, Julian Stanley of Johns Hopkins University encountered a mathematically bright 13-year-old boy. Drawing upon his knowledge of Hollingworth's work, Stanley administered the College Board's Scholastic Aptitude Test (SAT) to the child (Stanley, 1990). Stanley, being a psychometrician and methodologist with a passing interest in gifted education (Benbow & Lubinski, 2006), had previously administered tests above-level, but these endeavors had generated little interest (Stanley, 1951, 1954). The young teenager excelled at the SAT and eventually

earned a bachelor's and a master's degree at age 17 after being heavily accelerated in his education (Stanley & Benbow, 1981-1982). Within a few years, Stanley had found over 2000 middle school students who scored above the mean of high school seniors on the SAT-M (Stanley, 1976, p. 75). To accommodate those children's special educational needs, Stanley created a curriculum of accelerated mathematical instruction. This process—based on above-level testing—is called Talent Search and has spread to other universities around the United States (see S.-Y. Lee et al., 2008, for a review of the present state of Talent Search programs).

Stanley was familiar with Terman's longitudinal study of highly gifted children (Burks et al., 1930; Terman, 1926; Terman & Oden, 1947, 1959) and understood the importance of following up on the educational outcomes of the high ability children that he found through above-level testing (Stanley, 1990). Therefore, Stanley launched the Study for Mathematically Precocious Youth (SMPY) to study his high ability pupils (Stanley, 2005). Much of the research on above-level testing has come out of SMPY and Talent Search programs, and what little independent research there is on above-level testing is highly influenced by Stanley's work. This fact must be kept in mind when examining the literature on above-level testing.

Rationale of Above-Level Testing

As researchers have written about above-level testing, they have given several empirical or theoretical justifications for the practice. In my review of the literature, I have categorized these into four general claims about the benefits of above-level testing:

1. Above-level testing raises the test ceiling for gifted examinees which makes the observed scores of gifted students are more variable and discriminating when obtained from above-level tests.
2. Score reliability improves when gifted examinees are tested above-level.
3. Gifted pupils' scores are comparable to regular students for whom the tests were designed.
4. Regression toward the mean is reduced through above-level testing.

The following section of this chapter will examine the psychometric theory behind these claims and also evaluate relevant empirical studies in an effort to judge whether above-level testing is an empirically supported and theoretically justified practice.

Raising the Test Ceiling

The use of above-level testing has largely been driven by a practical need to examine the abilities of gifted children. The literature in gifted education is full of examples of bright children obtaining the highest possible score on regular tests (e.g., Gross, 2004; Ruf, 2005). Indeed, the oldest justification for above-level testing (Terman, 1926) was that it was needed to examine the abilities of children because regular tests were too easy for the gifted. Although the reasoning is old, the claim that above-level testing is needed to raise the test ceiling and examine students' real abilities has been echoed in more recent times (e.g., Assouline et al., 2009; Feldhusen, Proctor, & Black, 2002; Olszewski-Kubilius & Kulieke, 2008; Olszewski-Kubilius & S.-Y. Lee, 2011; Rogers, 2002; Stanley, 1977). In fact, raising the test ceiling is the most commonly stated rationale for above-level testing.

Without question, the empirical literature supports the view that the test ceiling for gifted children is raised through above-level testing (e.g., Achter, Lubinski, & Benbow, 1996; Keating, 1976; C. J. Mills & Barnett, 1992; Terman, 1926; VanTassell-Baska, 1986). In fact, I have been unable to find an example in the literature of a group of gifted children who have not obtained higher scores on a test that was at least two levels above their age group than the maximum scale score of the grade-level test. The fact that above-level testing has raised the test ceiling for high ability examinees is probably the most consistent finding presented in this literature review and one of the hardest to ignore.

However, there is no strong consensus about what constitutes an observed "ceiling effect," beyond obtaining the maximum score possible on a grade-level test. Validation studies on the cutoff scores for children to be eligible to take the SAT or ACT to apply for Talent Search programs have frequently found that children who score at the 95th percentile or higher on a grade-level test tend to obtain scores on an above-level test that would be approximately average for students four or more years older than them (Ebmeier & Schmulbach, 1989; Lupkowski-Shoplik & Swiatek, 1999; Olszewski-Kubilius, Kulieke, Willis, & Krasney, 1989; Olszewski-Kubilius & S.-Y. Lee, 2011; VanTassell-Baska, 1986).[2] More research needs to be done to investigate the exact

[2] It should be noted that 7.0% of Lupkowski-Shoplik and Swiatek's (1999) sample were tested two grade levels above their nominal grade, 35.8% were tested three levels above their nominal grade, and 67.2% were tested four or five grades above their nominal grade. Unsurprisingly, as the difference between grade and the test level increased, proportionally fewer students obtained a high enough score for admission into Talent Search. Olszewski-Kubilius & S.-Y. Lee (2011) found similar results when examining the impact of gifted students' age/nominal grade on above-level SAT, ACT, and EXPLORE scores.

purposes, populations, and conditions with which above-level testing should be attempted outside of a Talent Search setting.

Increasing Score Variability

Raising the ceiling is also important in gifted education research because a low test ceiling produces findings that may be plagued by restriction of range problems, which usually attenuate correlations, water down effect sizes, and cloud the interpretation of statistics (Kaplan & Saccuzzo, 2005). Moreover, a restriction of range makes examinees appear more alike than they really are, which causes problems in both research and practice (Johnsen & Corn, 2001). Warne (2009) gave the theoretical example of two first-grade students who score in the 99th percentile of math ability, saying, ". . . one of them may be able to do simple multiplication and the other one may be able to do pre-algebra. Even though their percentile score is the same, their mathematical abilities are different" (p. 50). This restriction of range is present in almost any score metric, although some metrics (like percentiles) have lower ceilings than others (such as scale scores or IQ-like scores).

Gifted education proponents have proposed above-level testing as a solution to the restriction of range problem often found in gifted education (Lupkowski-Shoplik, Benbow, Assouline, & Brody, 2003; Keating, 1975, 1976; Swiatek, 2007; VanTassel-Baska, 1996). Empirical evidence on above-level testing has supported claims about the increased variability of above-level test scores. For example, many studies associated with Talent Search programs have found that test scores were far more variable with above-level tests than with grade-level tests (e.g., Olszewski-Kubilius, 1998b;

VanTassell-Baska, 1986) and above-level test scores often form a distribution that is approximately normal (Keating & Stanley, 1972, p. 4; Lupkowski-Shoplik & Swiatek, 1999, p. 269). By raising the test ceiling, above-level tests also allow gifted children's test scores to become more variable and better manifest the differences among the gifted (Lubinski, Webb, Morelock, & Benbow, 2001; Olszewski-Kubilius & Kulieke, 2008; Olszewski-Kubilius & S.-Y. Lee, 2011).

The greater discrimination among gifted examinees of above-level tests is partially due to the increased variability among scores with above-level testing (e.g., Lupkowski-Shoplik et al., 2003; Olszewski-Kubilius & Kulieke, 2008; VanTassell-Baska, 1986). The importance of this improved discrimination among high ability students should not be understated. Benbow (1992), for example, has shown that above-level tests have the ability to detect differences among the top 1% of examinees and that the above-level test scores can make predictions about educational attainment, salary, and other important outcomes. Lubinski, et al. (2001) showed that the discrimination power of above-level tests even extends to the top .01% of ability. When one considers the poor discriminating power of regular grade-level tests among the top 5% of examinees, to be able to distinguish among the abilities in the top 1 in 10,000 students is a phenomenal property of above-level testing and one not to be treated lightly.

Improved Score Reliability

Advocates of above-level testing claim that above-level test scores are more reliable for their special populations than grade-level scores (Keating, 1975, 1976). The logic behind this claim is based on the fact that most grade-level tests are designed to

measure the largest possible number of students as efficiently as possible. This means that the majority of test items correspond to the middle-level ranges of ability. Because of the lower number of items corresponding to high levels of ability, the scores estimated from those items will usually be less reliable (Lohman & Korb, 2006; Minnema, Thurlow, Bielinski, & Scott, 2000). Therefore, more difficult tests will have more items corresponding to many gifted students' abilities, and the observed scores will have higher reliability than scores obtained from a grade-level test.

Kieffer, Reese, and Vacha-Haase (2010) used different logic to reach the same conclusion about grade-level tests generating poorly reliable data for gifted children. They stated that the constrained variance of gifted children's grade-level test scores theoretically drives down reliability coefficients. Because reliability can be understood as a squared correlation between true scores and observed scores, any constraints on the variance of observed scores will likely reduce reliability coefficients. Kieffer et al. (2010) provided a convincing theoretical example of how a grade-level test and a selected population (like gifted students) can combine to generate scores with very low reliability.

Despite the sound psychometric reasoning of these theoretical arguments and the support for them among researchers examining below-level test scores (e.g., Bielinski, Thurlow, Minnema, & Scott, 2000), the only reports of above-level reliability coefficients from a gifted education researcher that I have been able to find are from Stanley (1951). Even Stanley's report on reliability is of little use for today's researchers because of the age of the study. Stanley's (1951) study also suffers from the

fact that the coefficient is a split-half reliability coefficient corrected by the Spearman-Brown prophecy formula (in accordance with the accepted practice at the time). However, there is no evidence that the halves of the test were sufficiently equivalent. Also, Stanley used an instrument (the Nelson-Denny Reading Test) that has not since been used in above-level testing.

It seems that gifted education researchers quietly assume that the above-level tests they use will produce sufficiently reliable scores when administered to gifted students, despite the fact that these tests were not designed with such unusual examinees in mind. Test scores are a product of many different factors: sample characteristics, testing environment, test items, previous exposure that a child has had to test content, and many other issues. Because reliability is not a property of tests, but rather a property of test scores (Kieffer, et al., 2010; Thompson & Vacha-Haase, 2000; Vacha-Haase, Kogan, & Thompson, 2000), the assumption that above-level tests will produce high reliability coefficients may be erroneous. Above-level tests are administered to different populations under different conditions and for different reasons than when the same tests are administered as grade-level tests. For this reason alone, future researchers who conduct analyses on above-level test scores should report reliability information on their data. Indeed, current reporting standards in both education and psychology require *all* researchers to report the reliability of the data at hand (AERA, 2006; Wilkinson & the Task Force on Statistical Inference, 1999).

At least one researcher who was not directly concerned with gifted education has administered above-level tests and examined the ensuing reliability coefficients. Loyd

(1980) found that the most able students in her study obtained the most reliable scores with the highest level of the test she administered, even when the children were younger than the population that the test was designed for (p. 117). However, Loyd's exploration of reliability in above-level testing is incomplete because she still encountered ceiling effects that often prevented the most able students from obtaining highly reliable scores on some subtests (p. 97). Therefore, more research is needed to determine whether the assumptions on the reliability of above-level test scores are tenable.

Reliability coefficients are likely the most common measure of score reliability, but they are not the only one available to researchers. The standard error of measurement (SEM) is another viable option for reporting reliability information. However, because reliability coefficients and the SEM are algebraically related, the SEM still carries the assumption that it is constant across all score levels, which limits the usefulness of the SEM in examining the reliability of extreme scores. Researchers also have the option of reporting a conditional SEM, which varies according to observed score and is therefore better than a reliability coefficient or the regular SEM. The mechanics of producing a conditional SEM are beyond the scope of this article, but the interested reader should consult Kolen, Hanson, and Brennan (1992). However, the technical manuals for a few multi-level tests, such as the Cognitive Abilities Test (Lohman & Hagen, 2002, pp. 59-61), give conditional SEM values for different scores on different levels of the test, permitting researchers to estimate how much error would be reduced by administering a different test level.

Better Comparability and Use in Educational Planning

Despite the young age of some above-level testing examinees, many gifted education researchers believe that high ability students are often better compared to groups that consist of older children. In other words, children who are advanced cognitively should sometimes be compared to cognitive peers and not age peers. This is an implication of one definition of giftedness in which gifted children are understood as being in a more advanced stage of cognitive development than their age peers (Morelock, 1992). When a child's cognitive development is drastically out of sync with that of his or her age peers, that child has different educational needs than his or her age peers. Indeed, his or her needs may better resemble those of a regular developing older child (Morelock, 1992). Therefore, an above-level achievement test comparison to norms consisting of older children may provide better information and be more informative about the child's educational needs.

As researchers have interpreted above-level test scores, they have mostly come to the conclusion that such scores can be interpreted the same way that the scores would be interpreted for the test's norm population. For example, Gross (2004) administered above-level tests to her sample of highly gifted children (IQ 160+) and found that interpreting the test scores as if the children belonged to the older norm group was supported by her intense behavioral observations and interviews of her sample. This ease of interpretation makes sense under the theory that intellectual giftedness is merely a case of advanced cognitive development. It should be noted, however, that Gross used career interest inventories, personality tests, and educational planning tests in above-

level testing, and score interpretation of such tests may be radically different than above-level achievement test score interpretations.

The claim that above-level test scores from gifted children can be interpreted the same way as scores from a regular population taking the same test is bolstered by a study examining factor structure and measurement invariance between high school and gifted seventh grade students. Minor and Benbow (1996) found that the structure of test responses on the SAT-M was identical for both groups of students, as were the magnitude of the factor loadings and the item error variances. This study supports the claim that test results can be interpreted identically for high schoolers and gifted seventh graders, despite the age difference between the two groups. However, Minor and Benbow's study is flawed, because it relies on item parcels, which simulation studies have shown can distort item structure, hide a lack of invariance, and inflate goodness-of-fit statistics (Meade & Kroustalis, 2006; Nasser & Wisenbaker, 2003). Moreover, Minor and Benbow did not compare the invariance of item intercepts across groups, meaning that not all aspects of true measurement invariance have been investigated for *any* above-level test.

Regression Toward the Mean

Regression toward the mean is the statistical phenomenon where examinees who obtain extreme scores tend to obtain scores closer to the mean when retested. In other words, gifted students seem less gifted when retested and struggling students seem to improve when retested (on average). Regression toward the mean occurs any time two scores are not perfectly correlated (i.e., when $r \neq 1.0$ or -1.0). This imperfect correlation

can result from unreliable scores, the passage of time, or merely because two scores measure different constructs.

Regression toward the mean is a severe problem in gifted education. Lohman and Korb (2006) in their landmark article "Gifted Today but Not Tomorrow?" showed with real longitudinal data that about half of students who obtained scores in the top 3% of the Iowa Tests of Basic Skills composite battery did not obtain scores in the top 3% five years later (p. 465). Similarly, when Terman retested some children in his gifted sample about eight years after they were originally identified, he found that the average IQ had decreased. Some of these changes in scores ". . . were doubtless due to the statistical regression always found in a group of deviates selected on the basis of a fallible test . . ." (Burks et al., 1930, p. 45).

The formula for calculating the amount of regression to the mean is rather simple. First, one must obtain a predicted retesting *z*-score ($\hat{z}_2$) from the following equation:

$$\hat{z}_2 = r_{xy} \cdot z_1$$

where r_{xx} is the test-retest reliability of the scores, and z_1 is the *z*-score of the first obtained score. Thereafter, the amount of regression toward the mean is calculated by

$$|\hat{z}_2| - |z_1|$$

which can easily be converted back to the units in which that the original scores measured.[3]

[3] For a more detailed and technical treatise on the relationship between reliability, high ability, and regression toward the mean, see Ziegler and Zielger (2009).

Therefore, the amount of regression toward the mean is a result of two values: the original observed scores and the reliability of the observed scores. Regression toward the mean should be reduced by either (a) obtaining scores closer to the mean, or (b) increasing reliability. Theoretically, above-level tests serve both of these functions, because gifted children's scores are usually closer to the mean of the norm population of the above-level test (e.g., Barnett & Gilheany, 1996) and—as stated earlier—above-level tests should also raise reliability coefficients. However, the impact of above-level testing on regression toward the mean has not been empirically tested.

Other Research of Note on Above-Level Testing

Since the late 1970's, above-level testing has become a widely accepted practice in gifted education, due mostly to the promising results from Talent Search programs and the test scores' strong ability to predict outcomes important to stakeholders. Most of this evidence stems from SMPY. For example, Benbow (1992) showed that pre-adolescents' SAT scores are moderately good predictors of AP Calculus test scores, College Board Achievement Test scores, the number of math and science courses taken in high school, the selectivity of the college attended, and undergraduate GPA. Later follow-ups of the SMPY sample or subsets of the sample showed that the predictive power of above-level testing extended even further into the future. SMPY students who obtained high scores on above-level tests were later 25 times more likely than average to obtain a doctorate (Lubinski et al., 2001, p. 725). Also, the top quartile of Talent Search students were more likely than those in the bottom quartile to earn a higher income than average (effect size $h = .16$), acquire a patent ($h = .18$), and obtain tenure at a university

(h = .28) (all effect sizes from Wai, Lubinski, & Benbow, 2005, pp. 486, 487; see also Lubinski & Benbow, 2006). To say that these results are impressive would be an understatement, especially because some of these outcomes occurred decades after the above-level test scores were obtained. Oszewski-Kubilius (1998a) appropriately stated the usefulness of the SAT as an above-level instrument when she said, "Rarely has the field of education had such powerful predictive tools at its disposal" (p. 136).

Extensive research has been performed in order to determine when above-level testing is most appropriate for Talent Search purposes. This is because the tests are between two and five years above the child's grade level and it is in the child's and the program administrators' best interest to administer such a difficult test only if necessary. Empirical studies show that testing four or five levels above grade should only be done if the child can obtain a score at the 95th percentile or higher on a regular grade-level test (Ebmeier & Schmulbach, 1989; Lupkowski-Shoplik & Swiatek, 1999), although the standard may be lowered if the test level is closer to the student's grade or if the program isn't as intensive or selective as Talent Search (Olszewski-Kubilius et al., 1989).

Threlfall and Hargreaves (2008) conducted a study to see if 475 gifted 9-year-old children use the same problem solving strategies for math items as 230 average 13-year-old children. Giving both groups novel problems, the researchers examined the proportion of students in the groups who chose to use various problem solving strategies. Despite the large number of students in each group, Threlfall and Hargreaves did not find any statistically significant differences between the proportion of students who used each problem solving strategy. This lends credence to the belief that above-level test

scores can be interpreted for gifted students the same way that the test scores can be interpreted for the norm group. However, Threlfall and Hargreaves used item types that neither subject group had ever seen before, whereas in most above-level testing the older group would have been exposed to most—if not all—item types on an achievement test.

A final, more miscellaneous study on above-level testing should be noted. Pervasive evidence of gender differences among the top echelons of mathematical ability (e.g., Benbow & Stanley, 1980) prompted a study on item bias of the SAT-M with regards to gender (Benbow & Wolins, 1996). In the study, the researchers found that despite most items on the test being easier for the male gifted adolescents, there was no evidence of any meaningful item-level bias in the SAT-M. To date, this is the only study on item-level bias with above-level testing. Other group differences in above-level test scores (e.g., differences among ethnic groups) warrant further investigations of item bias in above-level testing.

Discussion

The research performed thus far in above-level testing has provided a firm foundation for research into how above-level tests function with gifted populations. The findings also have led to experimentation in above-level testing in non-academic domains (Achter et al., 1996; Gross, 2004). However, there are still some issues that remain unresolved. Most importantly, research on the psychometric properties of above-level test scores is mostly limited to the SAT and its subtests. Some work has been done on other Talent Search tests, such as EXPLORE (Colangelo, Assouline, & Lu, 1994; Lupkowski-Shoplik & Swiatek, 1999; Olszewski-Kubilius & Turner, 2002) and the

Secondary School Admissions Test (Lupkowski-Shoplik & Assouline, 1993; C. J. Mills & Barnett, 1992). But these studies do little beyond showing a raised test ceiling or establishing cutoffs on grade-level tests for eligibility to take an above-level test for Talent Search admission. Given the widespread endorsements of above-level testing of the gifted (e.g., Assouline et al., 2009; Colangelo, Assouline, & Gross, 2004; Gross, 1999; Rogers, 2002), more psychometric studies are needed to understand how items and tests "behave" when administered to a younger, gifted sample. Also, more tests should be evaluated for their suitability for above-level testing.

Evidence for validity of interpretations of above-level tests is also lacking in the published literature. Despite statistically identical structures and relatively similar interpretation of above-level testing scores, most researchers and practitioners who conduct above-level testing use above-level academic achievement tests as aptitude tests for younger, gifted students (e.g., Assouline et al., 2009; Lubinski & Benbow, 1994; Stanley, 1977). In other words, researchers are using tests of past learning (i.e., achievement tests) as estimators of future potential (i.e., aptitude tests).

Some readers may find a contradiction between using an achievement test in the service of evaluating aptitude and the claim that above-level test scores can be interpreted as if the gifted students were members of the older norm population. The contradiction is a real one, despite a conceptualization that the distinction between achievement and aptitude tests is unclear (e.g., Merwin & Gardner, 1962; Schmeiser & Welch, 2006; Zwick, 2006). Modern theorists recognize aptitude as a product of interest, motivation, affect, the specific environment, intelligence, meta-cognitive

abilities, and academic experiences (Corno et al., 2002). At most, above-level tests may measure the knowledge-based and reasoning aspects of academic aptitude. The exact degree to which a given above-level test measures aptitude or achievement may be the result of a wide variety of factors, some of which may be unique to each examinee (e.g., the test level, the age of the child, the opportunity to learn the more advanced material, test content). Further research is needed on this issue and whether above-level testing can equal or surpass traditional ability tests in measuring high levels of academic aptitude.

So what construct(s) do above-level academic achievement tests measure? At the very least, the SAT, ACT, EXPLORE and similar tests measure the suitability of participating in a Talent Search program. This interpretation of above-level test scores is likely beyond dispute. The only other specific interpretation that has been studied is as a measure of academic preparedness for acceleration. Unfortunately, the only studies that have examined this interpretation have been in conjunction with the Iowa Acceleration Scale (Assouline et al., 2009) and are not peer-reviewed (see Appendix D in Colangelo et al., 2004, for a summary of this research). The lack of an interpretation framework of above-level test scores outside of a Talent Search context may be one of the great stumbling blocks that prevent school personnel from using above-level testing more often.

There is also little understanding of the circumstances under which above-level tests should be administered outside of a Talent Search or grade acceleration context. Can above-level tests be used to identify gifted children in a local school district? Are

above-level tests useful for program evaluation or accountability purposes? Do above-level tests manifest racial bias that is absent when they are administered to regular samples? How can above-level testing impact day-to-day instruction in schools? Should practitioners distinguish between the test level administered to a gifted child and the norm group used for comparison when interpreting scores? What are the cognitive response process that a gifted child uses when answering above-level test items? These questions and others are in dire need of investigation before above-level testing becomes a common practice outside of Talent Search programs. Researchers could also explore more advanced psychometric questions, such as the possibility of growth modeling to measure academic progress, the investigation of above-level tests with item response theory methods, or the impact of linking methods on observed above-level test scores. Studies examining all of these issues would broaden understanding of exactly how above-level testing affects the psychometric properties and interpretation of scores.

Many of these new issues in above-level testing will require a change in research on how the practice has thus far been conducted. For example, improving the interpretation of above-level test scores and understanding what construct(s) they may be measuring may be difficult to determine with the SAT. A multi-level, vertically aligned test, such as the Iowa Test of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2001) would be a more appropriate instrument for this type of research, because the nationally representative norms and carefully documented item content at each test level would permit researchers to understand the relative influence of student ability and test content on above-level test scores. The ITBS and similar instruments would also be more

appropriate for studying growth modeling, program evaluation, and many other topics related to above-level testing.

Also, gifted education researchers will likely need to branch out from Talent Search samples in order to better understand above-level testing. The vast majority of the above-level testing research cited in this literature review is an outgrowth of Talent Search programs, which Matthews (2008) has criticized for several reasons: a total lack of random assignment or sampling, an operational definition that equates giftedness with a high test score, and a lack of economic or cultural diversity. All of these characteristics limit generalizability of Talent Search findings—including those reviewed in this article. To combat these problems, future researchers must use above-level testing with gifted non-Talent Search samples.

Alternatives to Above-Level Testing

Above-level testing is not the only feasible method of collecting high quality information about intellectually gifted children's abilities or achievement. Practitioners have the option of selecting tests with naturally high ceilings for purposes of identification. Traditional intelligence tests, such as the Stanford-Binet 5 or the Wechsler Intelligence Scale for Children — Fourth Edition, have high ceilings, sufficiently high reliability for intellectually gifted/high intelligence examinees, and a clear interpretive framework supported by a large body of research (Roid, 2003; Wechsler, 2003). The Screening Assessment for Gifted Elementary and Middle School Students — Second Edition also has a high ceiling and acceptable reliability in the gifted range (Johnsen & Corn, 2001).

For purposes of tracking learning and educational progress, however, options for evaluating intellectually gifted children are more limited. One possible alternative to above-level testing is to use computer adaptive testing (CAT; Gershon, 2005) to track a gifted child's progress through a curriculum. A suitable CAT assessment would need a large pool of items that span a continuum across several grade levels—which would likely make CAT financially unfeasible unless the local district or state already had such a system implemented as part of their regular assessment procedures. If practitioners do not wish to make cross-grade score comparisons, then content-based assessments are also a viable possibility. However, because many of these assessments do not meet the rigorous standards of psychometric practice, these may not be suitable for research or high-stakes decisions.

Conclusion

Overall, the research examined in this literature review supports the practice of above-level testing. As researchers and practitioners perform above-level testing, they can be assured that the basic assumptions behind the practice are psychometrically sound—especially as those assumptions relate to test ceilings and gifted students' score variability. However, further research is needed to investigate the reliability of above-level testing scores, the suitability of more instruments for above-level testing, regression toward the mean, the usefulness of the procedure in non-Talent Search settings, and the validity of score interpretations.

In this dissertation, I will attempt to shed light on some of the areas of above-level testing that are thus far uninvestigated. Specifically, I designed this study to examine five research questions:

1. What is the internal consistency reliability of the global battery, reading/language arts, and mathematics scores drawn from an above-level administration of an achievement test?
2. Do gifted children make larger achievement gains in overall, reading/language arts, and mathematics scores than average students in a more advanced grade?
3. Do demographic variables (gender, ethnicity, and SES) influence the initial scores or rate of overall, reading/language arts, and mathematics score growth of gifted students?
4. What is the relationship between initial above-level overall, reading/language arts, and mathematics scores and rate of score growth?
5. What percentage of overall, reading/language arts, and mathematics score variance is explainable through time, demographic variables, and cohort membership?

In addition to the knowledge that will be gained through the examination of these questions, this study is designed to overcome some of the criticisms that Matthews (2008) made of Talent Search research. For example, as described in the following chapter, the sample in this study will be less selective than a Talent Search sample and is more economically and ethnically diverse than many samples that have been described in previous research on above-level testing. Moreover, this studied occurred in the

context of a school district and regular gifted education practices—not the extracurricular setting of Talent Search programs or the rare pre-screening for grade acceleration. Through these research questions and design, I hope to gain a greater understanding of above-level test scores and their interpretation.

CHAPTER III

METHODS

Participants

The participants in this study were all the students at a gifted middle school magnet program located in a mid-sized district in the southern United States during the 2008-2009 and 2009-2010 school years. The students in the study were divided into four cohorts. Figure 1 shows the grade level for each cohort at each point in the study. Cohort 1 consisted of those students who were in the eighth grade during the first year of the study (i.e., the 2008-2009 school year). Cohort 2 consisted of students who were in the seventh grade during the first year of the study. Cohort 3 consisted of students who were in the sixth grade during the first year of the study. Cohort 4 consisted of students who were in the fifth grade during the first year of the study (i.e., they did not enter middle school until the second year).

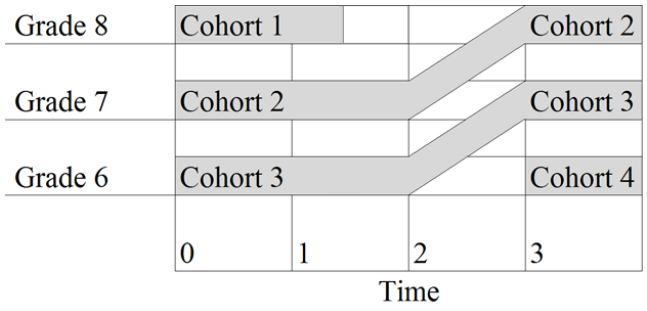

Figure 1 Relationship of time points in the study, cohort numbers, and grade levels. All time points are six months apart. Members of Cohort 1 were only present for time point 1. Members of Cohort 2 and 3 were present for all time points, but advanced grades at time point 2. Members of Cohort 4 were only part of the study at time point 3.

Because the magnet program that serves students in grades 6-8, Cohort 1 was only measured during the first year of the study (Time 0 and Time 1) and Cohort 4 was only measured in the second year (Time 3) of the study. Therefore, the research questions that address rate of score gains were only investigated with Cohorts 2 and 3. All other research questions were addressed with data from members of all four cohorts.

During Year 1 of the study, the program included 138 students. During the second year, the program included 170 students. The exact cohort sizes varied between 37 and 61 students, with cohorts tending to grow larger as time progressed. The exact size of each cohort at each point in the study is displayed in Figure 2. In total, 224 students were tested at least once, with 435 above-level tests administered in total. One hundred twenty-six students were White (56.2%), 72 were Hispanic (32.1%), 22 were African American (9.8%), three were Asian American (1.3%), and one was of unknown ethnicity (0.4%). One hundred students were male (44.6%), 123 were female (56.2%), and one student (0.4%) was of an unknown gender.

The cohorts varied in their gender and demographic makeup. Cohorts 2 and 4 were both 40.0% male and 60.0% female. In contrast, Cohort 1 (48.7% male and 51.3% female) and Cohort 3 (53.7% male and 46.3% female) had a much more even gender balance. Cohort 1 was noticeably less Hispanic than the other cohorts (20.5% compared to at least 31.5% for all other cohorts). Cohort 3, on the other hand, had proportionally far fewer African American students (just 3.7%) compared to the other cohorts (between 8.3% and 15.7%). Finally, Cohort 2 had a drastically lower proportion of White students (only 47.1%) compared to the other cohorts (all 53.3% or greater).

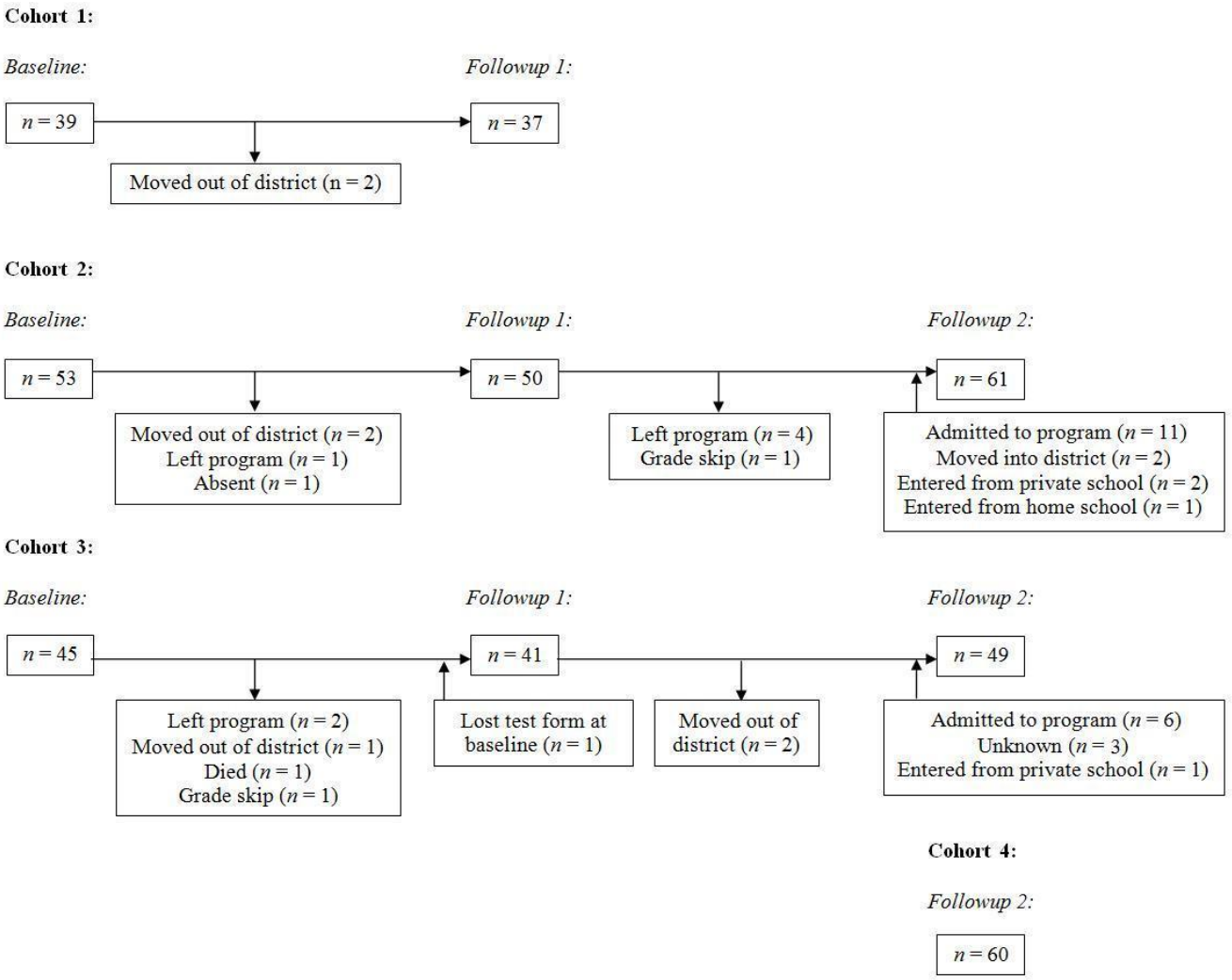

Figure 2 Subject flow through the study.

Instruments

The instruments used for this study were Form C of the Iowa Tests of Basic Skills (ITBS; Hoover, Dunbar, & Frisbie, 2001) and Form C of the Iowa Tests of Educational Development (ITED; Forsyth, Ansley, Feldt, & Alnot, 2001). These two tests are well-respected measures of academic achievement that permit comparisons across grades on a scaled metric. Comparisons can also be made across the two instruments because the ITED is merely an upward extension of the ITBS (Forsyth, Ansley, Feldt, & Alnot, 2003). At each testing point, students were given the ITBS or ITED test level that was designed for the grade two years above their actual grade (i.e., sixth grade students took the eighth grade test, seventh grade students took the ninth grade test, and eighth grade students took the tenth grade test) as part of normal procedure at the magnet program. In Year 1, students were administered the ITBS or ITED in November 2008 and May 2009. In Year 2, students received the ITBS or ITED only in May 2010. Year 2 only had one measurement time because of budget constraints that resulted from the current nationwide recession.

Coding and Statistical Power

Variables were coded as follows: the time points of the baseline (November 2008), first followup (May 2009), and second followup (May 2010) were coded as 0, 1, and 3, respectively. These values were chosen so that each unit represented six months and the spacing between values was proportional to the amount of time that passed between each testing, which is common practice in longitudinal studies (Hedeker &

Gibbons, 2006). Cohorts were dummy coded so students in Cohort 3 were the baseline group for all comparisons.

Because of the limited number of students who were tested in all three time points (n = 84), ethnicity groups were combined in order to increase statistical power. White and Asian American students were combined because these students were overrepresented compared to the district's general student population. African American and Hispanic students were also combined into a group of students in underrepresented ethnicities. This is consistent with the ethnic/racial makeup of most gifted programs nationwide (Konstantopoulos, Modi, & Hedges, 2001; McBee, 2006, 2010; Yoon & Gentry, 2009). Socioeconomic status was operationalized so that students who were eligible to receive free or reduced lunch were labeled as low-SES. Students who were not eligible to receive free or reduced lunch were combined into one SES category of middle- or high-SES students. Finally, the critical p-value (α) was changed from the traditional .05 value to .10 in order to increase statistical power and compensate for the relatively small sample size in this study.

Analysis

Research question 1 was investigated using KR20 values (Kuder & Richardson, 1937) to determine internal consistency reliability. KR20 values were calculated for each cohort at each measurement time and data were not combined across cohorts, test levels, or measurement occasions.

For research questions 2-5, which involved the investigation of student growth over time, HLM was used across the three time points. HLM is a necessary statistical

procedure in this case because the data points are dependent, with measurement occasions nested in persons. Moreover, HLM is widely recognized as an appropriate statistical method for examining change and growth within persons (Ferron, Hogarty, Dedrick, Hess, Niles, & Kromrey, 2008). There were three hierarchical linear models to answer research questions 2-5—one model each for the core battery, reading/language arts subtest, and mathematics subtest scores.

The model for each dependent variable was created through an exploratory step-up procedure. First, a baseline model with no predictors was created and called Model 1. When the dependent variable is the total above-level scale score, this model was defined with the level-1 equation of

$$Total_{ij} = \beta_{0j} + e_{ij}$$

and the level-2 equation of

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

which combine to form

$$Total_{ij} = \gamma_{00} + u_{0j} + e_{ij}$$

as a general equation in which γ_{00} represents the grand mean of all measurements across all time points, u_{0j} represents the deviation between the mean of a particular cluster of measurements (i.e., each person) from the grand mean, and e_{ij} represents the remaining level-1 error between the cluster mean and the individual measurement.

Model 1 can also be used to calculate the intraclass correlation (ICC), which is a measure of the amount of total variance that is between level-2 units (Raudenbush & Bryk, 2002). The ICC is calculated as

$$ICC = \frac{\tau_{00}}{\sigma^2 + \tau_{00}}$$

where τ_{00} is the between cluster (i.e., person) variance, σ^2 is the within cluster variance, and $\sigma^2 + \tau_{00}$ is the total variance observed across all measurements in the study. Because it is a percentage of total variance that is attributable to cluster, ICC can range from 0 to 1. Higher ICCs indicate that strong clustering effects and greater homogeneity within clusters. Because clusters in longitudinal studies are persons and the measurements are relatively close together, ICC values were expected to be high.

A second model with time (a level-1 variable) as the only predictor was built and called Model 2. When the dependent variable is the total battery score, Model 2 is represented by the level-1 equation

$$Total_{ij} = \beta_{0j} + \beta_{1j}(Time_{ij}) + e_{ij}$$

and the level-2 equations of

$$\beta_{0j} = \gamma_{00} + u_{0j}$$

and

$$\beta_{1j} = \gamma_{10} + u_{1j} \quad .$$

The first level-2 equation is identical to the level-2 equation in the HLM model that consists of no predictors (i.e., Model 1). The second level-2 equation produces a Beta coefficient for the time variable and represents the change in total battery scores for each time unit (i.e., six months) that passed.

The level-1 equation and two level-2 equations can be expressed together as

$$Total_{ij} = \gamma_{00} + \gamma_{10}(Time_{ij}) + u_{0j} + u_{1j}(Time_{ij}) + e_{ij}$$

in general form. The γ_{00}, u_{0j}, and e_{ij} terms are interpreted in Model 2 exactly as they were in Model 1 (as a grand mean, cluster mean deviation, and level-1 error, respectively). The γ_{10} term is interpreted as the average increase in score gain for each unit of time that passes (i.e., the slope of a line that would track a student's score gains). The u_{1j} represents the deviation of each person's slope coefficient from the overall average slope (γ_{10}).

It is important to distinguish between the γ values and the u terms, the e_{ij} term. The γs represent an overall mean fixed effect that applies to the sample in general. However, each level-2 unit (i.e., person in a longitudinal study) has its own u_{0j} and u_{1j} values. Moreover, each individual level-1 measurement has its own e_{ij} term for which the HLM computer program estimates the variance. These values can be used to examine the random effects—that is, the deviations from the average model—present in the data. Specifically, the variance of the e_{ij} is defined as the remaining level-1 variance, symbolized by σ^2. The variances and covariance of the u terms can be used to create a matrix that represents the variability and relationship of the random effects. This is called the G matrix and is represented as

$$G = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix}$$

where τ_{00} represents the variance of u_{0j} values, τ_{11} represents the variance of u_{1j} values, and τ_{01} and τ_{10} both represent the covariance between u values because the G matrix is symmetrical (Ferron et al., 2008; Raudenbush & Bryk, 2002). If G is

standardized to form a correlation matrix, then the diagonal terms are equal to 1.0 and the off-diagonal terms are converted into correlation values.

Thereafter, a series of models were investigated with time and a single level-2 variable as predictors. These models were named Models 3, 4, 5, and 6, each corresponding to a single level-2 independent variable in the study. For example, Model 6—which has only SES as a level-2 independent variable—consists of the level-1 equation of

$$Total_{ij} = \beta_{0j} + \beta_{1j}(Time_{ij}) + e_{ij}$$

and the level-2 equations:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(SES_j) + u_{0j}$$

and

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

which combine to form

$$Total_{ij} = \gamma_{00} + \gamma_{10}(Time_{ij}) + \gamma_{01}(SES_j) + u_{0j} + u_{1j}(Time_{ij}) + e_{ij}$$

as an equivalent general equation. Models 4-6 were also interpreted with an HLM effect size that is analogous to Cohen's *d*. This effect size is calculated with the following formula:

$$\delta = \frac{\gamma_{0j}}{\sqrt{\sigma^2 + \tau_{00}}}$$

where γ_{0j} is the fixed effect for a dichotomous independent variable coded 0 and 1 (Spybrook, 2008, p. 285).

Because there was an association between students' ethnicity and SES (r = -.354, $p < .001$), another model (Model 7) was investigated with time, ethnicity, and SES as predictors in order to investigate the relative strength of the two level-2 predictors when placed in the same model. Thereafter, a model with all available independent variables was created: Model 8. This model is very similar to the equations given above for Model 6 because both Model 6 and Model 8 have the same level-1 variable (time) and only differ in that Model 8 has three more level-2 independent variables than Model 6. Therefore, Model 8 is expressed with the level-1 equation of

$$Total_{ij} = \beta_{0j} + \beta_{1j}(Time_{ij}) + e_{ij}$$

and the level-2 equations of

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(Cohort1_j) + \gamma_{02}(Cohort2_j) + \gamma_{03}(Cohort4_j) + \gamma_{04}(Gender_j) + \gamma_{05}(Eth\_Und_j) + \gamma_{06}(SES_j) + u_{0j}$$

and

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

which combine to form

$$Total_{ij} = \gamma_{00} + \gamma_{10}(Time_{ij}) + \gamma_{01}(Cohort1_j) + \gamma_{02}(Cohort2_j) + \gamma_{03}(Cohort4_j) + \gamma_{04}(Gender_j) + \gamma_{05}(Eth\_Und_j) + \gamma_{06}(SES_j) + u_{0j} + u_{1j}(Time_{ij}) + e_{ij}$$

in general form.

After Models 1-8 were created and examined for all dependent variables, interactions between time and level-2 predictors were investigated. Although many interactions among independent variables were possible, only one interaction was

investigated at a time because investigating a large number of interactions with such a small sample may carve the dependent variable's variance into so many pieces that power is lost and statistical significance difficult to obtain (Raudenbush & Bryk, 2002). Any statistically significant interactions were then added to Model 8 to produce Model 9. Finally, a parsimonious model was created by eliminating any non-statistically significant fixed or random components from Model 9. All models in this study were analyzed with restricted maximum likelihood estimation using the EM algorithm in the computer program HLM 6.08 (Raudenbush, Bryk, & Congdon, 2009).

Research question 2 was answered by investigating the slope parameter (i.e., change over time) of the parsimonious models' equations and comparing the growth observed in cohorts 2 and 3 with the gain that would be expected from intellectual peers (i.e., older students), according to the norms published in the ITBS and ITED manuals (Forsyth et al., 2003; Hoover et al., 2003).

Research question 3 was answered through the same parsimonious HLM equations. A criterion was set *a priori* that any independent variable that has a statistically significant relationship was deemed to have an impact on either the rate of student score gains or initial student score. The importance of these independent variables was investigated with the change in level-2 Pseudo-R^2 as they are added to the model.

Research question 4 was investigated in the G matrix through the τ_{01} term in the G matrix. Because τ_{01} is unstandardized in the G matrix, it was converted into a correlation coefficient. A positive τ_{10} between these two values would indicate that

students who had a higher above-level test score at Time 1 also made the greatest gains in learning—which would be in accordance with theory (e.g., Carroll, 1993; Dai, 2010; Eisner, 2002; Gagné, 2005; Ruf, 2005).

Research question 5 was investigated with Pseudo-R^2 statistics that represented the proportional reduction in prediction error calculated for each level from the results of the HLM models described above. Multilevel models do not permit the estimation of true R^2 effect sizes that represent the proportion of variance explained by the independent variables for a variety of reasons. First, the possible presence of random effects in the models, the proportion of explained variance may vary from cluster to cluster (or in the case of this study, from person to person). Second, the partitioning of variance into two different levels prohibits the calculation of a single effect size that represents the proportion of explained variance. Finally, R^2 is an effect sized based upon ordinary least squares regression, whereas the HLM models are estimated through maximum likelihood or restricted maximum likelihood estimation algorithms (McCoach, 2010a).

The Pseudo-R^2 used in this study is from Raudenbush and Bryk (2002) and examines the decrease in σ^2 (for level 1 of the model) or τ_{00} (for level 2) that occurs when covariates are added to a model. All Pseudo-R^2 statistics in this study were calculated by comparing the models to the intercept-only model (Model 1), which had no predictors. Therefore, all Pseudo-R^2 statistics in this dissertation represent the decrease in the corresponding level's error variance when the covariate(s) are added to

the baseline model. More details about the Pseudo-R^2 can be found in Hox (2002), McCoach (2010a, 2010b), and Raudenbush and Bryk (2002).

CHAPTER IV

RESULTS

Descriptive Statistics

Descriptive statistics for all subtests and cohorts are displayed in Tables 1-3. The tables also display the descriptive statistics for the corresponding test level's national norms, which are taken from Hoover et al. (2003, p. 73) and Forsyth et al. (2003, pp. 57-58). All means and standard deviations are displayed in the scale score metric that permits comparisons across grades.

Tables 1-3 also show the skewness and kurtosis statistics for all above-level test administrations in this study. All skewness and the majority of kurtosis values are within the range of $\pm$ 1, indicating distributions that are approximately normal. Seven of the 66 kurtosis values (10.6%) are outside of the $\pm$ 1 range, with five of these being distributions for Cohort 2. However, only one kurtosis value is statistically different from 0 when $\alpha = .05$ (the vocabulary subtest distribution for Cohort 2 at Time 0). Moreover, the cohort's later skewness and kurtosis values for vocabulary subtests do not indicate a consistent pattern of kurtosis (-.775 at Time 1 and .824 at Time 3), which likely indicates that the extreme kurtosis value at Time 0 is likely due to random error.

Table 1
Descriptive Statistics for Gifted Grade 6 Cohorts 3 and 4 and National Grade 8 Norms[a]

	Reading	Language	Math	Total
Time 0, Gifted Grade 6: Cohort 3, Fall 2008[b]				
Mean	**248.9**	**248.3**	**238.7**	**245.2**
SD	**32.9**	**40.1**	**24.3**	**26.9**
Skewness	**-.060**	**-.059**	**.261**	**.480**
Kurtosis	**-.576**	**-.357**	**.006**	**-.449**
KR20	**.865**	**.903**	**.776**	**.931**
SEM	**12.09**	**12.49**	**11.5**	**7.07**
National Norms: Grade 8, Fall				
Mean	242.3	245.4	244.1	244.1
SD	32.8	41.0	31.6	32.6
KR20	.944	.957	.939	.980
SEM	7.76	8.50	7.80	4.61
Time 1, Gifted Grade 6: Cohort 3, Spring 2009[c]				
Mean	**268.8**	**257.4**	**256.0**	**260.5**
SD	**35.4**	**38.1**	**20.8**	**27.0**
Skewness	**-.190**	**-.310**	**.685**	**.019**
Kurtosis	**-.248**	**-.075**	**1.075**	**-.005**
KR20	**.880**	**.891**	**.725**	**.930**
SEM	**12.26**	**12.58**	**10.91**	**7.14**
Time 3, Gifted Grade 6: Cohort 4, Spring 2010[d]				
Mean	**256.9**	**249.8**	**241.5**	**249.7**
SD	**25.6**	**30.0**	**22.3**	**21.9**
Skewness	**-.315**	**-.280**	**-.997**	**-.571**
Kurtosis	**-.158**	**-.261**	**.981**	**.419**
KR20	**.800**	**.829**	**.750**	**.902**
SEM	**11.45**	**12.41**	**11.15**	**6.86**
National Norms: Grade 8, Spring				
Mean	248.8	251.6	251.0	251.0
SD	34.1	42.6	33.2	33.6
KR20	.950	.960	.949	.982
SEM	7.62	8.52	7.50	4.51

[a]Bold indicates scores of students in the present study at each time point. National norm statistics are from Hoover et al., 2003, p. 73.
[b]$n = 45$.
[c]$n = 41$.
[c]$n = 60$.

Table 2
Descriptive Statistics for Gifted Grade 7 Cohorts 2 and 3 and National Grade 9 Norms[a]

	Vocabulary	Reading Comprehension	Reading Total	Spelling	Revising Written Materials	Math Concepts & Problem Solving	Math Computation	Math Total	Total Battery
Time 0, Gifted Grade 7: Cohort 2, Fall 2008[a]									
Mean	**258.2**	**271.3**	**264.7**	**258.7**	**273.2**	**263.3**	**240.7**	**255.9**	**264.6**
SD	**23.7**	**31.1**	**25.6**	**34.0**	**36.6**	**30.2**	**31.8**	**28.2**	**26.7**
Skewness	**-.057**	**-.055**	**-.029**	**.170**	**-.544**	**-.052**	**.238**	**.220**	**-.190**
Kurtosis	**1.706**	**-.505**	**.161**	**-.334**	**-.078**	**1.112**	**.051**	**1.102**	**.557**
KR20	**.858**	**.867**	**.921**	**.842**	**.878**	**.845**	**.786**	**.893**	**.960**
SEM	**8.93**	**11.34**	**7.2**	**13.51**	**12.78**	**11.89**	**14.71**	**9.22**	**5.34**
National Norms: Grade 9, Fall									
Mean	251.9	252.4	252.2	254.7	254.7	254.2	254.5	254.3	253.7
SD	31.1	42.4	34.7	35.9	43.0	36.7	37.4	33.8	34.0
KR20	.908	.915	.950	.835	.911	.870	.853	.946	.972
SEM	9.43	12.36	7.76	14.58	12.83	13.23	14.34	7.85	5.69
Time 1, Gifted Grade 7: Cohort 2, Spring 2009[b]									
Mean	**268.4**	**288.1**	**278.3**	**263.4**	**286.9**	**275.1**	**247.6**	**266.0**	**277.1**
SD	**16.0**	**35.1**	**24.0**	**30.5**	**29.9**	**26.6**	**23.2**	**22.7**	**21.4**
Skewness	**-.176**	**.001**	**-.095**	**.282**	**-.251**	**-.669**	**-.029**	**-.416**	**-.348**
Kurtosis	**-.775**	**-.682**	**-1.052**	**-.570**	**-.771**	**-.273**	**-1.030**	**-.812**	**-.516**
KR20	**.793**	**.882**	**.912**	**.808**	**.826**	**.831**	**.553**	**.829**	**.937**
SEM	**7.28**	**12.06**	**7.12**	**13.36**	**12.47**	**10.94**	**15.51**	**9.39**	**5.37**
Time 3, Gifted Grade 7: Spring, 2010[c]									
Mean	**264.1**	**274.4**	**269.2**	**263.7**	**274.5**	**282.3**	**242.8**	**269.4**	**271.5**
SD	**20.6**	**35.7**	**26.8**	**33.0**	**41.8**	**27.4**	**33.6**	**26.0**	**27.3**
Skewness	**.065**	**.014**	**.147**	**.442**	**-.521**	**-.210**	**.539**	**.233**	**.004**
Kurtosis	**-.779**	**-.650**	**-.878**	**-.171**	**.146**	**-.500**	**.211**	**-.323**	**-.383**
KR20	**.864**	**.900**	**.934**	**.833**	**.910**	**.860**	**.821**	**.897**	**.963**
SEM	**7.60**	**11.29**	**6.89**	**13.49**	**12.54**	**10.25**	**14.22**	**8.34**	**5.25**
National Norms: Grade 9 Spring									
Mean	258.2	258.8	258.5	260.4	260.2	259.9	259.6	259.8	259.5
SD	32.7	44.4	35.8	37.0	43.3	38.0	39.0	34.9	34.5
KR20	.918	.921	.951	.864	.922	.902	.878	.958	.976
SEM	9.36	12.48	7.92	13.64	12.09	11.90	13.62	7.15	5.34

Note. The reading test consists of the vocabulary and reading comprehension subtests combined. The mathematics test consists of the match concepts and math computation subtests combined. The total battery consists of all items from all subtests combined.

[a]Bold numbers indicate scores of students in the present study at each time point. National norm statistics are from Forsyth et al., 2003, p. 57.

[b]$n = 53$.

[c]$n = 50$.

[d]$n = 49$.

Table 3
Descriptive Statistics for Gifted Grade 8 Cohorts 1 and 2 and National Grade 10 Norms[a]

	Vocabulary	Reading Comprehension	Reading Total	Spelling	Revising Written Materials	Math Concepts & Problem Solving	Math Computation	Math Total	Total Battery
Time 0, Gifted Grade 8: Cohort 1, Fall 2008[a]									
Mean	**265.0**	**275.8**	**270.3**	**265.3**	**275.8**	**272.0**	**248.3**	**265.0**	**271.1**
SD	**31.7**	**45.9**	**36.1**	**43.2**	**41.7**	**40.9**	**34.8**	**35.4**	**33.1**
Skewness	**-.340**	**-.598**	**-.555**	**.364**	**-.115**	**-.270**	**-.552**	**-.132**	**-.141**
Kurtosis	**.262**	**-.365**	**-410**	**-.544**	**-.136**	**-.969**	**-.136**	**-1.212**	**-.918**
KR20	**.894**	**.934**	**.951**	**.908**	**.911**	**.914**	**.835**	**.930**	**.973**
SEM	**10.32**	**11.79**	**7.99**	**13.10**	**12.44**	**11.99**	**14.14**	**9.37**	**5.44**
National Norms: Grade 10, Fall									
Mean	260.1	261.8	261.3	263.4	263.0	262.8	262.7	262.8	262.3
SD	33.0	44.9	36.0	37.3	44.0	38.5	39.3	35.4	35.1
KR20	.915	.918	.950	.852	.920	.892	.868	.954	.975
SEM	9.62	12.86	8.05	14.35	12.45	12.65	14.28	7.59	5.55
Time 1, Gifted Grade 8: Spring 2009[b]									
Mean	**271.7**	**292.1**	**283.2**	**277.8**	**293.3**	**285.9**	**256.0**	**257.9**	**285.6**
SD	**35.9**	**50.1**	**40.2**	**34.8**	**48.3**	**45.6**	**34.0**	**38.9**	**37.7**
Skewness	**-.493**	**-.659**	**-.723**	**.028**	**-.435**	**-.589**	**.335**	**-.281**	**-.554**
Kurtosis	**-.096**	**-.619**	**-.138**	**-.221**	**-.791**	**-.571**	**-.355**	**-.780**	**-.506**
KR20	**.916**	**.946**	**.982**	**.860**	**.936**	**.947**	**.825**	**.947**	**.981**
SEM	**10.40**	**11.64**	**5.39**	**13.02**	**12.22**	**10.50**	**14.22**	**8.96**	**5.20**
Time 3, Gifted Grade 8: Spring, 2010[c]									
Mean	**268.1**	**280.8**	**274.4**	**270.4**	**281.2**	**279.2**	**253.4**	**271.1**	**275.9**
SD	**28.8**	**42.8**	**32.9**	**32.9**	**49.0**	**34.3**	**28.9**	**29.9**	**30.3**
Skewness	**-.541**	**-.019**	**-.152**	**-.576**	**-.471**	**-.606**	**.165**	**-.469**	**-.358**
Kurtosis	**.824**	**-.940**	**-.437**	**.682**	**-.235**	**-.545**	**-.279**	**-.328**	**-.245**
KR20	**.882**	**.917**	**.940**	**.832**	**.899**	**.870**	**.766**	**.899**	**.964**
SEM	**9.89**	**12.33**	**8.06**	**13.48**	**15.57**	**12.37**	**13.98**	**9.50**	**5.75**
National Norms: Grade 10, Spring									
Mean	265.9	266.5	266.2	267.8	267.7	267.3	266.9	267.2	267.0
SD	34.1	46.2	36.9	38.5	45.1	39.5	40.5	36.5	36.0
KR20	.918	.921	.951	.864	.922	.902	.878	.958	.976
SEM	9.76	12.99	8.17	14.20	12.60	12.37	14.15	7.48	5.58

Note. The reading test consists of the vocabulary and reading comprehension subtests combined. The mathematics test consists of the match concepts and math computation subtests combined. The total battery consists of all items from all subtests combined.

[a]Bold numbers indicate scores of students in the present study at each time point. National norm statistics are from Forsyth et al., 2003, p. 57.

[b]$n = 53$.

[c]$n = 50$.

[d]$n = 61$, except for Revising Written Materials subtest descriptive statistics ($n = 60$).

Table 4
Cohort Group Mean Changes

Cohort	Time	Norm Mean	Norm SD	Above-Level Mean	*z*	Percentile
			Total Battery Score			
2	Baseline	253.7	4.0	264.6	.3206	62
2	Followup 2	267.0	36.0	275.9	.2472	60
3	Baseline	241.1	32.6	245.2	.0337	51
3	Followup 2	259.5	34.5	271.5	.3483	65
			Reading Score			
2	Baseline	252.2	34.7	264.7	.3615	64
2	Followup 2	266.2	36.8	274.4	.2235	59
3	Baseline	242.3	32.8	248.9	.2012	58
3	Followup 2	258.5	35.8	269.2	.2995	62
			Math Score			
2	Baseline	254.3	33.8	255.9	.0473	52
2	Followup 2	273.7	37.8	271.1	.1068	54
3	Baseline	244.1	31.6	238.7	-.1709	43
3	Followup 2	259.8	34.9	269.4	.2747	61

Note. Norm means and SD's are from Hoover et al. (2003, p. 73) and Forsyth et al. (2003, pp. 57-58).

The test score distribution was also closer to the middle range of the test level that students took than had they obtained the same ITBS/ITED scale scores on a grade-level test. When these scores are converted to *z*-scores by dividing by the norm group's mean and standard deviation, the result is a standardized score that shows the number of standard deviations that the scores are from the norm group's mean. Table 4 shows that the *z*-scores obtained from the above-level testings were all between -.25 and +.59, with the average above-level *z*-score being +.35 for reading scores, +.01 for math scores, and +.27 for total battery scores. Assuming that the vertical equating of the ITBS and ITED test levels is of high enough quality that the students would have received the similar scale scores on a grade-level version of the ITBS and ITED, the mean grade-level *z*-

scores of the gifted students would have been +1.04 for reading, +.70 for math, and +.94 for the total battery. Thus, the gifted students' scale scores were less extreme with the above-level testing than they would have been for a grade-level test. If moving the gifted students' score distribution towards the middle of the test level's score range is a goal of above-level testing in order to improve score reliability, then above-level testing would seem to be successful.

Given the concern about regression to the mean expressed in Chapter II, it was investigated separately. The number of students who showed a score decline from one time point to another is displayed in Table 5. The results are somewhat surprising. Almost half of students (40.3%) who were tested at least twice showed at least one decline in reading scores during the course of the study. Math results were similar, with 34.4% of students demonstrating a score decline. Total battery score declines were not as common, with only 25.8% of students showing a total battery score decline during the study. Because overall battery scores are a composite of the subtests, the lower rate of total battery score declines may be due to declines in one subtest (such as reading) being compensated for by gains or maintenance in another subtest (such as spelling or mathematics). In total, 57.2% of the students showed at least one score decline on the reading, mathematics, or total battery during the study. Interestingly, score declines were most common between Followup 1 and Followup 2, during which time students were advanced a grade, and therefore a test level. The change in test level may have some impact on the common score declines observed in this study.

Table 5
Number and Percentage of Students Who Showed a Score Decline

	Test Score		
Time Period	Reading	Math	Total Battery
Baseline to Followup 1	16 (12.7%)	25 (20.0%)	13 (10.4%)
Followup 1 to Followup 2	36 (42.9%)	21 (25.3%)	21 (25.3%)
Baseline to Followup 2	14 (16.9%)	8 (9.8%)	11 (13.4%)
Total	52 (40.3%)	44 (34.4%)	33 (25.8%)

Note. 72 of 128 (57.2%) students who were tested at least twice showed at least one score decline.
Note. *n* varies from 82 to 128.

To investigate the possibility of regression toward the mean, an independent samples *t*-test was conducted for each pair of scores to determine if students who exhibited score declines had higher initial scores than students whose scores did not decline. Of the nine pairs of *t*-tests that were conducted, only the students who showed a decline in reading scores between the first and second followups had higher initial (i.e., from Followup 1) scores than those who didn't show a decline ($d = .58$, $p = .012$). All other mean differences were not statistically significant ($p \geq .360$). Therefore, the presence of score declines for most subtests and tests was not related to initial score.

Research Question 1: Internal Consistency Reliability

The first research question for this study was: What is the internal consistency reliability of the global battery, reading/language arts, and mathematics scores drawn from an above-level administration of an achievement test? The internal consistency reliability coefficients for the ITBS/ITED above-level test administrations and from the test manuals are displayed in Tables 1-3. KR20 values for the norm groups for the

eighth, ninth, and tenth grade levels are also displayed in Tables 1-3. These data are taken from Hoover et al. (2003, p. 73) and Forsyth et al. (2003, pp. 57-58).

Internal consistency reliability coefficients (KR20) ranged from .725 to .931 for sixth grade students, .553 to .963 for seventh grade students, and .766 to .982 for eighth-grade students. Of the 66 coefficients, one (1.51%) was below .700, six (9.09%) were between .700 and .799, 30 (45.45%) were between .800 and .899, and the remaining 29 (43.93%) coefficients were .900 or higher. The distribution of KR20 values had a skewness value of -1.645, which is in accordance with what is known about the distribution of internal consistency coefficients (Feldt, 1965; Rodriguez & Maeda, 2006; Warne, 2011). The total battery scores had the highest reliability (all above .900), which is unsurprising, given the larger number of items in the total battery.

As the tables show, the norm group's KR20 estimates were usually higher than the KR20 values generated by the above-level test scores. In total, only eight above-level KR20 values exceeded the corresponding reliability coefficients for the norm groups—and all eight coefficients were from Cohort 1's scores.

Cronbach's alpha values were used to calculate the standard error of measurement (SEM) values in Tables 1-3. For gifted students in the seventh and eighth grades, SEMs values were very similar to what is seen in the norm sample, which reflects the similar KR20 and standard deviations values of both groups. However, for the gifted sixth graders, the SEM values were much higher than those seen in the eighth grade norm groups. This means that gifted sixth grade students' scores are less precise than other students' scores in the study and is likely the result of the shorter test length for the eighth

grade ITBS test level and the resulting lower KR20 values for the sixth grade gifted students.

Hierarchical Linear Models

HLM results examining the nature of score gains are displayed in Tables 7-9. As stated in Chapter III, nine models are displayed in each table: one random intercept model with no predictors (Model 1); one model with only time as a predictor (Model 2); four random-coefficients regression models with time and one additional level-2 predictor (Models 3-6); a random-coefficients model with time, SES, and ethnicity as predictors (Model 7); a random-coefficients model with time and all four level-2 predictors (Model 8); and a final intercepts- and slopes-as-outcomes model that is equal to Model 8 with an additional interaction (Model 9). The tables also contain Pseudo-R^2 for both level-1 and level-2 variance, the fixed and random parameter estimates for all models, the deviance, and a statistical χ^2 difference test of model improvement (based on the deviance).

As stated in the previous chapter, the independent variables are dummy coded so that Cohort 3 represents the reference group for the cohorts, males are the reference group for the gender variable, overrepresented ethnicities (i.e., White and Asian American students) are the reference group for the ethnicity variable, and middle- and high-SES students (defined as those not participating in a free or reduced lunch program) are the reference group for the SES variable. Time is coded so that the baseline test administration in November 2008 is time 0, the first follow-up (May 2009) is coded as 1,

and the second follow-up (May 2010) is coded as 3. The correlation table of the level-2 independent and dependent variables for the HLM models are displayed in Table 6.

Table 6
Correlation of Dependent Variables and Level-2 Independent Variables

	Total Score	Reading Score	Math Score	Cohort	Gender	Eth_Und	SES
Total Score	1.000						
Reading Score	0.887**	1.000					
Math Score	0.838**	0.629**	1.000				
Cohort	-0.313**	-0.210**	-0.293**	1.000			
Gender	0.038	-0.025	-0.046	-0.014	1.000		
Eth_Und	-0.323**	-0.396**	-0.216*	-0.033	0.063	1.000	
SES	-0.103*	-0.110*	-0.034	-0.031	0.074	0.354**	1.000

[a]Negative correlation indicates that older children have higher scores. [b]Negative correlation indicates that children from underrepresented ethnicities score lower than children from overrepresented ethnicities.
* $p < .05$ ** $p < .01$

Table 7
HLM Analysis Results (Total Battery Score)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Fixed Effect	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Intercept (γ_{00})	264.56	254.26	248.77	252.42	263.15	257.65	263.36	255.06	253.36
	(1.91)***	(2.30)***	(3.61)***	(3.43)***	(2.89)***	(2.97)***	(3.19)***	(4.50)***	(4.64)***
Time (γ_{10})		6.10 (0.66)***	7.35 (0.68)***	6.11 (0.66)***	6.11 (0.66)***	6.07 (0.66)***	6.12 (0.66)***	7.47 (0.67)***	8.47 (0.87)***
Cohort1 (γ_{01})			25.86 (6.64)***					24.36 (6.23)***	24.71 (6.26)***
Cohort2 (γ_{02})			11.97 (4.69)**					12.53 (4.47)***	12.88 (4.47)***
Cohort4 (γ_{03})			-21.15					-20.85	-20.83
			(4.61)***					(4.32)***	(4.33)***
Gender (γ_{04})				3.27 (4.21)				5.06 (3.39)	4.92 (3.41)
Eth_Und (γ_{05})					-21.17		-20.95	-20.25	-16.19
					(3.87)***		(4.13)***	(3.55)***	(4.17)***
SES (γ_{06})						-7.28 (4.15)*	-0.68 (4.17)	-1.48 (3.47)	-1.63 (3.49)
Time x Eth_Und Interaction (γ_{11})									-2.21 (1.26)*
Random Effect	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
Intercept Random effect (u_0)	25.19***	28.66***	25.42***	28.77***	27.35***	28.51***	27.42***	24.23***	24.21***
Time random effect ($\sqrt{\tau_{11}}$)	—	3.59***	3.43***	3.57***	3.57***	3.56***	3.56***	3.46**	3.34**
					Variance Components				
σ^2	280.95	145.12	141.94	145.23	146.40	145.45	146.40	142.82	140.80
τ_{00}	634.48	821.24	645.99	827.63	748.13	812.80	751.97	586.88	586.13
					Effect Sizes (Pseudo-R^2)				
Level-1	—	48.34%	49.48%	48.31%	47.89%	48.23%	47.89%	49.16%	49.88%
Level-2	—	—	21.34%	-0.78%	8.90%	1.03%	8.43%	28.54%	28.63%
					Deviance				
	3994.97	3924.12	3843.19	3916.97	3891.04	3914.53	3888.17	3797.70	3790.62
ΔDeviance (df)	—	70.85[a] (1)***	80.93[b] (3)***	7.15[b] (1)***	33.08[b] (1)***	9.59[b] (1)***	2.87[c] (1)*	126.42[b] (6)***	7.08[d] (1)***

[a]Comparison model is Model 1. [b]Comparison model is Model 2. [c]Comparison model is Model 5. [d]Comparison model is Model 8.
* $p \leq .10$, ** $p \leq .05$, *** $p \leq .01$.

Total Battery Score Results

Table 7 shows the nine HLM models with the total scale score as a dependent variable. Model 1—the random intercept model with no predictors—had a deviance of 3994.97 and an ICC of .693. Unsurprisingly, this was the highest deviance, because the addition of any level-1 or level-2 predictor caused the model to fit the data better than Model 1 does. The addition of the time independent variable reduced level-1 variance (σ^2) by 48.34%. Because time was the only level-1 independent variable, the Pseudo-R^2 remained approximately constant for the following seven models. In addition to the high Pseudo-R^2, the importance of time as a level-1 variable is shown in a change of deviance of 70.85 ($\Delta\chi^2 = 70.85$, $p < .001$), indicating that Model 2 is a major improvement over Model 1. For every six months of time that passed, total ITBS/ITED battery scores increased by an average of 6.10 points (as indicated by the γ_{10} value in Model 2). This increase in scores as time passes is unsurprising, given the longitudinal nature of the study and the presumption that student scores should increase as they spend more time in school. The random effect for time was also statistically significant, with a u_1 value of 3.59 ($p < .001$). This indicates that the rate of total battery score growth varied across students. Because u_1 is interpreted as a standard deviation of slopes (and the square root of τ_{11} when there is only one level-1 variable), an estimate for the range of slopes in the sample can be calculated. This is possible by assuming that the individual students' slopes are normally distributed around the γ_{10} value. Therefore, it is likely that 95% of student slopes are between -0.93 and 13.14 points ($6.10 \pm 1.96 * 3.59$ points) gained

every six months, which shows that students' rate of total score growth varies widely and a small number of students' scores even declined over the course of the study.

Models 3 through 6 each consist of one level-1 predictor (time) and one level-2 predictor: cohort, gender, and ethnicity, and SES respectively. The cohort dummy code variables explained the most level-2 variance: 21.34%. This was expected, because it is assumed that students who have been in school longer (Cohort 1) would obtain higher scores on an academic test like the ITBS/ITED than students who had fewer years in formal schooling. Members of Cohort 1—who had two years more of schooling than members of the baseline group—had above-level scores that were on average 25.86 points higher at baseline than the scores of students in Cohort 3. The γ_{02} value was 11.97, indicating that Cohort 2 students had ITBS/ITED scores that were almost 12 points higher at the baseline measurement time than the scores from Cohort 3. Finally, the γ_{03} value indicates that Cohort 4 members had scores that were on average 21.15 points lower than the scores from Cohort 3 at baseline. However, it is important to note that Cohort 4 was not tested at the baseline because the students were still in the fifth grade at the time. This value is imputed on the basis of the scores obtained from Cohort 4 at the second followup.

Of the demographic variables, ethnicity was the most powerful predictor, with a γ_{05} value of -21.17 (δ = -.71, $p < .001$) and level-2 variance reduced (τ_{00}) reduced by 8.90%. This indicates that students from underrepresented ethnicities (African Americans and Hispanics) had scores that were over twenty points lower than the scores of students from overrepresented ethnicities (Whites and Asian Americans). The other

two independent variables were of little, if any, importance. SES explained 1.03% of level-2 variance according to Model 6 and a γ_{06} value of -7.28 (δ = -.22, p = .080). Gender had a small γ_{04} (3.27 in Model 4, which corresponds to δ = .10; p = .438) and reduced the level-2 total ITBS/ITED score variance by a negligible amount: -0.78%. This latter finding is an anomaly; variance theoretically can never be negative because it is a squared statistic. However, negative variance values are sometimes mathematically possible (such as in Cronbach's α or in commonality analysis) and it is generally accepted that if the values are close to zero, then they should be interpreted as being equal to zero (McBee, 2010; McCoach, 2010a; Thompson, 2003, 2006). This information, combined with the weak statistical significance of gender as a level-2 predictor (p = .438) means that in this student gender plays no detectable role in initial observed above-level test scores.

Models 7 and 8 show theoretically important combinations of covariates and their relative importance in explaining observed total above-level test score variance. Model 7 shows that although SES and ethnicity are statistically significant predictors when entered alone into the HLM models (see Models 5 and 6), but when they are both part of the model, SES explains almost no additional variance and has a very small γ_{06} value of -0.68 (p = .871). In Model 8, it can be seen that both gender and SES are statistically insignificant predictors with large standard errors (p = .137 and 670, respectively).

Finally, Model 9 shows all the level-2 predictors with the only statistically or practically significant cross-level interaction, which is a time x ethnicity interaction. In total, Model 9 reduces level-1 variance by 49.88% and level-2 variance by 28.63%.

A more parsimonious model based on Model 9 can be generated by retaining only those independent variables that are statistically significant. This results in a level-1 equation of

$$Total_{ij} = \beta_{0j} + \beta_{1j}(Time_{ij}) + e_{ij}$$

and level-2 equations of

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(Cohort1_j) + \gamma_{02}(Cohort2_j) + \gamma_{03}(Cohort4_j) + \gamma_{04}(Eth\_Und_j) + u_{oj}$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(Eth\_Und_j) + u_{1j}$$

which combine to form

$$Total_{ij} = \gamma_{00} + \gamma_{10}(Time_{ij}) + \gamma_{01}(Cohort1_j) + \gamma_{02}(Cohort2_j) + \gamma_{03}(Cohort4_j) + \gamma_{04}(Eth\_Und_j) + \gamma_{11}(Eth\_Und_j)(Time_{ij}) + u_{0j} + u_{1j}(Time_{ij}) + e_{ij}$$

as a general equation. This model had a deviance of 3801.46, a level-1 Pseudo-R^2 of 49.86%, and a level-2 Pseudo-R^2 of 29.96%. The estimates and standard errors of each covariate for the parsimonious model are displayed in Table 8.

Interpreting Table 8 is relatively straightforward. The γ_{00} intercept term (254.89) represents the score at Time 0 for a male student in Cohort 3 from an overrepresented ethnicity. Compared to a student in Cohort 3, a student in Cohort 2 would have a score of γ_{02} (14.06) points higher at the first test administration, which would be 268.95 points (254.89 + 14.06 = 268.95). If the student in Cohort 2 were from an underrepresented ethnicity (γ_{05}), then the average score would be 252.37 (254.89 + 14.06 – 16.58 = 252.37).

The fixed effect for time (γ_{10}) is used to calculate the rate of average growth for students in the study. The γ_{10} value of 8.46 indicates that for every six months that passed, students scores increased by an average of 8.46 points. Therefore, a male student in Cohort 3 from an overrepresented ethnicity would be expected to have a score gain of 25.44 during the course of the study, which would lead him to have a score of 280.27 (254.89 + 3*8.46 = 280.27) at the second followup. In comparison, an average student in the norms group would be expected to have a score increase of 13.29 points (from fall of grade 9 to the spring of grade 10) or 15.43 points (from the fall of grade 8 to the spring of grade 9; Forsyth et al., 2003, p. 73; Hoover et al., 2003, p. 57-58).

The interaction between time and ethnicity (γ_{11}) indicates that members of underrepresented ethnic groups have a slower rate of score growth than children in overrepresented groups. Students in underrepresented groups would be expected to have an average score gain of 18.69 points (3*8.46 – 3*2.23 = 18.69) for a final score of 257.00 points (254.89 – 16.58 + 3*8.46 + 3*-2.23 = 257.00) during the 18 months of the study. It is interesting that in spite of the slower rate of score gains that students from underrepresented ethnicities had compared to their White and Asian American peers, they still demonstrated larger score gains than the average student in the norm group would over the course of 18 months.

Figure 3 shows some of the results from the parsimonious model. The figure reflects both the difference in initial starting scores between ethnicity groups (reflected in the different intercepts) and the interaction between time and ethnicity (shown in the different slopes). Norm group scores and growth trends are also shown for comparison purposes.

Table 8
HLM Parsimonious Models

	Dependent Variable		
	Total Battery Score	Reading Score	Math Score
Fixed Effect	Estimate (SE)	Estimate (SE)	Estimate (SE)
Intercept (γ_{00})	254.89 (3.87)***	266.76 (5.13)***	247.03 (3.38)***
Time (γ_{10})	8.46 (0.46)***	4.70 (1.13)***	9.45 (0.79)***
Cohort1 (γ_{01})	25.00 (6.15)***	19.15 (6.10)***	21.98 (6.37)***
Cohort2 (γ_{02})	14.06 (4.43)***	10.63 (4.66)**	9.95 (4.12)**
Cohort4 (γ_{03})	-19.92 (4.36)***	-13.49 (3.60)***	-25.41 (4.17)***
Gender (γ_{04})		-3.71 (4.30)	
Eth_Und (γ_{05})	-16.58 (4.01)***	-26.26 (3.60)***	-9.54 (4.13)**
Time x Eth_Und Interaction (γ_{11})	-2.23 (1.26)*		-2.98 (1.39)**
Time x Gender Interaction (γ_{11})		2.45 (1.40)*	
Random Effect	Estimate	Estimate	Estimate
Intercept random effect (u_0)	23.98***	23.09***	23.16***
Time random effect ($\sqrt{\tau_{11}}$)	3.31**		
	Variance Components		
σ^2	140.87	243.51	209.21
τ_{00}	575.18	533.32	536.40
	Effect Sizes (Pseudo-R^2)		
Level-1	49.86%	25.57%	40.43%
Level-2	29.96%	35.31%	28.57%
	Deviance		
	3801.46	3902.88	3862.33
ΔDeviance (df)[a]	193.51 (7)***	149.97 (7)***	173.65 (6)***
ΔDeviance (df)[b]	10.84 (2)***	3.32 (1)*	9.98 (2)***

[a]Comparison model is Model 1. [b]Comparison model is Model 9.
* $p < .10$, ** $p < .05$, *** $p < .01$

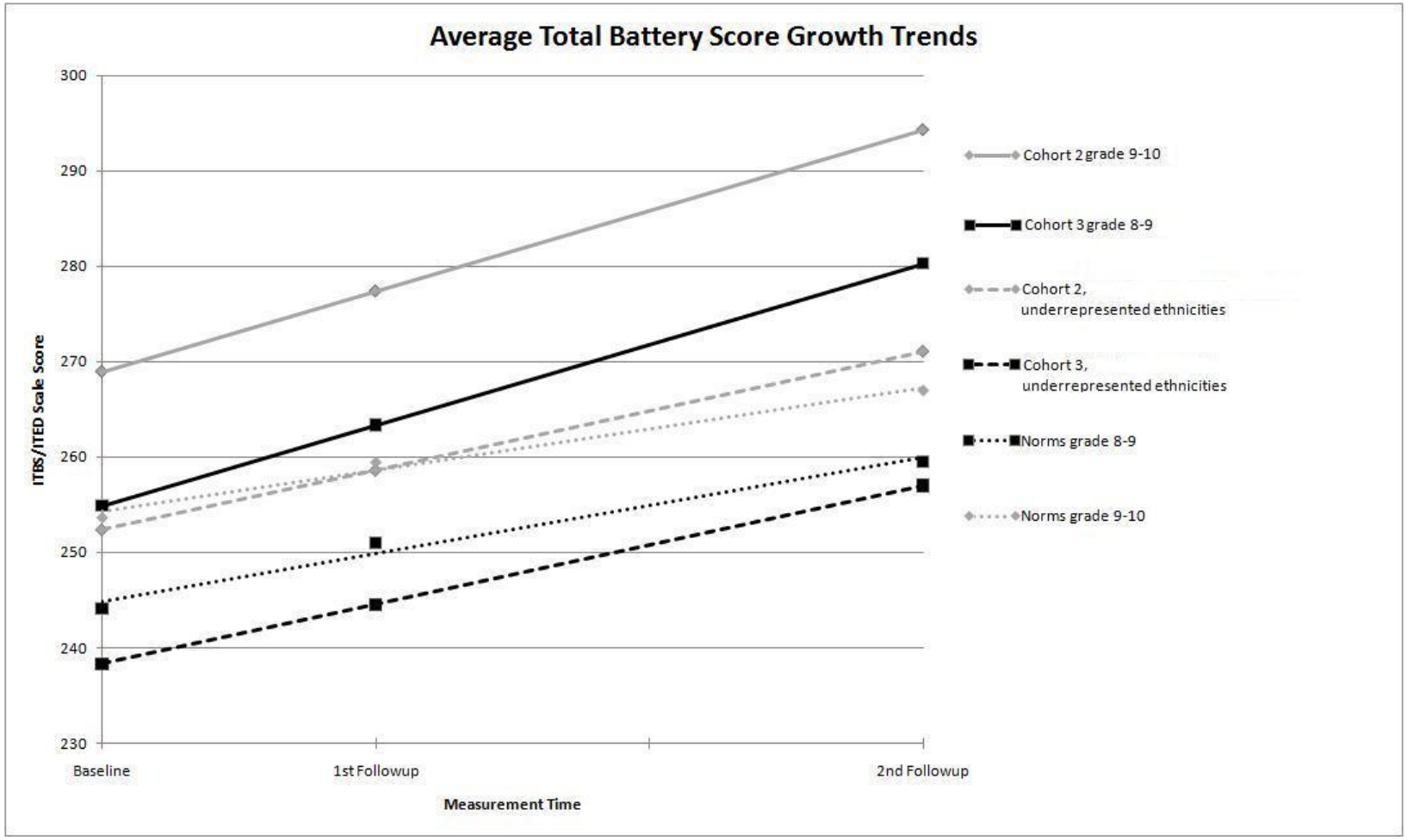

Figure 3 Average total battery score growth trends for above-level cohorts and norm groups.

Table 9
HLM Analysis Results (Reading Score)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Fixed Effect	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Intercept (γ_{00})	267.39	258.89	254.95	259.20	270.40	263.01	270.53	264.53	267.14
	(1.98)***	(2.42)***	(4.26)***	(3.50)***	(2.77)***	(2.94)***	(3.08)***	(4.91)***	(5.33)***
Time (γ_{10})		4.91 (0.71)***	6.07 (0.74)***	4.92 (0.71)***	4.91 (0.70)***	4.87 (0.71)***	4.91 (0.71)***	6.16 (0.73)***	4.71 (1.13)***
Cohort1 (γ_{01})			20.81 (6.19)***					19.45 (6.12)***	19.18 (6.08)***
Cohort2 (γ_{02})			6.67 (6.65)					9.77 (4.71)**	9.82 (4.69)**
Cohort4 (γ_{03})			-15.65					-13.77	-13.84
			(5.05)***					(4.52)***	(4.52)***
Gender (γ_{04})				-0.57 (4.20)				1.04 (3.50)	2.46 (1.40)
Eth_Und (γ_{05})					-26.82		-26.68	-26.00	-25.68
					(3.76)***		(3.98)***	(3.81)***	(3.78)***
SES (γ_{06})						-8.85 (4.15)**	-0.45 (3.98)	-0.99 (3.59)	-0.71 (3.58)
Time x Gender Interaction (γ_{11})									2.46 (1.40)*
Random Effect	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
Intercept Random effect (u_0)	25.85***	28.71***	27.21***	28.76***	25.72***	28.56***	25.79***	24.34***	24.16***
Time random effect ($\sqrt{\tau_{11}}$)	—	1.09	1.09	1.09	1.09	1.09	1.09	1.09	1.12
					Variance Components				
σ^2	325.84	248.78	243.07	248.71	248.39	249.11	248.71	242.14	242.73
τ_{00}	668.13	824.45	740.17	827.18	661.35	815.86	827.18	592.64	583.65
					Effect Sizes (Pseudo-R^2)				
Level-1	—	23.65%	25.40%	23.67%	23.77%	23.55%	23.78%	25.69%	25.51%
Level-2	—	—	10.22%	-0.33%	19.78%	1.04%	19.30%	28.12%	29.21%
					Deviance				
	4052.85	4013.42	3966.65	4010.55	3965.50	4006.03	3959.05	3903.18	3899.56
ΔDeviance (df)	—	39.43[a] (1)***	46.77[b] (3)***	2.87[b] (1)*	47.92[b] (1)***	7.39[b] (1)***	6.45[c] (1)***	109.82[b] (6)***	3.62[d] (1)*

[a]Comparison model is Model 1. [b]Comparison model is Model 2. [c]Comparison model is Model 5. [d]Comparison model is Model 8.
* $p \leq .10$, ** $p \leq .05$, *** $p \leq .01$.

Reading Score Results

HLM models are displayed in Table 9 for the ITBS/ITED reading scores. Model 1's ICC for reading scores was found to be .672 and the deviance was 4052.85. The addition of time as a level-1 predictor produced a statistically significant lower deviance ($\Delta\chi^2 = 39.43$, $p < .001$), indicating that time should be included in the model. This model modification caused a 23.65% reduction in level-1 error variance, which was much lower than the Model 2 Pseudo-R^2 observed for the total battery score dependent variable (48.34%), indicating that time had a lower impact on observed reading scores. This interpretation is reinforced by the γ_{10} value of 4.91 points, which is much lower than the γ_{10} of 6.10 points observed for total ITBS/ITED scores. The random effect for time, however, was small and statistically insignificant (u_1 = 1.09 for Models 2-8; u_1 = 1.12 for Model 9; $p > .500$ for all models). This indicates that the reading scores increased at approximately the same rate for all students in the sample. The γ_{10} value of 4.91 indicates that during the course of the study, the average male student in Cohorts 2 and 3 gained 14.73 points (4.91*3 = 14.73), which is similar to the score gains from the corresponding norm groups: 16.19 points for norm groups from fall of grade 8 to spring of grade 9 and 14.02 points for norm groups from fall of grade 9 to the spring of grade 10 (Hoover, et al., 2003, p. 63; Forsyth et al., 2003; pp. 57-58).

Models 3 through 6 examined the level-2 predictors individually. Model 3 showed that the addition of the cohort variables in the model caused level-2 variance to decrease by 10.22%. Again, this is a much smaller impact than the reduction in level-2 variance that was observed for the total score Model 3 (Pseudo-R^2 = 21.34%). Model 4

showed that gender had no statistically or practically significant impact on above-level reading scores (γ_{04} = -0.57, p = .892, δ = -.02, level-2 Pseudo-R^2 = -0.33%). The strongest level-2 independent variable was that of ethnicity (γ_{05} = -26.82, p < .001, δ = -.89, level-2 Pseudo-R^2 = 19.78%). This indicates that students from underrepresented ethnicities had scores that were on average 26.82 points lower than the scores of students from overrepresented ethnicities. This magnitude of this difference is indicated by the fact that in the norm samples the difference between reading scores of ninth grade students in the fall and twelfth grade students in the spring is 25.75 points (Forsyth et al., 2003, pp. 57, 60). In other words, the score gap between ethnic groups in this study is approximately the same as the gap between students who are just beginning their high school careers and those who are about to graduate. Model 6 examined the impact of student SES on reading scores and found a practically and statistically significant effect (γ_{06} = -8.85, p < .001, δ = -.27, Pseudo-R^2 = 1.04%) that indicated that low-SES students had scores 8.85 points lower than students from middle- or high-SES homes. However, when SES and ethnicity were combined in Model 7, the SES had no unique predictive power above that of ethnicity.

Model 8—which consists of all independent variables—largely confirms the results of the previous models. Individually testing the different level-2 variables' interaction with time produced only one statistically significant interaction, which was with gender. This model is displayed in Table 9 as Model 9. The interaction γ_{11} value is 2.46 (p = .09), indicating that female students had an additional gain of 2.46 points every six months compared to the male students.

A more parsimonious model based on Model 9 can be generated by retaining the statistically significant independent variables. The most parsimonious model for reading scores has a level-1 equation of

$$Reading_{ij} = \beta_{0j} + \beta_{1j}(Time_{ij}) + e_{ij}$$

and level-2 equations of

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(Cohort1_j) + \gamma_{02}(Cohort2_j) + \gamma_{03}(Cohort4_j) + \gamma_{04}(Gender) + \gamma_{05}(Eth\_Und_j) + u_{0j}$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(Gender_j)$$

which combine to form

$$Reading_{ij} = \gamma_{00} + \gamma_{10}(Time_{ij}) + \gamma_{01}(Cohort1_j) + \gamma_{02}(Cohort2_j) + \gamma_{03}(Cohort4_j) + \gamma_{04}(Gender_j) + \gamma_{05}(Eth\_Und_j) + \gamma_{11}(Gender_j)(Time_{ij}) + u_{0j} + e_{ij}$$

as a general equation. It is important to note that even though a gender main effect is statistically insignificant in Models 5, 8, and 9, it is still included in the model because the interaction between time and gender is part of the model, and it is best practice to retain non-statistically significant predictors when they are part of a statistically significant interaction (Thompson, 2006). This model had a deviance of 3902.88, a level-1 Pseudo-R^2 of 25.57%, and a level-2 Pseudo-R^2 of 35.31%. The model estimates and standard errors of each covariate are displayed in Table 9.

One important difference between the results in Tables 7 and 9 is that the HLM reading score models explain much less variance than the models for the total battery score do. All reading models have a level-1 explained variance of less than 26%, while all models for the total battery scores have a level-1 explained variance of at least 47%. The difference in variance explained shows that high above-level reading scores are less influenced by the number of years of schooling (as represented by the cohort variable) and demographic variables. Conversely, above-level reading scores may be more influenced by personal preference and individual psychological variables not included in the HLM models examined in this study.

The important aspects of the parsimonious model for reading scores are shown in Figure 4. In addition to showing the differences between ethnic groups' baseline scores, it also shows gender differences in initial score and growth rates. The figure also includes the norm group scores and growth trends, which shows that all groups' scores and growth rates are comparable or higher than the older norm groups'. Similar to the results from Model 2, the parsimonious model indicated that the average male student from Cohorts 2 and 3 gained 14.1 points in their reading scores during the 18 months of the study. Females gained an average of 21.45 points.

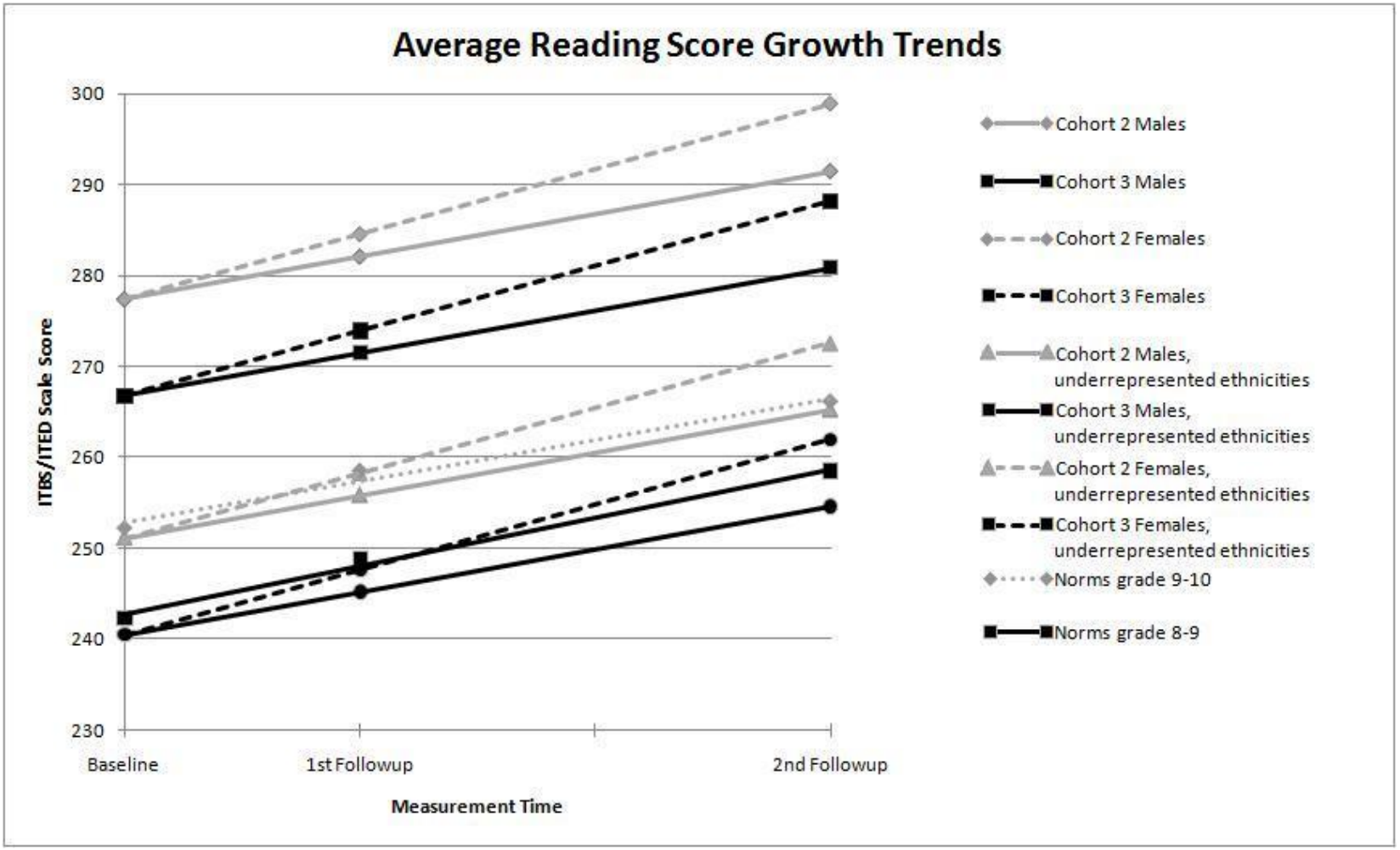

Figure 4 Average reading score growth trends for above-level cohorts and norm groups.

Table 10
HLM Analysis Results (Math Score)

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9
Fixed Effect	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)	Estimate (SE)
Intercept (γ_{00})	257.77 (1.88)***	247.09 (2.31)***	243.99 (2.98)***	248.25 (3.24)***	253.49 (2.93)***	248.27 (2.91)***	252.57 (3.13)***	249.27 (3.90)***	246.95 (3.96)***
Time (γ_{10})		6.30 (0.72)***	8.00 (0.74)***	6.30 (0.72)***	6.38 (0.72)***	6.30 (0.72)***	6.41 (0.72)***	8.12 (0.74)***	9.44 (0.79)***
Cohort1 (γ_{01})			22.24 (6.59)***					21.14 (6.38)***	21.69 (6.38)***
Cohort2 (γ_{02})			7.76 (4.27)*					9.18 (4.22)**	9.60 (4.22)**
Cohort4 (γ_{03})			-26.48 (4.33)***					-25.62 (4.24)***	-25.58 (4.25)***
Gender (γ_{04})				-2.10 (4.11)				-0.65 (3.44)	-0.79 (3.46)
Eth_Und (γ_{05})					-15.61 (3.94)***		-16.50 (4.26)***	-15.75 (3.64)***	-10.25 (4.40)**
SES (γ_{06})						-2.56 (4.11)	2.71 (4.30)	2.37 (3.69)	2.14 (3.69)
Time x Ethnicity Interaction (γ_{11})									-2.92 (1.40)**
Random Effect	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
Intercept Random effect (u_0)	23.89***	27.40***	24.45***	27.45***	27.09***	27.44***	27.19***	24.28***	24.12***
Time random effect ($\sqrt{\tau_{11}}$)	—	3.31	2.71	3.33	3.32	3.30	3.32	2.88	2.53
					Variance Components				
σ^2	351.19	201.18	200.63	201.09	201.66	201.27	201.60	200.49	198.23
τ_{00}	570.97	750.95	597.87	753.60	733.76	753.11	739.19	589.66	581.66
					Effect Sizes (Pseudo-R^2)				
Level-1	—	42.72%	42.87%	42.74%	42.58%	42.69%	42.59%	42.91%	43.56%
Level-2	—	—	30.39%	-0.35%	2.29%	-0.29%	1.57%	21.48%	22.54%
					Deviance				
	4035.98	3974.80	3889.90	3968.03	3953.93	3967.91	3950.65	3860.97	3852.35
ΔDeviance (df)	—	61.18[a] (1)***	84.9[b] (3)***	6.77[b] (1)***	20.87[b] (1)***	6.89[b] (1)***	3.28[c] (1)*	113.83[b] (6)***	8.62[d] (1)***

[a]Comparison model is Model 1. [b]Comparison model is Model 2. [c]Comparison model is Model 5. [d]Comparison model is Model 8.
* $p \leq .10$, ** $p \leq .05$, *** $p \leq .01$.

Math Score Results

The results for the various HLM models that had ITBS/ITED math scores as a dependent variable are shown in Table 10. Model 1, which contained no predictors, had a deviance of 4035.98 and an ICC of .619, which was the lowest ICC of any of the three dependent variables. The addition of time as a level-1 independent variable improved model fit ($\Delta\chi^2$ = 61.18, p < .001, level-1 Pseudo-R^2 = 42.72%). Again, this was unsurprising, given the longitudinal nature of the study. The fixed effect for time (γ_{10}) in Model 2 was 6.30 points (p < .001), indicating that students gained 18.9 points in math scores over the course of the 18 months of the study. The random effect for time (u_1) was notable, but not statistically significant (u_1 = 3.31, p = .232), indicating that there statistically the students' growth in math scores was statistically equal.

The individual impact of the four level-2 independent variables was examined in Models 3 through 6. Model 3 produced similar results for math scores as it did when total ITBS/ITED scores or reading scores were the dependent variable, with a 30.39% reduction in level-2 variance (σ^2). Similarly, Model 4 showed a small and statistically insignificant fixed effect for gender (γ_{04} = -2.10, p = .610, δ = -0.07, level-2 Pseudo-R^2 = -0.35%), indicating that there were no real differences between males and females in above-level mathematics scores at the initial time point. Like the previous two dependent variables, the impact of ethnicity as a level-2 predictor in Model 5 was large compared to the other demographic independent variables (γ_{05} = -15.61, p < .001, δ = -0.51, Pseudo-R^2 = 2.29%). This negative γ_{05} value indicates that students from underrepresented ethnicities obtain lower scores than students from overrepresented

ethnicities. It is important to note, though, that both the γ values and δ effect size both show that the score difference between ethnic groups is not as great for mathematics as it is for reading or the overall battery.

The impact of SES on above-level math scores was examined through Model 6. Unlike the other dependent variables, SES was found to have a small fixed effect value and no statistically significant impact on above-level math scores (γ_{06} = -2.56, p = .533, δ = -0.08) in Model 6. Moreover, the Pseudo-R^2 value was negative (-0.29%), indicating that SES had no impact on above-level math scores in the sample's gifted students. This finding of SES was consistent a consistent aspect of all models that included SES as a predictor (Models 6-9).

Model 8, which included all of the independent variables considered in the previous models, produced results that were consistent with Models 1-7. Afterwards, the interactions between time and the level-2 independent variables were examined. The only interaction that was found to be statistically significant was an interaction between time and ethnicity (γ_{11} = -2.92, p < .001). Including this interaction led to a statistically improved model ($\Delta\chi^2$ = 8.62, p = .033) and means that students from underrepresented ethnicities had score increases that were 2.92 points lower every six months than the overrepresented students' 6.30 (γ_{10}) point gains.

Based on these findings, the most parsimonious model for the above-level math reading scores consists of a level-1 equation of

$$Math_{ij} = \beta_{0j} + \beta_{1j}(Time_{ij}) + e_{ij}$$

and level-2 equations of

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(Cohort1_j) + \gamma_{02}(Cohort2_j) + \gamma_{03}(Cohort4_j) + \gamma_{04}(Eth\_Und_j) + u_{oj}$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(Eth\_Und_j)$$

which combine to form

$$Math_{ij} = \gamma_{00} + \gamma_{10}(Time_{ij}) + \gamma_{01}(Cohort1_j) + \gamma_{02}(Cohort2_j) + \gamma_{03}(Cohort4_j) + \gamma_{04}(Eth\_Und_j) + \gamma_{11}(Eth\_Und_j)(Time_{ij}) + u_{0j} + e_{ij}$$

as a general equation. This model had a deviance of 3862.33, a level-1 Pseudo-R^2 of 40.43%, and a level-2 Pseudo-R^2 of 28.57%. The model estimates and standard errors of each covariate are displayed in Table 10. Figure 5 shows some of the results from the parsimonious model. The figure reflects both the difference in initial starting scores between ethnicity groups (reflected in the different intercepts) and the interaction between time and ethnicity (shown in the different slopes). Norm group scores and growth trends are also shown for comparison purposes.

The rate of scores gains in the parsimonious models reveals that gifted students in the study made greater improvements in math than the corresponding norm groups. The average student from an overrepresented ethnicity gained 28.35 points (9.45*3 = 28.35) between the baseline testing and second followup. Due to the interaction effect in the parsimonious model, students from underrepresented ethnicities had an average gain of 19.41 points. Nevertheless, the norm groups were expected to gain 15.72 points (between fall of grade 8 and spring of grade 9) and 12.89 points (between fall of grade 9

and spring of grade 10)—a score gain that is noticeably smaller than what was observed in this study.

Research Question 2: Rate of Score Gains

The second research question for this study is: Do gifted children make larger achievement gains in overall, reading/language arts, and mathematics scores than average students in a more advanced grade? This is calculated by multiplying the slope of the HLM equations by 3 and comparing the result to the difference in the norms' means for measurements 18 months apart (see Forsyth et al., 2003, pp. 57-58; Hoover et al., 2003, p. 73). For gifted students, the total battery average score gain was 25.41 points for students from overrepresented ethnicities and 18.78 points for students from underrepresented ethnicities. The norm groups, in comparison, gained 15.38 points between the fall of grade 8 and the spring of grade 9 (which corresponds to Cohort 3) and 13.29 points between the fall of grade 9 and the spring of grade 10 (which corresponds to Cohort 2).

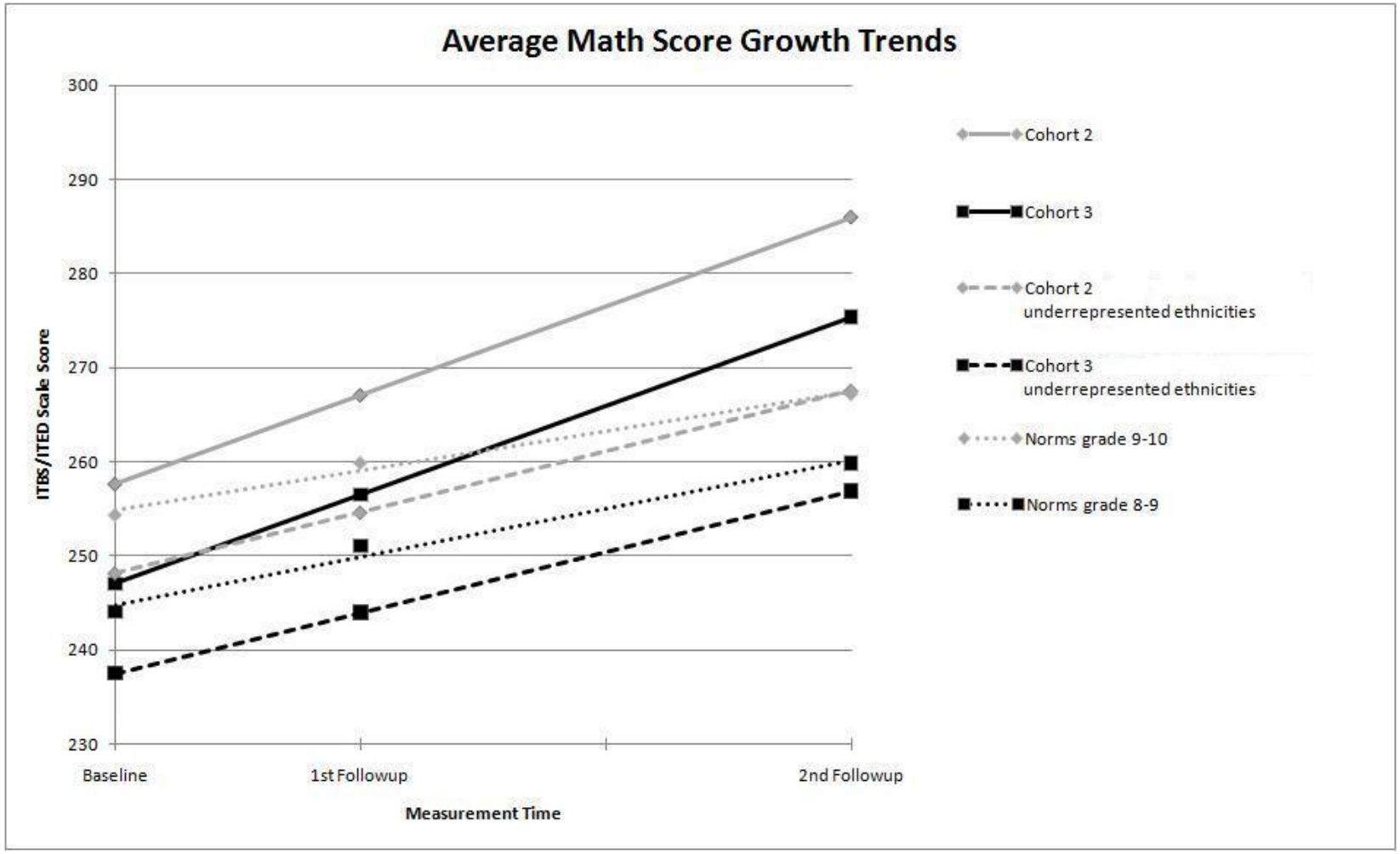

Figure 5 Average math score growth trends for above-level cohorts and norm groups.

Table 4 shows some cohort differences in the growth rate. Comparing the *z*-scores between from the cohorts at baseline and the second followup shows that Cohort 2 had mean changes that were much smaller than the changes for Cohort 3. Overall, Cohort 2's *z*-scores decreased between the baseline and the second followup in all three score areas: total battery, reading, and mathematics. Cohort 3, on the other hand, showed gains in all three areas.

Table 4 is illuminating in that it shows cohort differences in growth. However, it should be remembered that both cohorts had a number of students enter and leave the study between the two time points. Therefore, the results in Table 9 may not reflect actual growth, but rather shifts in cohort membership over time. The HLM time parameter estimates are thus more interpretable as real measures of growth. The differences between the norm groups' scores and the HLM growth results are displayed in Figures 3-5.

Figures 3-5 show a few illuminating aspects of the results. First, even the groups with the lowest mean scores (e.g., underrepresented ethnicities and—in reading—males) had higher rates of growth than the norm groups who were two grades more advanced in their education. Second, despite the fact that these groups underperformed compared to the baseline groups, they still obtained scores that were competitive with the older norm groups (although the means scores did not always surpass those of the older norm groups.

Research Question 3: Demographic Variable Impact

The third research question for this study was: Do demographic variables (gender, ethnicity, and SES) influence the initial scores or rate of overall, reading/language arts, and mathematics score growth of gifted students? The answer to this question, based on the parsimonious models in Table 8, is similar for total scores and mathematics scores, but not for reading scores.

For gifted students taking the ITBS/ITED above-level, ethnicity had a statistically and practically significant impact on their observed scores. For total scores, students from underrepresented ethnicities had a score that was 16.58 points lower than other students' scores at the baseline testing. Moreover, this score gap increased by an average of 2.23 points every six months. For mathematics scores, the initial gap was smaller, with students from underrepresented ethnicities scoring 9.54 points lower at baseline. Yet, the gaps in mathematics also continued to grow at a rate of 2.98 points every six months as students advanced through their schooling.

Ethnicity also had an association with initial score gaps for reading scores, with students from underrepresented ethnicities scoring 26.26 points lower than their classmates from overrepresented ethnicities. However, there was no time x ethnicity interaction for reading scores, indicating that this score gap did not change throughout the course of the study. Reading score results were also influenced by gender—an outcome not observed in mathematics and total battery scores. Although there were no initial differences in male and female students' scores at the baseline assessment, the

female students' scores increased by 2.45 points more every six months than did the male students' scores.

SES had a statistically significant relationship between observed scores for reading and total battery scores, as indicated in Model 6 for Tables 7 and 8. However, Model 7 in the same tables shows that after ethnicity is taken into account, SES has no additional influence on observed above-level test scores.

For the reading scores, the male students in this study gained 14.13 points, while female students gained 21.51 points. The corresponding norm groups gained only 16.19 points between the fall of grade 8 and the spring of grade 9 and 14.02 points between the fall of grade 9 and the spring of grade 10.

Students from overrepresented ethnicities showed a gain of 28.32 points in math scores and students from underrepresented ethnicities showed an 18-month gain of 19.56 points. The corresponding norm groups gained only 15.72 points between the fall of grade 8 and the spring of grade 9 and 12.89 points between the fall of grade 9 and the spring of grade 10.

Research Question 4: Intercept-Slope Correlations

The fourth research question for this study was: What is the relationship between initial above-level overall, reading/language arts, and mathematics scores and rate of score growth? This question was answered through examination of the G matrix produced by the final parsimonious HLM models.

Because it would be impractical to report the matrices of all models examined in this dissertation, they will not be reported in full. However, Table 11 displays the slope-

intercept correlations for all models. To produce the unstandardized G matrix, the interested reader may obtain the τ_{00} from Tables 7-9. The τ_{11} value can be derived from information in Tables 7-9 by squaring the u_1 value. To produce the covariance between the slope and intercept, the correlation reported in Table 11 should be multiplied by the corresponding $\sqrt{\tau_{00}}$ and u_1 values.

Table 11
Slope-Intercept Correlations (Standardized τ_{01} Values) for HLM Models (n = 84)

	ITBS/ITED Test		
Model	Total Battery	Reading	Math
1	—	—	—
2	-.006	-.305**	.040
3	-.116	-.476***	-.184
4	-.019	-.288**	.038
5	-.128	-.431***	-.104
6	-.008	-.379***	.043
7	-.127	-.433***	-.112
8	-.287**	-.616***	-.362***
9	-.252*	-.613***	-.305**
Parsimonious Model	-.211	—[a]	—[a]

[a]Model does not produce a correlation, because the model does not include a u_1 term.
* $p < .05$, ** $p < .01$, *** $p < .001$

For total battery above-level scores, Models 2-7 show no statistically significant relationship between initial student score and slope. However, Models 8 and 9 show a statistically significant negative relationship between the intercept and slopes (r = -.287, -.252, respectively). The parsimonious model for above-level total scores produces an intercept-slope r = -.211 (p = .057). This negative relationship indicates that students

with lower initial scores had greater score gains than students with higher initial scores. This relationship is show in Figure 6, which displays the total battery score gains and decreases over time for the 221 students who were tested at least twice during this study. The negative relationship between initial score and slope is contrary to prevailing theory that high achieving (or high ability) students learn faster and make greater academic gains than lower achieving peers (e.g., Carroll, 1993; Dai, 2010; Eisner, 2002; Gagné, 2005; Ruf, 2005). These results may indicate a persistence of regression toward the mean in the above-level test scores.

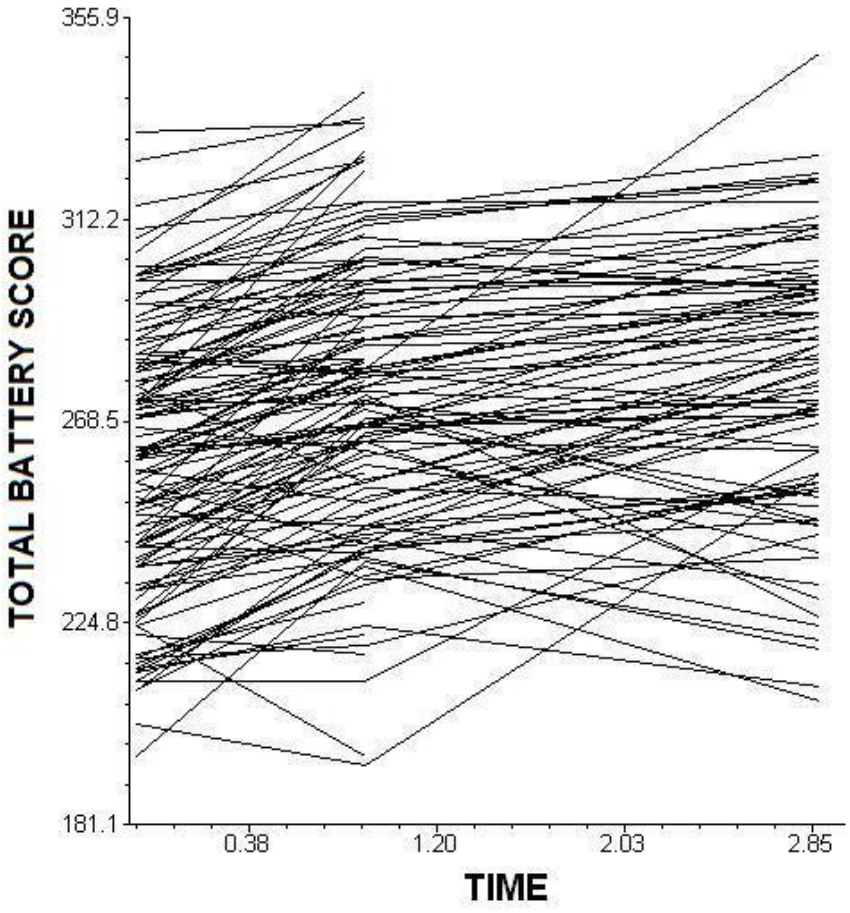

Figure 6 Total battery score changes over time (*n* = 221).

HLM models for above-level reading subtest scores also indicated a negative relationship between intercepts and slopes, although these inverse relationships were much stronger than those observed for the total battery scores. Indeed, the relationship between intercept and slope for the reading score Model 9 is r = -.613. Despite the strong relationship between initial scores and growth rate, the covariance between initial score and growth rate is not included in the parsimonious model because the students' growth rates (u_1) were not found to be statistically different and were therefore constrained to be equal.

The slope-intercept relationship for above-level math scores was not statistically different from zero for Models 2-7. However, for Models 8 and 9, the correlation between slope and intercept is r = -.362 and r = -.305, respectively. Again, this demonstrates an inverse relationship between initial student score and the rate of score gains. The above-level math scores also did not have an intercept-slope correlation for the parsimonious models, because the model lacked a u_1 term, which means that the model constrains all individual slopes to be equal.

Research Question 5: Effect Sizes

The fifth research question was: What percentage of overall, reading/language arts, and mathematics score variance is explainable through time, demographic variables, and cohort membership? As mentioned in Chapter III, the multilevel nature of HLM does not permit a true effect size to be calculated. Instead, a Pseudo-R^2 was calculated to represent the percentage of reduction in σ^2 (for level-1) or τ_{00} (for level-2) compared

to Model 1, which has no predictors in either level. Instead of a "variance accounted for," it is more properly understood as the percentage of reduction in the error variance.

Table 8 shows the Pseudo-R^2 values for the parsimonious models. For level-1 variance, time had a larger impact on total battery scores (49.86%) and mathematics scores (40.43%) than it did on reading scores (25.57%). However, demographic variables had a similarly strong magnitude of impact on the reduction in level-2 variance. The level-2 Pseudo-R^2 was for 29.96%, 35.31%, and 28.57% for total, reading, and mathematics scores, respectively.

CHAPTER V

DISCUSSION AND CONCLUSION

As described in Chapter I, the research questions for this study were:

1. What is the internal consistency reliability of the global battery, reading/language arts, and mathematics scores drawn from an above-level administration of an achievement test?
2. Do gifted children make larger achievement gains in overall, reading/language arts, and mathematics scores than average students in a more advanced grade?
3. Do demographic variables (gender, ethnicity, and SES) influence the initial scores or rate of overall, reading/language arts, and mathematics score growth of gifted students?
4. What is the relationship between initial above-level overall, reading/language arts, and mathematics scores and rate of score gains?
5. What percentage of overall, reading/language arts, and mathematics score variance is explainable through time, demographic variables, and cohort membership?

This chapter will examine each question, which will then be followed by a general discussion of findings.

Research Question 1: Internal Consistency Reliability

Tables 1-3 in Chapter IV display the internal consistency reliability coefficients for the above-level test scores and comparison coefficients for the test level norms. Of the 66 above-level KR20 values are reported in the tables, only eight (12.1%) were

higher than the corresponding KR20 value for the test level's norm group scores. Further, the vast majority of the KR20 coefficients, were in acceptable ranges for basic research (Cortina, 2002), although most (87.9%) were lower than their counterparts in the norm groups. This is probably due to unique sample characteristics, including greater homogeneity than is likely observed in the norm groups.

Research Question 2: Rate of Score Gains

Gifted students taking above-level tests demonstrated higher score gains than what would be expected for the average student who would normally take those test levels. The HLM models indicated that on average, the gifted students in the study made statistically significantly greater score gains across the 18 months of the study than the typical student in the norm group. For total battery scores, the average gifted student from an overrepresented ethnicity in this study gained 25.38 points, while the average student who was two years' more advanced in school would gain only 13.29 (from the fall of grade 9 to the spring of grade 10) or 15.43 points (from the fall of grade 8 to the spring of grade 9). Even the students who would be expected to gain fewer standard score points due to the presence of interaction effects (e.g., male students in reading and students from underrepresented ethnicities for total battery and mathematics scores) made greater mathematics and total score gains than the average student in the norm groups. This finding coincides with decades of previous findings on the rate of gifted students' learning and progress through the academic curriculum (e.g., Corno et al., 2002; Gottfredson, 1997b; Gross, 2004; Stanley & Benbow, 1981-1982; Terman, 1926; van Wagenen, 1925). Reading score gains were not as pronounced in the gifted group as

in the norm group. The gifted males in the study gained an average of 14.1 points, which is approximately equal to the 16.19 and 14.02 point gains for the two corresponding norm groups. Gifted females tested above-level, however, gained an average of 21.45 points, due to the presence of the time x gender interaction effect.

Research Question 3: Demographic Variable Impact

Three demographic variables were examined to determine their influence on both initial scores and rate of growth. Results indicated that ethnicity was the strongest predictor of initial above-level test scores—about three times more powerful than SES when predicting total battery and reading scores and six times more powerful when predicting above-level math scores. Further, when SES and ethnicity were combined into the same HLM equations, the explanatory power of SES almost completely vanished. To say that these results are disappointing would be an understatement because it implies that the observed differences in above-level scores are more due to cultural and/or developmental differences and not economic differences. The relative strength of the impact of SES and ethnicity on intellectual ability or academic achievement is subject to much debate in the literature. Some previous researchers find SES to be a more powerful determinant of group differences than ethnicity (e.g., Carman & Taylor, 2010). On the other hand, other researchers (e.g., Konstantopoulous et al., 2001) find the opposite to be true—as I do in this study. The issue is further clouded by the fact that low academic ability or intelligence often acts as a cause for many poor life outcomes—including poverty (Gottfredson, 1997a, 1997b, 1998)—and that poverty and

a poor environment can depress intellectual and academic development (Brooks-Gunn & Duncan, 1997; Gottfredson, 1997a).

Gender was not a statistically significant predictor of initial above-level test scores for total battery, reading, or mathematics ITBS/ITED tests. However, gender and time interacted to produce different growth rates in reading for males and females. For every six months that passed, females gained 2.92 points more than males on above-level reading tests. This interaction effect for gender is not completely unexpected. Gifted girls find reading to be more interesting than gifted boys do (Olszewski-Kubilius & Turner, 2002), a tendency that also manifests itself among the general school populations (Francis, 2000). This interest in reading could easily translate into a higher above-level test score, whether the test is measuring aptitude or achievement (Corno et al., 2002).

Similar interactions were found between ethnicity and time for the total battery and reading scores, indicating that ethnicity was a moderator variable for those outcomes. Students from overrepresented ethnicities gained an additional 2.21 for total ITBS/ITED scores and 2.92 points for reading scores per half year, respectively.

Research Question 4: Intercept-Slope Correlations

Table 11 shows the correlations between initial score and the rate of score gains for all models considered in this study. For the above-level reading scores, the correlation was negative in all models. However, for total battery and mathematics scores, the correlation was close to 0 for Models 1-7, but then became negative in Models 8 and 9. (In Model 9, r = -.252 for total battery scores, -.613 for reading scores,

and -.305 for mathematics scores.) Therefore, the students with the lowest scores made the greatest score gains, which does not coincide with previous theory (e.g., Carroll, 1993; Dai, 2010; Eisner, 2002; Gagné, 2005; Gottfredson, 1997b).

The presence of regression toward the mean may account for some of the negative correlation, but likely not all of it. Additional factors that may contribute to this finding include (a) a ceiling effect may still be present for the highest scoring students, (b) the gifted program is not serving the needs of the brightest students but is serving the needs of the moderately gifted, (c) the above-level test scores do not demonstrate sufficient test-retest reliability to track score gains, or (d) the above-level test scores do not perform as expected and the negative score-gain correlation is a manifestation of a unique psychometric phenomenon. The data at hand cannot reveal which factor or combination of factors the cause of the negative correlation between initial score and rate of score change.

Research Question 5: Effect Sizes

The level-2 Pseudo-R^2 values in Tables 4-6 show the amount of reduced level-2 variance of above-level scores from demographic variables. For all the total battery and mathematics scores outcome variables, cohort membership explained the most level-2 variance: 21.34% and 30.39%, respectively. However, for the above-level reading scores, ethnicity was the most powerful predictor (19.78%).

Normally, one would expect that the cohort variables would be the most powerful demographic variable for all above-level outcomes in this study because cohorts differed in the number of years of schooling that children have received.

However, by the middle school years, a high reading level is largely a result of early success with reading, self-directed practice, and personal choices to engage in reading (Petersen, Kolen, & Hoover, 1989; Stanovich, 1986), especially among gifted students (J. R. Mills & Jackson, 1990). Thus an additional year or two of schooling may translate into a much weaker score advantage (and therefore, a smaller Pseudo-R^2) in reading than it would in mathematics—where self-instruction is much more difficult. Instead, the most independent variable that produced the largest level-2 Pseudo-R^2 was ethnicity, with 19.78%.

This study also shows a large advantage that students from overrepresented ethnicities (Whites and Asian Americans) have over students from ethnicities that are underrepresented in the gifted magnet program (Hispanics and African Americans), both in terms of initial scores and in the greater score gains that overrepresented students demonstrate in total scores and mathematics subtest scores. The presence of score gaps between ethnicities on educational achievement or intellectual ability tests is widespread (e.g., Forsyth et al., 2003; Gottfredson, 1997a, 2000; Herrnstein & Murray, 1996; Hoover et al., 2003; J. Lee, 2002), including in gifted education research (e.g., S.-Y. Lee & Olszewski-Kubilius, 2006; McBee, 2006, 2010; Olszewski-Kubilius & S.-Y. Lee, 2011; Yoon & Gentry, 2009). This study merely joins the large body of research showing a substantial score difference between ethnic groups. The existence of these score gaps is not controversial, but the cause(s) of such gaps are (Kaplan & Saccuzzo, 2005). Explanations range from genetic or biological factors (Plomin & Petril, 1997; Rowe, 1997) to mostly environmental causes (Scarr & Weinberg, 1976). Other

researchers claim that the true causes of score gaps are unknown or ascribe score gaps to the product of a vague genetic-biological-environment interaction (e.g., Gottfredson, 1997a; Herrnstein & Murray, 1996; J. Lee, 2002; Neisser et al., 1996). The data from this study do not shed any light on the causes of these score gaps. Indeed, the score gaps in this study may be merely due to local influences—such as a differential selection in admission to the program or in the types of families in each group who choose to send their child to the program.

Gender had little effect on explained level-2 variance. For all models in which it was the only level-2 predictor (Model 4 in Tables 4-6), gender had a negative Pseudo-R^2. Thus, as a main effect, gender holds no predictive power. This is unsurprising because the educational and achievement gap that formerly existed between boys and girls has effectively closed (Francis, 2000), and in some areas females have surpassed males in educational achievement (Deary, Strand, Smith, & Fernandes, 2007). Moreover, in most studies of intellectual ability, males and females have equal group means or any differences are very small (Gottfredson, 2003; Olszewski-Kubilus & S.-Y. Lee, 2011). The negligible effects of gender in this study coincide with such previous findings. However, (as discussed in Chapter IV) gender provided a statistically significant interaction with time to produce differential levels of growth in reading scores, with females gaining 2.92 more points than males every six months. Therefore, the initial equality between genders in reading may change over time, with females having higher observed scores than males.

General Discussion

In general, the ITBS and ITED above-level test scores "behave" in much the same way as the SAT, ACT, and other previously researched above-level test scores do. As was expected, the test ceiling was higher in the above-level testing condition than it would be on a grade-level test. This, in turn, led to the observed scores being more variable than would be expected on a grade-level test, which also improves discrimination and evaluation of each individual gifted child. In addition to the test ceiling being raised, the tests scores usually demonstrated high levels of reliability. The generally high reliability of above-level test scores is new empirical evidence that supports one of the most frequently cited reasons for conducting above-level testing.

Because of theoretical claims of above-level testing's capability of reducing regression toward the mean, the pattern, magnitude, and causes of score declines were investigated. In this study, a majority (57.2%) of the students who were tested at least twice showed a score decline in reading, math, or the total battery. Thus, score declines are surprisingly common—even when gifted students are tested above-level. However, in eight of nine comparisons, the score declines that occurred between two testings did not have a statistically significant relationship with the score at the first testing. Therefore, it is not possible to state the exact cause of score declines in this study. However, given the importance of regression toward the mean in the identification of giftedness (Lohman & Korb, 2006), these results show that relationship between regression toward the mean and above-level testing may be a fruitful area of investigation.

Implications

This study has important psychometric implications. Much of the descriptive data from this study is consistent with previous reports of above-level test data, including the existence of large ethnicity group differences, the lack of statistically significant gender differences, and the approximate normality of the above-level test score distributions. The fact that much of this information on the ITBS and ITED corresponds to previous Talent Search findings on the SAT (Barnett & Gilheany, 1996; Keating & Stanley, 1972; Olszewski-Kubilius & S.-Y. Lee, 2011), ACT (S.-Y. Lee & Olszewski-Kubilius, 2006; Olszewski-Kubilius & S.-Y. Lee, 2011), SSAT (Lupkowski-Shoplik & Assouline, 1993), and EXPLORE (Colangelo et al., 1994; Olszewski-Kubilius & S.-Y. Lee, 2011) is encouraging and suggests that above-level achievement tests perform in similar ways, even when they are administered to a group that is not as selective as those who usually apply for Talent Search programs.

Some findings are a unique contribution to the above-level testing literature, such as the level of KR20 reliability coefficients and the degree of regression toward the mean of above-level test scores. This study provides the first psychometric evaluation of these issues and can lay the foundation for further examinations of the psychometric properties of above-level test scores.

Apart from psychometric issues, this study also has practical implications. First, the study provides a possible outline for evaluating a gifted program. Program evaluation has historically been a weak area of gifted education research and practice (Borland, 2003; Gallagher, 2006; VanTassel-Baska, 2006), partially due to a lack of

instruments that effectively measure high levels of educational progress. This study is the first to show that above-level testing can be used to track individual progress and overcome the problems of using traditional measures of academic achievement in gifted program evaluation.

Another practical implication of this study is that it can provide some guidance to district personnel who wish to implement above-level testing. The gifted education literature provides almost no guidance on when above-level testing should be implemented outside of a Talent Search or grade skipping context. For most practitioners, this lack of guidance may be an impediment to using above-level testing in their gifted programs. This study's example of a specific test, age-grade discrepancy, and results can give practitioners a starting point for making plans to implement above-level testing in their districts for identification, evaluation, and educational planning. This study also provides guidance on how to use above-level test scores, which few districts currently do, even when scores are available (Swiatek & Lupkowski-Shoplik, 2005).

Limitations

Internal validity. There are several threats to internal validity that arise from the fact that all sample members were enrolled at a single gifted magnet program in a single district. These will be interpreted in the threat to internal validity framework provided by Cook and Campbell's (1979). First, this study may be threatened by history effects in which different events that happened in the school or district impacted the cohorts differently. For example, as the program became better established, it is possible

that district personnel learned better which types of students were best suited for it. This would lead to Cohort 4 and Cohort 2 (which had a large number of student admitted to the program after the first followup) to have a different composition compared to the other cohorts. Similarly, the students in Cohort 1 could have been exceptionally bright and motivated compared to other gifted students in the district because many of them were willing to change schools during their final year of middle school and leave their home campuses and friends.

Differential selection could also be a threat to internal validity. This may partially explain the score differences between overrepresented and underrepresented ethnicities in the sample. There is strong pressure from the school district's state office of education to have the composition of gifted programs reflect the ethnic makeup of the district as a whole, which may cause district personnel to admit children from underrepresented groups who are not as academically advanced as other students. (See Lewis, DeCamp-Fritson, Ramage, McFarland, & Archwamety, 2007, for an example of the changes to program admissions criteria that could lead to different standards for underrepresented groups. See Ford, 2003, and Warne, 2009, for suggestions to increase the diversity of a gifted program without lowering admissions standards.)

The age of the students could have contributed to history and differential selection issues associated with this study. Because the students were all at least aged 11 years or more when they entered the program, they had several years of educational history which could have created (or magnified) group differences that were discovered in this study. For example, the district policies on gifted identification have changed

multiple times during the students' educational careers, and children labeled as gifted under current policy may vary greatly from those children who were identified several years previously. I requested data on how each child was identified and labeled as gifted, but the school district personnel said that the data were not available. In the future, I hope to conduct this same study with younger children (perhaps as young as the first grade) who were all identified under a single policy in order to lessen these history and differential selection effects.

Experimental mortality could also be a threat to the internal validity of this study. However, experiment mortality likely had a small impact on the internal validity of the study. Figure 2 shows the reasons why students left the gifted program. As can be clearly seen, most students who left did so for reasons that likely had nothing to do with their academic ability (i.e., moving out of the district, death, absent on test day, test form lost). Nine students left the program during the course of the study. Two of these left the study because they skipped a grade, and the other seven left for because they were struggling with the gifted curriculum or for social reasons. However, these were a minority of study dropouts, and I do not believe that these seven students had a large impact on the analysis of above-level test scores.

There is also the problem of a small sample size for some research questions. For example, conclusions about the rate of score gains were based on data from the 84 students who were measured at all three time points. Therefore, it is necessary to replicate this study with much larger sample sizes in the future.

External validity. The possible threats to external validity are no less severe than the threats to internal validity. Again, many of these threats arise from the presence of a single convenience sample in this study. The threats to validity discussed here are also drawn from the Cook and Campbell (1979) framework.

The convenience sample in this study severely limits the extent to which these findings can be generalized to other gifted student populations or programs. Even if one limits the target population to gifted middle school students in this specific district, the results may not be completely generalizable because there are two gifted magnet programs in the district and a majority of gifted students attend neither program. However, the fact that some of the above-level testing results coincide with many previous studies using more selective Talent Search samples and different instruments (e.g., SAT, EXPLORE, ACT) is encouraging and may provide a logical basis for some tentative generalization to other samples.

On the other hand, the fact that this study was conducted over the course of two school years in a typical gifted middle school magnet program may make the results more applicable to the real world. Practitioners who encounter this study may likely recognize aspects of this study that are found in many gifted programs throughout the country. Perhaps seeing above-level testing applied to a real school situation (instead of a Talent Search environment) could prompt practitioners to consider the practice for the gifted students in their districts.

Statistically, the final parsimonious models in the study were mostly exploratory in nature. Like all exploratory statistical procedures, there exists the possibility that the

results here are overfitted to this specific sample and capitalize on chance. I recommend that the parsimonious models from this study be used again with a separate sample in order to judge their applicability to other groups.

Other study limitations. Another limitation of this study is that the gifted children at this magnet program do not seem to be as elite as is often seen in gifted education research. The children's average scaled scores when expressed as grade-level percentile are quite low for gifted students: between the 80th and 92nd percentile for reading, the 58th and 84th percentile for math, and the 77th and 87th percentile for overall scale scores. Although these percentiles may be depressed through the equipercentile equating procedure used in the development of the ITBS and ITED (Forsyth et al., 2003; Holland & Dorans, 2006; Hoover et al., 2003; Kolen, 1981), they still indicate that this sample is not as selective as what appears in most above-level testing studies.

Also, some subgroups of interest (i.e., African Americans and Hispanics) were combined because individually the groups were too small for the statistical tests to have much power. It is possible that Hispanics and African Americans have different above-level score profiles (e.g., growth rates, score gains, and distributions). Similarly, the small number of Asian American students (just 3 out of the 225) prevents in-depth analysis about a substantively interesting overrepresented group. Other variables that could be potentially interesting, such as whether a student was bilingual, could also not be included in the analyses, because there were not enough students for powerful statistical tests to be conducted.

Another limitation of this study is that it does not address whether the above-level test scores or reliability coefficients obtained here were higher than the scores or coefficients that would be obtained from a grade-level test. In order to make such a judgment, the same sample of gifted students would need to take a grade-level and an above-level test—which did not happen in this study. A study in which counterbalanced grade-level and above-level test forms were administered to a sample would provide this information. The current study is also limited by examining just one type of reliability—internal consistency reliability. Although internal consistency reliability is the most commonly examined reliability type (Hogan, Benjamin, & Brezinski, 2000; Thompson, 2003), there are other sources of measurement error that could be investigated with other testing research designs. Other measurements of reliability, such as the conditional standard error of measurement, may also prove to be valuable to investigate.

The findings on above-level score declines are also problematic the focus of this study was score growth. However, score declines happened nonetheless, despite the passage of 6-12 months between testings, which is enough time for real learning to occur that could mask any regression toward the mean. A study in which the testing intervals are much shorter—perhaps two weeks or less—would be more informative. Nevertheless, it is enlightening that over half of participants who were tested at least twice showed at least one score decline, which indicates that above-level testing may not solve the problem of regression toward the mean among gifted students.

Finally, this study may be limited by the linking and equating procedures that the ITBS and ITED creators used to create a scale score that permits comparisons across test

levels. All equating procedures introduce some amount of error into score comparisons across different tests, test forms, or levels (Holland & Dorans, 2006; Skaggs & Lissitz, 1986). In general, the equipercentile equating method has performed well in many different contexts (e.g., Kolen, 1981; O'Brien & Tohn, 1984; Yin, Brennan & Kolen, 2004), including in vertical equating across ITBS levels when the examinees have high levels of ability for their age (Harris & Hoover, 1987). However, the impact of equipercentile grading on above-level test score interpretation is unknown.

Further Research

Above-level testing has the potential to be a fruitful avenue of research, mostly because psychometric research on the practice has been directed almost exclusively towards basic issues. One advanced psychometric issue would be to investigate how above-level testing scores are impacted by different scaling methods because cross-level score comparisons are very sensitive to the scaling method used to align test levels (Kolen, 2006). A study like this would help disentangle the effects of actual student achievement and artifacts from the test construction and scaling process.

Another useful study would be to examine the factor structure of above-level scores and compare the factor structure for gifted adolescents with the factor structure for the older group for whom the test was designed. Minor and Benbow (1996) conducted such a study with items from the SAT-M, but it is problematic because they (a) created item parcels in order to have more normal data distributions for their confirmatory factor analyses, and (b) they did not test item intercepts when examining measurement invariance across groups. A study that corrects these flaws by taking into

account the dichotomous nature of the items and includes item intercepts in the test of invariance would be more likely to detect differences in factor structure and item properties across groups. The data generated for this sample do not permit such a comparison because the sample sizes are too small and data do not include average ability students in the grades that correspond to the test levels.

The possibility of item bias in above-level testing has also been insufficiently examined. Benbow and Wolins (1996) conducted a study investigating item bias among seventh- and eighth-graders taking the SAT and found no substantial levels of item bias between genders. This study was important because Talent Search populations have always been majority male (e.g., Lubinski & Benbow, 1994; S.-Y. Lee & Olszewski-Kubilius, 2006), and Benbow and Wolin's findings indicated that SAT gender bias was not a cause of the gender imbalance in Talent Search populations. However, other potential types of bias, such as bias against different ethnic groups, have not been subject to investigation. Such tests of bias are needed because Talent Search populations have also been nonrepresentative of ethnicity of the general population from which they come, with Asian Americans and Whites usually overrepresented and other ethnic groups underrepresented (e.g., S.-Y. Lee & Olszewski-Kubilus, 2006; Olszewski-Kubilius & S.-Y. Lee, 2011). It would also be important to test different tests, such as the ITBS, ACT, and EXPLORE, for gender and ethnicity bias in items or tests. I was not able to examine item gender or ethnicity bias in this study because of a small sample size for this type of study.

Factor analyses—both exploratory and confirmatory—of above-level testing data should become standard practice in gifted education research. Unfortunately, factor analysis was not possible with these data because the sample size was not large enough. I find it disappointing that only a single factor analysis has ever been performed on above-level testing data (Minor & Benbow, 1996), despite the fact that above-level testing has been a widely accepted practice in gifted education for at least 25 years. Factor analyses provide important information about test structure and interpretation that is not obtainable from any other analysis or practice. With the myriad possible combinations of test, test level, type of giftedness, sample age, etc., I propose that any researcher who uses above-level tests should provide information of factor structure of their particular data, provided that the sample size is large enough. These factor analyses should also be conducted and reported as part of the larger effort to assess the validity of above-level testing.

Another possible future research line would be to strengthen the external validity evidence of above-level test scores and the interpretations of those scores as provided by researchers like Swiatek (2007). For example, a study that examines the correlation between above-level test scores and other criteria such as IQ tests, algebra readiness, and AP tests taken at a young age would be impressive.

Future research could also examine the relationship between above-level test score and learning speed. A set of regression equations that could predict the probability that a child could master advanced coursework in a limited amount of time would be helpful for educational planning. Currently, no empirical investigations have been

performed to determine the relationship between above-level test scores and the rate of educational acceleration that a child could handle, although some researchers have produced educated guesses based on extensive experience (e.g., Olszewski-Kubilius, 1998b; Rogers, 2002; Ruf, 2005; VanTassel-Baska, 1984).

Finally, another interesting area of research would be to compare the psychometric research on above-level testing with the research on below-level testing (see Ayrer & McNamara, 1973, for an early example of below-level testing). There are similarities in the measurement problems that special education and gifted education researchers grapple with when using grade-level tests, such as restriction of range problems and a high measurement error (e.g., Roberts, 1976), so it is likely that psychometric issues related to above-level testing have corollaries in below-level testing. However, below-level testing has largely fallen out of favor in special education because of interpretation difficulties and because some assumptions of below-level testing advocates have been strongly questioned by empirical research (e.g., Bielinski et al., 2000; Minnema et al., 2000, 2001). This is a stark contrast from gifted education, where above-level testing enjoys widespread support, and what little psychometric research there is on the practice is favorable. Articles about the commonalities and differences between the above- and below-level testing and why the research supports the former practice but not the latter would be illuminating.

Conclusion

In conclusion, this study provides evidence that administering ITBS or ITED test levels to a group of gifted students who are two years younger than the norm group

produces results that are similar to what has been reported in administering other achievement tests to more extreme groups of gifted children. Specifically, above-level testing raised the test ceiling, test scores became more variable than would be expected with a gifted sample taking a grade-level test, and observed score reliability was high. Moreover, the test scores can be used to track individual progress, although there is some evidence that regression toward the mean may still be a problem for some examinees, even with above-level testing.

The study also found that some student demographic characteristics had an influence on both above-level test scores and the rate of score growth. Ethnicity was found to be a powerful influence on the initial scores for the reading, mathematics, and total battery and to be a moderating variable for growth of mathematics and total scores. Gender also was a moderator variable for reading score growth, but did not produce any statistically significant main effects. SES had a statistically significant relationship with above-level test scores, but SES provided little unique explained variance above and beyond what ethnicity provided.

Finally, the correlation between initial score and rate of score growth was negative. This means that the highest scoring students were the ones who demonstrated the smallest gains over the course of the study. The cause of this theoretically unexpected finding is unknown, but it may reflect local characteristics or the remnants of a ceiling effect.

This study was designed to be a starting point for a future line of research, and this chapter provides just a few possibilities for future research. There is so little in-

depth, high quality psychometric research on above-level testing that no study could possibly be the last word on the subject. Moreover, this study has raised new questions about when to test above-level outside of a Talent Search context, program evaluation, regression toward the mean, and other issues. Answering new questions about above-level testing will take more research and studies on a wide array of gifted populations.

The field of gifted education is ready for more psychometric research. I believe that the field is undergoing a revolution in methodology and statistics, as demonstrated by several recent works (e.g., Matthews, Gentry, McCoach, Worrell, Matthews, & Dixon, 2008; Shore, 2006; Thompson & Subotnik, 2010; VanTassel-Baska, 2006). Above-level testing is an ideal battlefront for this revolution because the practice is so widely accepted, yet poorly understood. As understanding of psychometric issues and above-level testing grow, researchers and practitioners may become more thoughtful about all of their psychometric data, which may improve the quality of research and practice.

REFERENCES

Achter, J. A., Lubinski, D., & Benbow, C. P. (1996). Multipotentiality among the intellectually gifted: "It was never there and already it's vanishing." *Journal of Counseling Psychology, 43*, 65-76. doi: 10.1037/0022-0167.43.1.65

Almack, J. C., & Almack, J. S. (1921). Gifted pupils in the high school. *School & Society, 14*, 227-228.

American Educational Research Association. (2006). Standards for reporting on empirical social science research in AERA publications. *Educational Researcher, 35*(6), 33-40. doi: 10.3102/0013189x035006033

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Assouline, S., Colangelo, N., Lupkowski-Shoplik, A., Lipscomb, J., & Forstadt, L. (2009). *Iowa acceleration scale manual* (3rd ed.). Scottsdale, AZ: Great Potential Press.

Ayrer, J. E., & McNamara, T. C. (1973). Survey testing on an out-of-level basis. *Journal of Educational Measurement, 10*, 79-84. doi: 10.1111/j.1745-3984.1973.tb00785.x

Barnett, L. B., & Gilheany, S. (1996). The CTY Talent Search: International applicability and practice in Ireland. *High Ability Studies, 7*, 179-190. doi: 10.1080/0937445960070208

Benbow, C. P. (1992). Academic achievement in mathematics and science of students between ages 13 and 23: Are there differences among students in the top one percent of mathematical ability? *Journal of Educational Psychology, 84*, 51-61. doi: 10.1037/0022-0663.84.1.51

Benbow, C. P., & Lubinski, D. (2006). Julian C. Stanley Jr. (1918-2005). *American Psychologist, 61*, 251-252. doi: 10.1037/0003-066X.61.3.251

Benbow, C. P., & Stanley, J. C. (1980). Sex differences in mathematical ability: Fact or artifact? *Science, 210*, 1262-1264. doi: 10.1126/science.7434028

Benbow, C. P., & Wolins, L. (1996). The utility of out-of-level testing for gifted seventh and eighth graders using the SAT-M: An examination of item bias. In C. P. Benbow & D. Lubinski (Eds.), *Intellectual talent: Psychometric and social issues* (pp. 333-346, 413-417). Baltimore, MD: Johns Hopkins University Press.

Bielinski, J., Thurlow, M., Minnema, J., & Scott, J. (2000). *How out-of-level testing affects the psychometric quality of test scores. Out-of-level testing report 2*. Retrieved from ERIC database. (ED449174)

Borland, J. H. (2003). Evaluating gifted programs: A broader perspective. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (3rd ed., pp. 293-307). Boston: Allyn and Bacon.

Brooks-Gunn, J., & Duncan, G. J. (1997). The effects of poverty on children. *The Future of Children, 7*(2), 55-71. doi: 10.2307/1602387

Burks, B. S., Jensen, D. W., & Terman, L. M. (1930). *Genetic studies of genius: Vol. III. The promise of youth: Follow-up studies of a thousand gifted children*. Stanford, CA: Stanford University Press.

Carman, C. A., & Taylor, D. K. (2010). Socioeconomic status effects on using the Naglieri Nonverbal Ability Test (NNAT) to identify the gifted/talented. *Gifted Child Quarterly, 54*, 75-84. doi: 10.1177/0016986209355976

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York, NY: Cambridge University Press.

Colangelo, N., Assouline, S. G., & Gross, M. U. M. (Eds.). (2004). *A nation deceived: How schools hold back America's brightest students* (Vol. 2). Iowa City, IA: University of Iowa.

Colangelo, N., Assouline, S. G., & Lu, W.-H. (1994). Using EXPLORE as an above-level instrument in the search for elementary student talent. In N. Colangelo, S. G. Assouline & D. L. Ambroson (Eds.), *Talent development: Proceedings from the 1993 H. B. and Jocelyn Wallace National Research Symposium on Talent Development* (pp. 281-297). Dayton, OH: Ohio Psychology Press.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimental designs: Design and analysis issues for field experiments*. Chicago, IL: Rand McNally.

Corno, L., Cronbach, L. J., Kupermintz, H., Lohman, D. F., Mandinach, E. B., Porteus, A. W., & Talbert, J. E. (2002). *Remaking the concept of aptitude: Extending the legacy of Richard E. Snow*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cortina, J. M. (2002). Big things have small beginnings: An assortment of "minor" methodological misunderstandings. *Journal of Management, 28*, 339-362. doi: 10.1016/s0149-2063(02)00131-9

Dai, D. Y. (2010). *The nature and nurture of giftedness*. New York, NY: Teachers College Press.

Deary, I. J., Strand, S., Smith, P., & Fernandes, C. (2007). Intelligence and educational achievement. *Intelligence, 35*, 13-21. doi: 10.1016/j.intell.2006.02.001

Ebmeier, H., & Schmulbach, S. (1989). An examination of the selection practices used in the Talent Search Program. *Gifted Child Quarterly, 33*, 134-141. doi: 10.1177/001698628903300402

Eisner, E. W. (2002). The kind of schools we need. *Phi Delta Kappan, 83*, 576-583.

Feldhusen, J. F., Proctor, T. B., & Black, K. N. (2002). Guidelines for grade advancement of precocious children. *Roeper Review. Special Issue: A quarter century of ideas on ability grouping and acceleration, 24*, 169-171. doi: 10.1080/02783198609553000

Feldt, L. S. (1965). The approximate sampling distribution of Kuder-Richardson reliability coefficient twenty. *Psychometrika, 30*, 357-370. doi: 10.1007/BF02289499

Ferron, J. M., Hogarty, K. Y., Dedrick, R. F., Hess, M. R., Niles, J. D., & Kromrey, J. D. (2008). Reporting results from multilevel analyses. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 391-426). Charlotte, NC: Information Age Publishing.

Ford, D. Y. (2003). Equity and excellence: Culturally diverse students in gifted education. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (pp. 506-520). Boston: Allyn and Bacon.

Forsyth, R. A., Ansley, T. N., Feldt, L. S., & Alnot, S. D. (2001). *Iowa Tests of Educational Development*. Itasca, IL: Riverside Publishing Company.

Forsyth, R. A., Ansley, T. N., Feldt, L. S., & Alnot, S. D. (2003). *Iowa Tests of Educational Development guide to research and development*. Itasca, IL: Riverside Publishing.

Francis, B. (2000). The gendered subject: Students' subject preferences and discussions of gender and subject ability. *Oxford Review of Education, 26*, 35-48. doi: 10.1080/030549800103845

Gagné, F. (2005). From noncompetence to exceptional talent: Exploring the range of academic achievement within and between grade levels. *Gifted Child Quarterly, 49*, 139-153. doi: 10.1177/001698620504900204

Gallagher, J. J. (2006). According to Jim Gallagher: How to shoot oneself in the foot with program evaluation. *Roeper Review, 28*, 122-124. doi: 10.1080/02783190609554350

Gershon, R. C. (2005). Computer adaptive testing. *Journal of Applied Measurement, 6*, 109-127.

Gottfredson, L. S. (1997a). Mainstream science on intelligence: An editorial with 52 signatories, history, and bibliography. *Intelligence, 24*, 13-23. doi: 10.1016/S0160-2896(97)90011-8

Gottfredson, L. S. (1997b). Why *g* matters: The complexity of everyday life. *Intelligence, 24*, 79-132. doi: 10.1016/S0160-2896(97)90014-3

Gottfredson, L. S. (1998, Winter). The General Intelligence Factor. *Scientific American Presents*, 24-29.

Gottfredson, L. S. (2000). Skill gaps, not tests, make racial proportionality impossible. *Psychology, Public Policy, and Law, 6*, 129-143. doi: 10.101317//1076-8971.6.1.129

Gottfredson, L. S. (2003). The science and politics of intelligence in gifted education. In N. Colangelo & G. A. Davis (Eds.), *Handbook of gifted education* (3rd ed., pp. 24-40). Boston: Allyn and Bacon.

Gross, M. U. M. (1999). Small poppies: Highly gifted children in the early years. *Roeper Review, 21*, 207-214. doi: 10.1080/02783199909553963

Gross, M. U. M. (2004). *Exceptionally gifted children* (2nd ed.). New York, NY: Routledge.

Harris, D. J., & Hoover, H. D. (1987). An application of the three-parameter IRT model to vertical equating. *Applied Psychological Measurement, 11*, 151-159. doi: 10.1177/014662168701100203

Hedeker, D., & Gibbons, R. D. (2006). *Longitudinal data analysis*. Hoboken, NJ: John Wiley & Sons.

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

Hogan, T. P., Benjamin, A., & Brezinski, K. L. (2000). Reliability methods: A note on the frequency of use of various types. *Educational and Psychological Measurement, 60*, 523-531. doi: 10.1177/00131640021970691

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187-220). Westport, CT: Praeger Publishers.

Hollingworth, L. S. (1926). *Gifted children: Their nature and nurture*. New York, NY: Macmillan.

Hollingworth, L. S. (1942). *Children above 180 IQ, Stanford-Binet: Origin and development*. Yonkers-on-Hudson, NY: World Book.

Hoover, H. D., Dunbar, S. B., & Frisbie, D. A. (2001). *Iowa tests of basic skills, forms A, B, and C*. Itasca, IL: Riverside Publishing Company.

Hoover, H. D., Dunbar, S. B., Frisbie, D. A., Oberley, K. R., Ordman, V. L., Naylor, R. J., . . . Shannon, G. P. (2003). *Iowa Tests of Basic Skills guide to research and development*. Itasca, IL: Riverside Publishing.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum Associates.

Johnsen, S. K., & Corn, A. L. (2001). *Screening assessment for gifted elementary and middle school students examiner's manual*. Austin, TX: PRO-ED.

Kaplan, R. M., & Saccuzzo, D. P. (2005). *Psychological testing: Principles, applications, and issues* (6th ed.). Belmont, CA: Thomson Wadsworth.

Keating, D. P. (1975). Testing those in the top percentiles. *Exceptional Children, 41*, 435-436.

Keating, D. P. (1976). Discovering quantitative precocity. In D. P. Keating (Ed.), *Intellectual talent: Research and development* (pp. 23-31). Baltimore: Johns Hopkins University Press.

Keating, D. P., & Stanley, J. C. (1972). Extreme measures for the exceptionally gifted in mathematics and science. *Educational Research, 1*, 3-7.

Kelley, T. L. (1923). A new method for determining the significance of differences in intelligence and achievement scores. *Journal of Educational Psychology, 14*, 321-333. doi: 10.1037/h0072213

Kieffer, K. M., Reese, R. J., & Vacha-Haase, T. (2010). Reliability generalization methods in the context of giftedness research. In B. Thompson & R. F. Subotnik (Eds.), *Methodologies for conducting research on giftedness* (pp. 89-111). Washington, DC: American Psychological Association.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18*, 1-11. doi: 10.1111/j.1745-3984.1981.tb00838.x

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155-186). Westport, CT: Praeger Publishers.

Kolen, M. J., Hanson, B. A., & Brennan, R. L. (1992). Conditional standard errors of measurement for scale scores. *Journal of Educational Measurement, 29*, 285-307. doi: 10.1111/j.1745-3984.1992.tb00378.x

Konstantopoulos, S., Modi, M., & Hedges, L. V. (2001). Who are America's gifted? *American Journal of Education, 109*, 344-382. doi: 10.1086/444275

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of reliability. *Psychometrika, 2*, 151-160. doi: 10.1007/BF02288391

Lee, J. (2002). Racial and ethnic achievement gap trends: Reversing the progress toward equity? *Educational Researcher, 31*(1), 3-12. doi: 10.3102/0013189X031001003

Lee, S.-Y., Matthews, M. S., & Olszewski-Kubilius, P. (2008). A national picture of Talent Search and Talent Search educational programs. *Gifted Child Quarterly, 52*, 55-69. doi: 10.1177/0016986207311152

Lee, S.-Y., & Olszewski-Kubilius, P. (2006). Talent search qualifying: Comparisons between talent search students qualifying via scores on standardized tests and via parent nomination. *Roeper Review, 28*, 157-166. doi: 10.1080/02783190609554355

Lewis, J. D., DeCamp-Fritson, S. S., Ramage, J. C., McFarland, M. A., & Archwamety, T. (2007). Selecting for ethnically diverse children who may be gifted using Raven's Standard Progressive Matrices and Naglieri Nonverbal Abilities Test. *Multicultural Education, 15*(1), 38-42.

Lohman, D. F. (2005). The role of nonverbal ability tests in identifying academically gifted students: An aptitude perspective. *Gifted Child Quarterly, 49*, 111-138. doi: 10.1177/001698620504900203

Lohman, D. F., & Hagen, E. P. (2002). *CogAT Form 6 research handbook*. Itasca, IL: Riverside Publishing.

Lohman, D. F., & Korb, K. A. (2006). Gifted today but not tomorrow? Longitudinal changes in ability and achievement during elementary school. *Journal for the Education of the Gifted, 29*, 451-484. doi: 10.4219/jeg-2006-245

Loyd, B. H. (1980). *Functional level testing and reliability: An empirical study.* Doctoral dissertation, University of Iowa, Iowa City, IA.

Lubinski, D., & Benbow, C. P. (1994). The study of mathematically precocious youth: The first three decades of a planned 50-year study of intellectual talent. In R. F. Subotnik & K. D. Arnold (Eds.), *Beyond Terman: Contemporary longitudinal studies of giftedness and talent* (pp. 255-281). Westport, CT: Ablex.

Lubinski, D., & Benbow, C. P. (2006). Study of Mathematically Precocious Youth after 35 years: Uncovering antecedents for the development of math-science expertise. *Perspectives on Psychological Science, 1*, 316-345. doi: 10.1111/j.1745-6916.2006.00019.x

Lubinski, D., Webb, R. M., Morelock, M. J., & Benbow, C. P. (2001). Top 1 in 10,000: A 10-year follow-up of the profoundly gifted. *Journal of Applied Psychology, 86*, 718-729. doi: 10.1037/0021-9010.86.4.718

Lupkowski-Shoplik, A. E., & Assouline, S. G. (1993). Identifying mathematically talented elementary students: Using the lower level of the SSAT. *Gifted Child Quarterly, 37*, 118-123. doi: 10.1177/001698629303700304

Lupkowski-Shoplik, A., Benbow, C. P., Assouline, S. G., & Brody, L. E. (2003). Talent searches: Meeting the needs of academically talented youth. In N. Colangelo &

G. A. Davis (Eds.), *Handbook of gifted education* (3rd ed., pp. 204-218). Boston: Allyn and Bacon.

Lupkowski-Shoplik, A., & Swiatek, M. A. (1999). Elementary student talent searches: Establishing appropriate guidelines for qualifying test scores. *Gifted Child Quarterly, 43*, 265-272. doi: 10.1177/001698629904300405

Madsen, I. N. (1920). High-school students' intelligence ratings according to the Army Alpha test. *School & Society, 11*, 298-300.

Madsen, I. N., & Sylvester, R. H. (1919). High-school students' intelligence ratings according to the Army Alpha test. *School & Society, 10*, 407-410.

Matthews, M. S. (2008). Talent Search programs. In J. A. Plucker & C. M. Callahan (Eds.), *Critical issues and practices in gifted education* (pp. 641-654). Waco, TX: Prufrock Press.

Matthews, M. S., Gentry, M., McCoach, D. B., Worrell, F. C., Matthews, D., & Dixon, F. (2008). Evaluating the state of a field: Effect size reporting in gifted education. *Journal of Experimental Education, 77*, 55-68. doi: 10.3200/JEXE.77.1.55-68

McBee, M. T. (2006). A descriptive analysis of referral sources for gifted identification screening by race and socioeconomic status. *Journal of Secondary Gifted Education, 17*, 103-111. doi: 10.4219/jsge-2006-686

McBee, M. (2010). Modeling outcomes with floor or ceiling effects: An introduction to the Tobit model. *Gifted Child Quarterly, 54*, 314-320. doi: 10.1177/0016986210379095

McCoach, D. B. (2010a). Dealing with dependence (part II): A gentle introduction to hierarchical linear modeling. *Gifted Child Quarterly, 54*, 252-256. doi: 10.1177/0016986210373475

McCoach, D. B. (2010b). Hierarchical linear modeling. In G. R. Hancock & R. O. Mueller (Eds.), *The reviewer's guide to quantitative methods in the social sciences* (pp. 123-140). New York, NY: Routledge.

Meade, A. W., & Kroustalis, C. M. (2006). Problems with item parceling for confirmatory factor analytic tests of measurement invariance. *Organizational Research Methods, 9*, 369-403. doi: 10.1177/1094428105283384

Merwin, J. C., & Gardner, E. F. (1962). Development and application of tests of educational achievement. *Review of Educational Research, 32*, 40-50. doi: 10.2307/1169202

Mills, C. J., & Barnett, L. B. (1992). The use of the Secondary School Admission Test (SSAT) to identify academically talented elementary school students. *Gifted Child Quarterly, 36*, 155-159. doi: 10.1177/001698629203600306

Mills, J. R., & Jackson, N. E. (1990). Predictive significance of early giftedness: The case of precocious reading. *Journal of Educational Psychology, 82*, 410-419. doi: 10.1037//0022-0663.82.3.410

Minnema, J., Thurlow, M., Bielinski, J., & Scott, J. (2000). *Past and present understandings of out-of-level testing: A research synthesis. Out-of-level testing report 1*. Retrieved from ERIC database. (ED446409)

Minnema, J. E., Thurlow, M. L., Bielinski, J., & Scott, J. K. (2001). Past and current research on out-of-level testing of students with disabilities. *Assessment for Effective Intervention, 26*(2), 49-55. doi: 10.1177/073724770102600208

Minor, L. L., & Benbow, C. P. (1996). Construct validity of the SAT-M: A comparative study of high school students and gifted seventh graders. In C. P. Benbow & D. Lubinski (Eds.), *Intellectual talent: Psychometric and social issues* (pp. 347-361). Baltimore, MD: Johns Hopkins University Press.

Morelock, M. J. (1992). Giftedness: The view from within. *Understanding Our Gifted, 4*(3), 11-15.

Nasser, F., & Wisenbaker, J. (2003). A Monte Carlo study investigating the impact of item parceling on measures of fit in confirmatory factor analysis. *Educational and Psychological Measurement, 63*, 729-757. doi: 10.1177/0013164403258228

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., . . . Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51*, 77-101. doi: 10.1037/0003-066X.51.2.77

O'Brien, M. L., & Tohn, D. (1984). Applying and evaluating Rasch vertical equating procedures for out-of-level testing. Retrieved from ERIC database. (ED246071)

Olszewski-Kubilius, P. (1998a). Research evidence regarding the validity and effects of talent search educational programs. *Journal of Secondary Gifted Education, 9*, 134-138.

Olszewski-Kubilius, P. (1998b). Talent search: Purposes, rationale, and role in gifted education. *Journal of Secondary Gifted Education, 9*, 106-113.

Olszewski-Kubilius, P., & Kulieke, M. J. (2008). Using off-level testing and assessment for gifted and talented students. In J. VanTassel-Baska (Ed.), *Alternative assessments with gifted and talented students* (pp. 89-106). Waco, TX: Prufrock Press.

Olszewski-Kubilius, P. M., Kulieke, M. J., Willis, G. B., & Krasney, N. S. (1989). An analysis of the validity of SAT entrance scores for accelerated classes. *Journal for the Education of the Gifted, 13*, 37-54.

Olszewski-Kubilius, P., & Lee, S.-Y. (2011). Gender and other group differences in performance on off-level tests: Changes in the 21st century. *Gifted Child Quarterly, 55*, 54-73. doi: 10.1177/0016986210382574

Olszewski-Kubilius, P., & Turner, D. (2002). Gender differences among elementary school-aged gifted students in achievement, perceptions of ability, and subject preference. *Journal for the Education of the Gifted, 25*, 233-268. doi: 10.4219/jeg-2002-279

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221-262). New York: American Council on Education.

Plomin, R., & Petrill, S. A. (1997). Genetics and intelligence: What's new? *Intelligence, 24*, 53-77. doi: 10.1016/s0160-2896(97)90013-1

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2009). HLM 6.08 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

Roberts, A. O. H. (1976). *Out-of-level testing. ESEA Title I evaluation and reporting system. Technical paper no. 6*. Retrieved from ERIC database. (ED169126)

Rodriguez, M. C., & Maeda, Y. (2006). Meta-analysis of coefficient alpha. *Psychological Methods, 11*, 306-322. doi: 10.1037/1082-989X.11.3.306

Rogers, K. B. (2002). *Re-forming gifted education: How parents and teachers can match the program to the child*. Scottsdale, AZ: Great Potential Press.

Roid, G. H. (2003). *Stanford-Binet Intelligence Scales, Fifth Edition, Technical Manual*. Itasca, IL: Riverside Publishing.

Rowe, D. C. (1997). A place at the policy table? Behavior genetics and estimates of family environmental effects on IQ. *Intelligence, 24*, 133-158. doi: 10.1016/s0160-2896(97)90015-5

Ruf, D. L. (2005). *Losing our minds: Gifted children left behind*. Scottsdale, AZ: Great Potential Press.

Scarr, S., & Weinberg, R. A. (1976). IQ test performance of Black children adopted by White families. *American Psychologist, 31*, 726-739. doi: 10.1037/0003-066x.31.10.726

Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307-353). Westport, CT: Praeger Publishers.

Shore, B. M. (2006). Yogi Berra's Chevy truck: A report card on the state of research in the field of gifted education. *Gifted Child Quarterly, 50*, 351-353. doi: 10.1177/001698620605000409

Skaggs, G., & Lissitz, R. W. (1986). IRT test equating: Relevant issues and a review of recent research. *Review of Educational Research, 56*, 495-529. doi: 10.2307/1170343

Spybrook, J. (2008). Power, sample size, and design. In A. A. O'Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data* (pp. 273-311). Charlotte, NC: Information Age Publishing.

Stanley, J. C., Jr. (1951). On the adequacy of standardized tests administered to extreme norm groups. *Peabody Journal of Education, 29*, 145-153.

Stanley, J. C. (1954). Identification of superior learners in grades ten through fourteen. In H. Robinson (Ed.), *Promoting maximal reading growth among able learners* (Supplementary Educational Monographs No. 81, pp. 31-34). Chicago, IL: University of Chicago Press.

Stanley, J. C. (1976). The case for extreme educational acceleration of intellectually brilliant youths. *Gifted Child Quarterly, 20*, 66-75. doi: 10.1177/001698627602000120

Stanley, J. C. (1977). Rationale of the Study of Mathematically Precocious Youth (SMPY) during its first five years of promoting educational acceleration. In J. C. Stanley, W. C. George & C. H. Solano (Eds.), *The gifted and the creative: A*

fifty-year perspective (pp. 75-112). Baltimore: The Johns Hopkins University Press.

Stanley, J. C. (1990). Leta Hollingworth's contributions to above-level testing of the gifted. *Roeper Review, 12*, 166-171. doi: 10.1080/02783199009553264

Stanley, J. C. (2005). A quiet revolution: Finding boys and girls who reason exceptionally well and/or verbally and helping them get the supplemental educational opportunities they need. *High Ability Studies, 16*, 5-14. doi: 10.1080/13598130500115114

Stanley, J. C., & Benbow, C. P. (1981-1982). Using the SAT to find intellectually talented seventh graders. *College Board Review*, (122), 2-7, 26-27.

Stanovich, K. E. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly, 21*, 360-406. doi: 10.1598/rrq.21.4.1

Stedman, L. M. (1924). *Education of gifted children*. Yonkers-on-Hudson, NY: World Book.

Swiatek, M. A. (2007). The talent search model: Past, present, and future. *Gifted Child Quarterly, 51*, 320-329. doi: 10.1177/0016986207306318

Swiatek, M. A., & Lupkowski-Shoplik, A. (2005). An evaluation of the elementary student Talent Search by families and schools. *Gifted Child Quarterly, 49*, 247-259. doi: 10.1177/001698620504900306

Terman, L. M. (1926). *Genetic studies of genius: Vol. I. Mental and physical traits of a thousand gifted children.* (2nd ed.). Stanford, CA: Stanford University Press.

Terman, L. M., & Fenton, J. C. (1921). Preliminary report on a gifted juvenile author. *Journal of Applied Psychology, 5*, 163-178. doi: 10.1037/h0074962

Terman, L. M., & Oden, M. H. (1947). *Genetic studies of genius: Vol. IV. The gifted child grows up: Twenty-five years' follow-up of a superior group*. Stanford, CA: Stanford University Press.

Terman, L. M., & Oden, M. H. (1959). *Genetic studies of genius: Vol. V. The gifted group at mid-life: Thirty-five years' follow-up of the superior child*. Stanford, CA: Stanford University Press.

Thompson, B. (2003). Understanding reliability and coefficient alpha, really. In B. Thompson (Ed.), *Score reliability: Contemporary thinking on reliability issues* (pp. 3-23). Thousand Oaks, CA: Sage.

Thompson, B. (2006). *Foundations of behavioral statistics*. New York, NY: The Guilford Press.

Thompson, B., & Subotnik, R. F. (2010). *Methodologies for conducting research on giftedness*. Washington, DC: American Psychological Association.

Thompson, B., & Vacha-Haase, T. (2000). Psychometrics is datametrics: The test is not reliable. *Educational and Psychological Measurement, 60*, 174-195. doi: 10.1177/00131640021970448

Threlfall, J., & Hargreaves, M. (2008). The problem-solving methods of mathematically gifted and older average-attaining students. *High Ability Studies, 19*, 83-98. doi: 10.1080/13598130801990967

Vacha-Haase, T., Kogan, L. R., & Thompson, B. (2000). Sample compositions and variabilities in published studies versus those in test manuals: Validity of score reliability inductions. *Educational and Psychological Measurement, 60*, 509-522. doi: 10.1177/00131640021970682

van Wagenen, M. J. (1925). A comparison of the mental ability and school achievement of the bright and dull pupils in the sixth grade of a large school system. *Journal of Educational Psychology, 16*, 186-192. doi: 10.1037/h0067265

VanTassel-Baska, J. (1984). The talent search as an identification model. *Gifted Child Quarterly, 28*, 172-176. doi: 10.1177/001698628402800406

VanTassel-Baska, J. (1996). Contributions of the talent-search concept to gifted education. In C. P. Benbow & D. Lubinski (Eds.), *Intellectual talent: Psychometric and social issues* (pp. 236-245). Baltimore, MD: Johns Hopkins University Press.

VanTassel-Baska, J. (2006). NAGC symposium: A report card on the state of research in the field of gifted education. *Gifted Child Quarterly, 50*, 339-341. doi: 10.1177/001698620605000406

VanTassell-Baska, J. (1986). The use of aptitude tests for identifying the gifted: The talent search concept. *Roeper Review, 8*, 185-189. doi: 10.1080/02783198609552970

Wai, J., Lubinski, D., & Benbow, C. P. (2005). Creativity and occupational accomplishments among intellectually precocious youths: An age 13 to age 33

longitudinal study. *Journal of Educational Psychology, 97*, 484-492. doi: 10.1037/0022-0663.97.3.484

Warne, R. T. (2009). Comparing tests used to identify ethnically diverse gifted children: A critical response to Lewis, DeCamp-Fritson, Ramage, McFarland, & Archwamety. *Multicultural Education, 17*(1), 48-53.

Warne, R. T. (2011). A reliability generalization of the Overexcitability Questionnaire-Two (OEQII). Manuscript submitted for publication.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children--Fourth Edition technical and interpretive manual*. San Antonio, TX: The Psychological Corporation.

Wilkinson, L., & the Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54*, 594-604. doi: 10.1037/0003-066X.54.8.594

Witty, P. A., & Jenkins, M. D. (1935). The case of "B"--a gifted negro girl. *The Journal of Social Psychology, 6*, 117-124. doi: 10.1080/0022545.1935.9921630

Yin, P., Brennan, R. L., & Kolen, M. J. (2004). Concordance between ACT and TED scores from different populations. *Applied Psychological Measurement, 28*, 274-289. doi: 10.1177/0146621604265034

Yoakum, C. S., & Yerkes, R. M. (1920). *Army mental tests*. New York, NY: Henry Holt and Company.

Yoon, S. Y., & Gentry, M. (2009). Racial and ethnic representation in gifted programs: Current status of and implications for gifted Asian American students. *Gifted Child Quarterly, 53*, 121-136. doi: 10.1177/0016986208330564

Ziegler, A., & Ziegler, A. (2009). The paradoxical attenuation effect in tests based on classical test theory: Mathematical background and practical implications for the measurement of high abilities. *High Ability Studies, 20*, 5-14. doi: 10.1080/13598130902860473

Zwick, R. (2006). Higher education admissions testing. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 647-679). Westport, CT: Praeger Publishers.

APPENDIX

ITEM STATISTICS

The research questions in this dissertation were mostly concerned with the students in the study: their initial scores, their growth rates, the impact of demographic variables on their scores, etc. Only one research question—about internal consistency reliability—was directly concerned with a psychometric issue. However, I recognize that above-level item properties may also be of interest to some readers. Hence, in this appendix I will the item difficulty index (IDI) for the two different groups in order to examine how above-level testing impacts IDI values. The IDI is merely the proportion of the sample that answered the item correctly, which means that—counterintuitively—easier items have higher IDI values (Allen & Yen, 1979; Kaplan & Saccuzzo, 2005).

From a psychometric perspective, item statistics are important to examine when instruments and items are administered to a sample that differs from the population for which the test was developed (Crocker & Algina, 2002). Moreover, item statistics can shed light on the validity of using the ITBS and ITED as above-level tests to evaluate the educational progress of gifted children. Validation of above-level instruments is necessary because current testing standards dictate that test users who use an instrument for a purpose for which it was not originally designed and validated must conduct validation studies themselves (AERA et al., 1999).

Methods

The data in this appendix are drawn from the above-level test administrations described in Chapter III of this dissertation. There were three test levels administered to

the students in this study. The eighth-grade level of the ITBS contained a total of 142 items on three subtests; the reading subtest contained 37 items, the writing subtest contained 59 items, and the math subtest contained 46 items. All of these subtests are also used to generate a total battery score (Hoover et al., 2003).

The ninth- and tenth- grade level of the ITED contained 240 items on six subtests; the vocabulary subtests contained 40 items, the reading comprehension subtest contained 44 items, the spelling subtest contained 30 items, the revising written materials subtest contained 56 items, the mathematics concepts & problem solving subtest contained 40 items, and the mathematics computation subtest contained 30 items. The vocabulary and reading comprehension subtests combine to generate a total reading subscore and the mathematics concepts & problem solving and mathematics computation subtests combine to generate a total mathematics subscore. Like the eighth-grade ITBS test level, all of the items on the ITED levels combine to generate a total battery subscore. The spelling and revising written materials subtests do not contribute to any other scores besides the total battery score (Forsyth et al., 2003).

As Chapter III showed, each test level was administered three times as students during the course of the study. The baseline administration was in fall 2008 and two followups were in spring 2009 and 2010. Therefore, the item statistics from the baseline measurement were compared to only the ITBS and ITED test level fall norms, while the item statistics from the two followup administrations were compared to the test levels' spring norms. The IDI values come from the class item response records, which is one of the reports provided by the test publisher, Riverside Publishing.

Because of the small size of the cohorts (ranging from $n = 37$ to $n = 61$), I combined the cohorts that took the same test levels during the spring of different years. This led to a total of six sets of above-level item statistics: one for each test level in the fall and one for each test level in the spring. For these comparisons n ranged from 39 to 101.

Analysis

Item statistics were calculated for the above-level test to be comparable to the norm item statistics provided by the test publisher. As stated above, IDI is merely a proportion of students who answered the item correctly. This information is easily calculated for the gifted sample and is compared with IDIs on the score reports issued from the test publisher.

Results

IDIs for the gifted students and the norm students are displayed in Tables A1-A15. The information from these tables is also displayed in Figures A1-A30. The figures show a high correlation between gifted IDIs and norm group IDIs for the majority of the tests. For 22 of the subtests, the correlation between the two sets of IDIs was quite high ($r > .700$, $p < .001$), indicating that the same items tended to be difficult for both the norm group and the younger gifted students. Conversely, the same items were usually easy for both groups of test takers. The major exceptions to this trend was in the spelling and revising written materials subtests. For both levels (ninth and tenth grade ITED levels) and both administrations (fall and spring), the IDIs of the two set of IDIs did not have a statistically significant correlation ($p \geq .101$).

The mean and standard deviation IDIs for each subtest are reported in Table A16. The table also displays results from two-tailed *t*-tests between the gifted and norm IDIs and standardized effect sizes (Cohen's *d*) indicating the size of the difference between group IDIs. Because there are 30 *t*-tests in Table A16, a Bonferroni correction was use to adjust α and to lessen Type I error. Six of the *t*-tests were shown to be statistically significant at the adjusted α (.002). These were both administrations of both levels of the revising written materials subtests and both administrations of the ninth grade level of the reading comprehension subtest. The effect size for all of these differences was quite large (d = .617 to .825), indicating large differences in IDIs between the younger gifted students and the students in the norm groups. For five of the statistically significant difference, the subtests were easier for the younger gifted sample than it was for the students in the norm sample. The only test that was easier for the norm group was the grade 10 revising written materials subtest that was administered in the spring (d = .761, p < .001).

Discussion

As the figures indicate, the majority of the IDIs did not change drastically when they were used in above-level testing. This result suggests that these items do not function very differently for a younger gifted population than they do for a sample that the tests were designed for. The similarity of IDIs also suggests that some academic achievement subtests—particularly in reading and mathematics—may be interpreted in a similarly, whether they are administered to a traditional population or in an above-level fashion. However, this is just one piece of validity information for above-level testing

and more study is needed on above-level testing items to gather more validity evidence in order to make more sure interpretations of above-level test scores.

For the spelling and revising written materials subtests, however, the correlations between the two sets of IDIs were very low ($|r| \leq .216$). This lack of relationship between IDIs for these subtests may indicate that either (a) the items function differently when administered to a younger gifted sample, or (b) local curriculum and education practices have altered which items are difficult and which are easy for the gifted students in the sample. Until this study is duplicated with another gifted sample, it is impossible to say which of these two options is more likely. However, in a discussion about this finding with the school district official who is in charge of the gifted program, we came to an agreement that (b) is more likely.

I had hoped to also make similar comparisons of the item discrimination index values of the two groups. Measured with a point-biserial correlation (r_{pbis}), which is the correlation between the item score and the total scale score the item discrimination index measures the degree to which items distinguish between high and low scorers on a test (Crocker & Algina, 2002; Kaplan & Saccuzzo, 2005). However, I contacted Riverside Publishing for r_{pbis} values for every item individually, but that information was not available for Form C of the tests (L. Nawojski, personal communication, January 12, 2011). Therefore, it is not possible to examine how items' discriminatory properties change as they are used in above-level testing. This is disappointing because item discrimination indexes are one of the most basic statistics used to evaluate items (Crocker & Algina, 2002; Kaplan & Saccuzzo, 2005).

As far as IDIs are concerned, though, it is interesting that the vast majority of IDIs were similar for both groups, whether that similarity was measured by mean difficulty for the entire test (Table A16) or by correlations between groups (Figures A1 through A30). Although further study is needed into the issue of above-level item statistics, this appendix provides some new information about item functioning in above-level testing.

Table A1
ITBS Reading Subtest Item Difficulty Indexes, Gifted Grade 6 and National Grade 8 Norm Groups

Item #	Fall Gifted IDI	Fall Norm IDI	Spring Gifted IDI	Spring Norm IDI
1	.67	.71	.77	.74
2	.60	.54	.47	.58
3	.62	.59	.53	.62
4	.78	.76	.81	.79
5	.80	.76	.92	.79
6	.82	.86	.96	.89
7	.51	.41	.56	.45
8	.44	.44	.49	.48
9	.58	.73	.79	.76
10	.84	.76	.78	.79
11	.58	.57	.67	.61
12	.33	.46	.48	.50
13	.51	.53	.62	.57
14	.33	.23	.26	.25
15	.56	.38	.62	.41
16	.69	.64	.75	.67
17	.56	.63	.78	.66
18	.33	.50	.47	.54
19	.91	.72	.88	.75
20	.82	.71	.79	.74
21	.33	.36	.46	.39
22	.82	.50	.73	.54
23	.49	.48	.61	.52
24	.78	.51	.83	.55
25	.53	.48	.59	.52
26	.60	.72	.86	.75
27	.62	.54	.81	.58
28	.76	.71	.89	.74
29	.80	.69	.85	.72
30	.53	.50	.56	.54
31	.76	.63	.75	.66
32	.53	.55	.63	.59
33	.56	.54	.76	.58
34	.64	.67	.72	.70
35	.62	.52	.64	.56
36	.51	.52	.71	.56
37	.64	.71	.77	.74

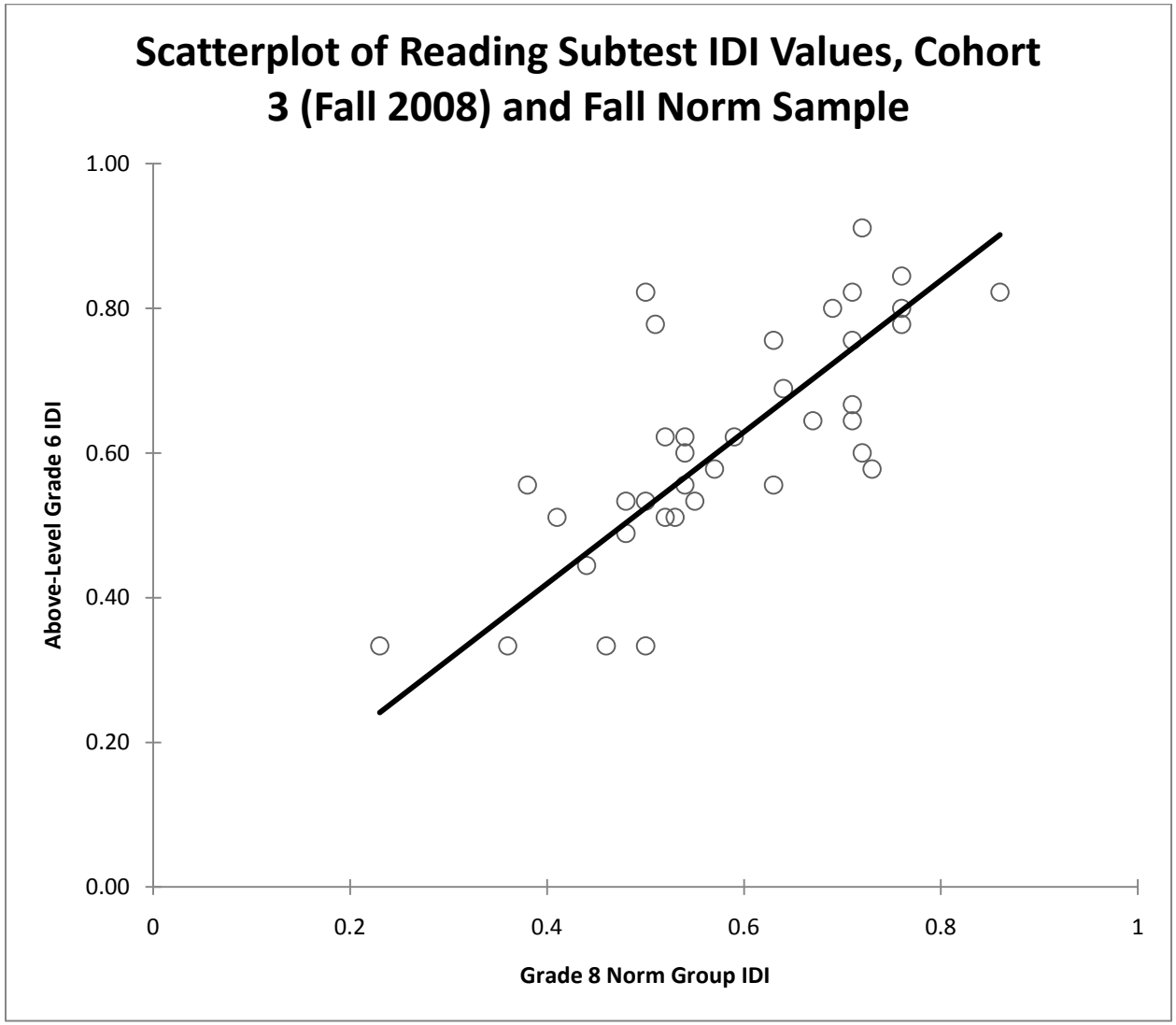

Figure A1 Scatterplot of reading subtest IDI values, Cohort 3 (Fall 2008) and Fall Norm Sample. IDI values correlation is $r = .748$ ($p < .001$).

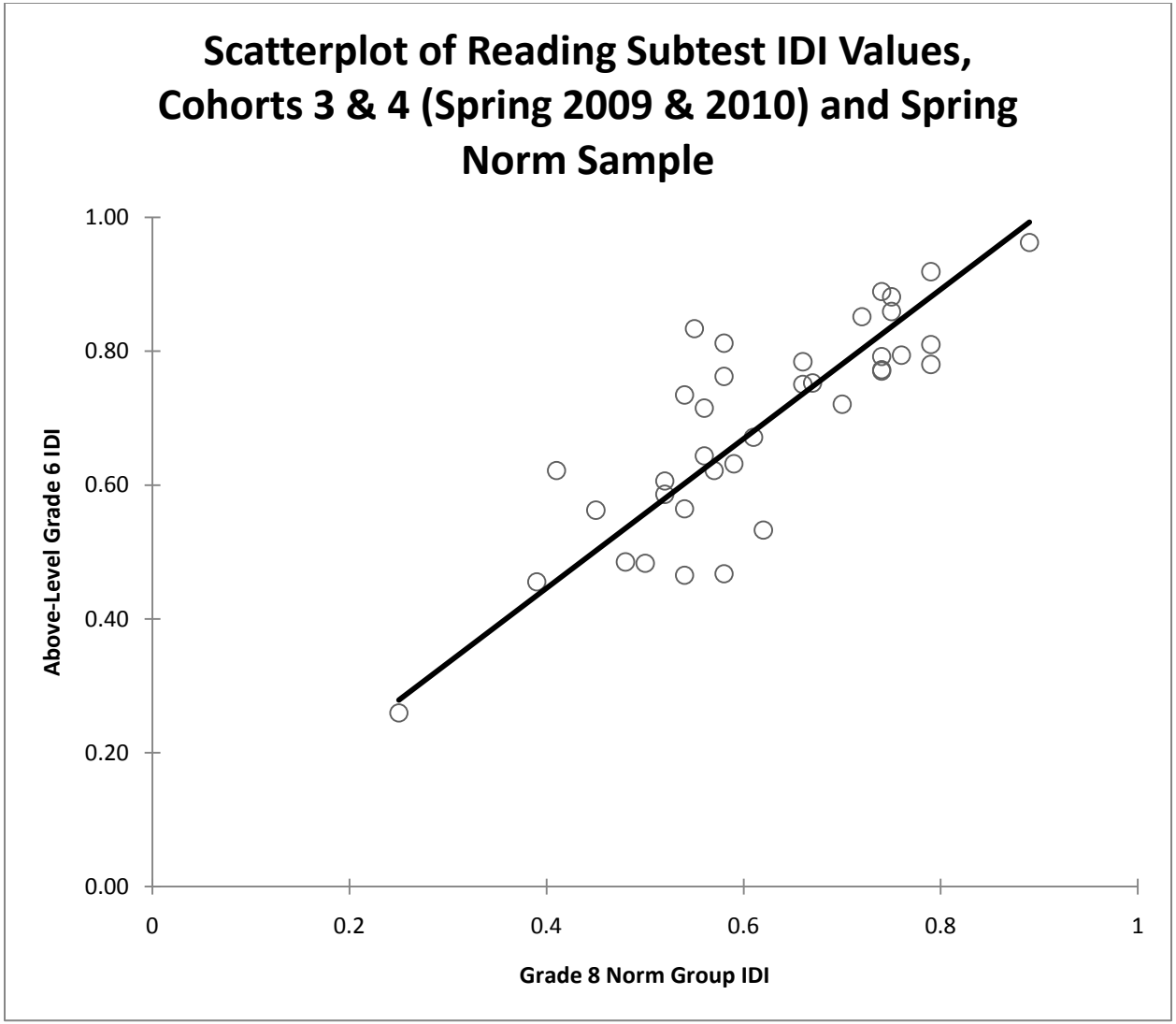

Figure A2 Scatterplot of reading subtest IDI values, Cohorts 3 & 4 (Spring 2009 & 2010) and Spring Norm Sample. IDI values correlation is $r = .838$ ($p < .001$).

Table A2
ITBS Language Subtest Item Difficulty Indexes, Gifted Grade 6 and National Grade 8 Norm Groups

Item #	Fall Gifted IDI	Fall Norm IDI	Spring Gifted IDI	Spring Norm IDI
1	.78	.82	.86	.85
2	.67	.57	.63	.60
3	.71	.60	.70	.63
4	.71	.67	.64	.70
5	.49	.65	.65	.68
6	.71	.61	.70	.34
7	.40	.34	.25	.37
8	.40	.38	.42	.41
9	.53	.51	.53	.54
10	.60	.56	.55	.59
11	.36	.50	.49	.53
12	.29	.35	.19	.38
13	.76	.65	.71	.68
14	.38	.42	.42	.45
15	.24	.32	.12	.35
16	.69	.61	.77	.64
17	.76	.77	.80	.80
18	.71	.79	.89	.82
19	.69	.71	.84	.74
20	.47	.43	.42	.46
21	.58	.51	.59	.54
22	.67	.53	.59	.56
23	.64	.52	.51	.54
24	.42	.47	.33	.50
25	.62	.56	.64	.59
26	.31	.44	.32	.47
27	.73	.49	.70	.52
28	.62	.62	.55	.65
29	.56	.41	.43	.44
30	.73	.57	.66	.60
31	.76	.72	.76	.75
32	.51	.41	.39	.44
33	.80	.78	.74	.81
34	.58	.58	.69	.61
35	.60	.52	.55	.54
36	.42	.28	.29	.31
37	.78	.58	.63	.61
38	.82	.75	.82	.78
39	.11	.23	.16	.25
40	.69	.57	.72	.60
41	.42	.47	.47	.50
42	.60	.54	.63	.57
43	.36	.36	.10	.39
44	.78	.66	.83	.69
45	.36	.26	.31	.28
46	.56	.56	.66	.59
47	.58	.68	.76	.71
48	.60	.66	.72	.69
49	.47	.47	.54	.50
50	.20	.23	.28	.25
51	.53	.57	.64	.60
52	.38	.47	.49	.50
53	.64	.64	.78	.67

54	.49	.52	.59	.54
55	.27	.34	.28	.37
56	.47	.63	.73	.66
57	.18	.23	.32	.25
58	.47	.61	.69	.64
59	.24	.40	.50	.43

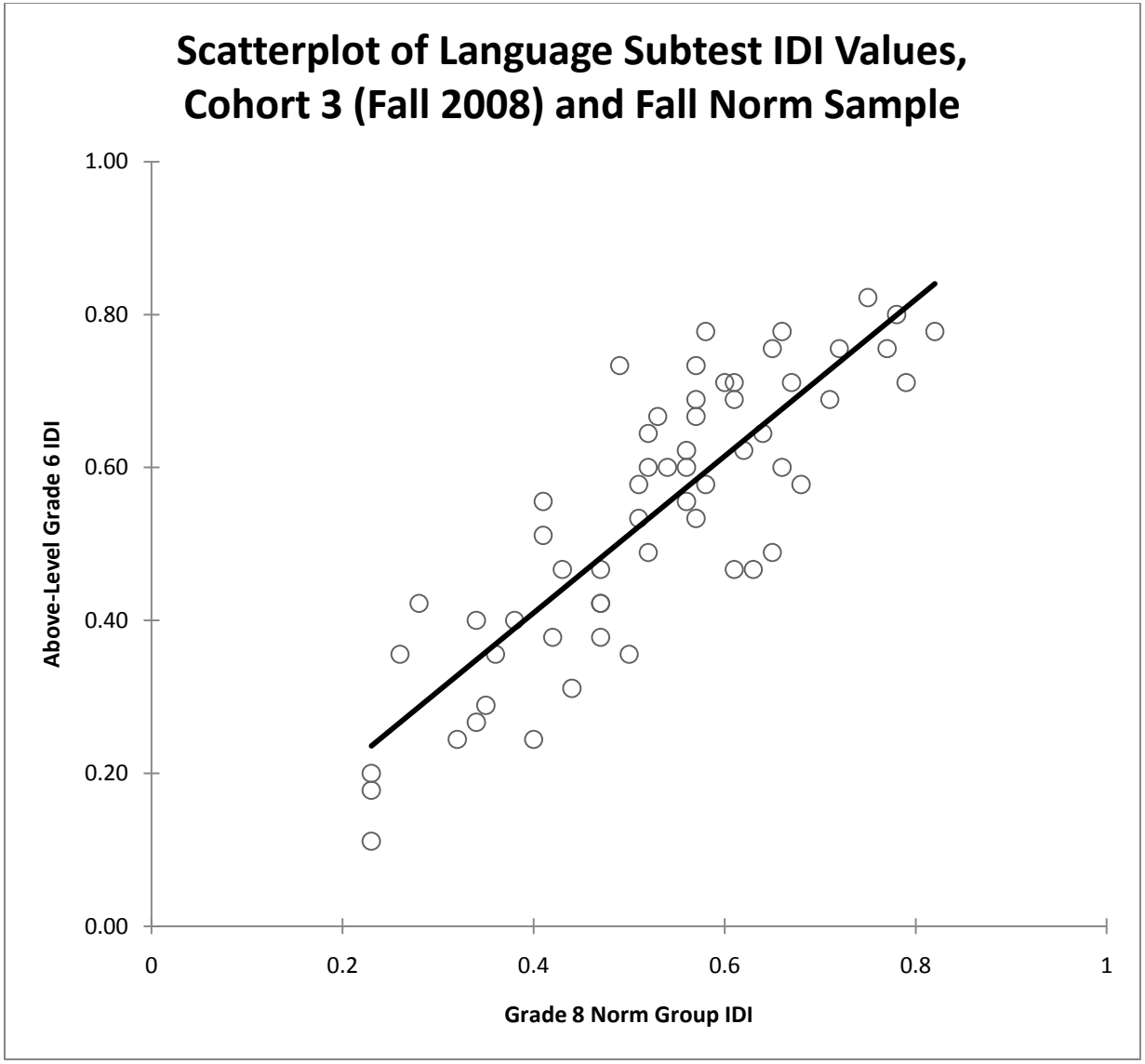

Figure A3 Scatterplot of language subtest IDI values, Cohort 3 (Fall 2008) and Fall Norm Sample. IDI values correlation is $r = .846$ ($p < .001$).

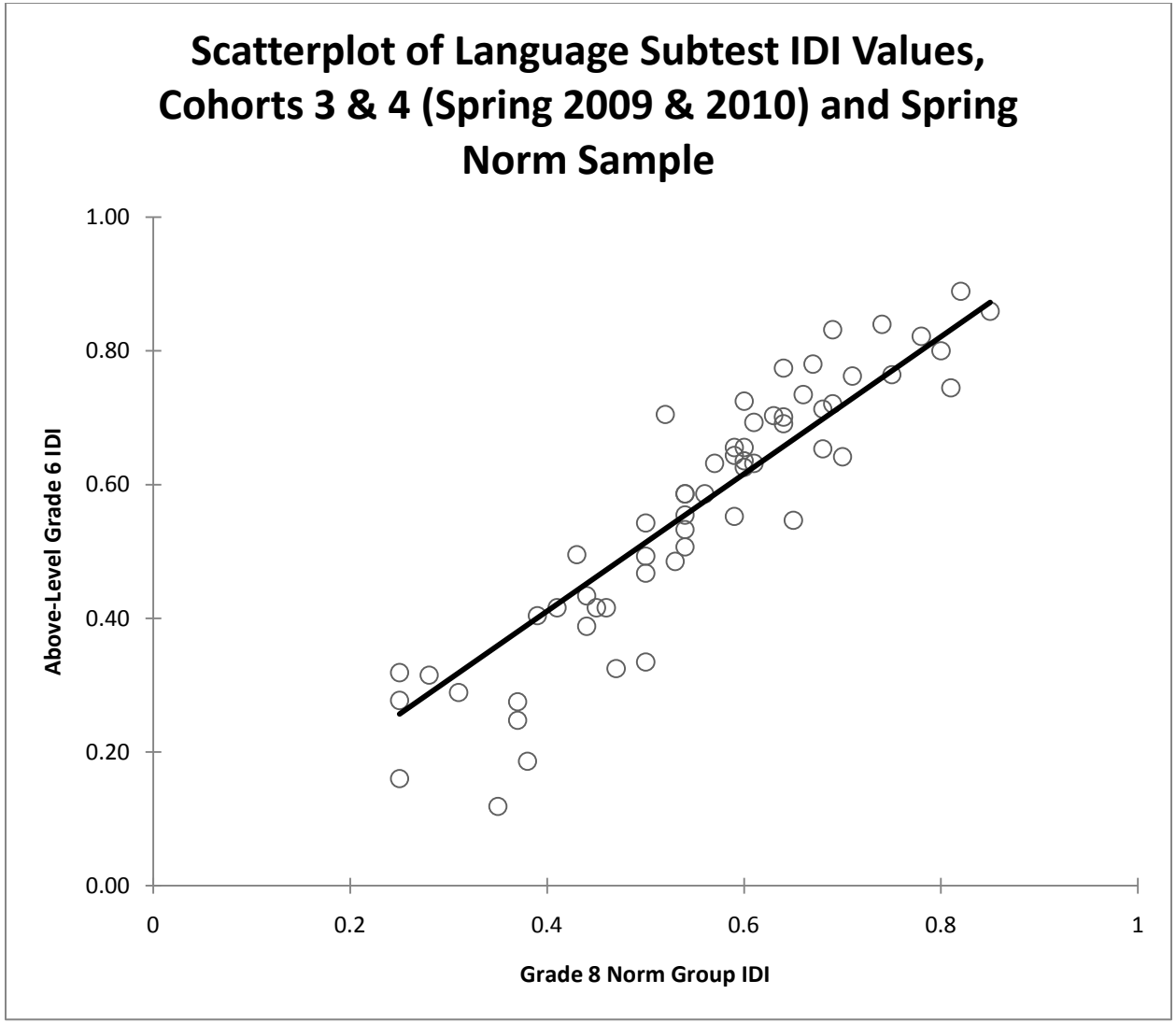

Figure A4 Scatterplot of language subtest IDI values, Cohorts 3 & 4 (Spring 2009 & 2010) and Spring Norm Sample. IDI values correlation is $r = .920$ ($p < .001$).

Table A3
ITBS Math Subtest Item Difficulty Indexes, Gifted Grade 6 and National Grade 8 Norm Groups

Item #	Fall Gifted IDI	Fall Norm IDI	Spring Gifted IDI	Spring Norm IDI
1	.60	.55	.54	.59
2	.40	.30	.44	.33
3	.96	.83	.96	.86
4	.96	.67	.93	.70
5	.71	.65	.83	.68
6	.44	.44	.55	.47
7	.56	.50	.70	.53
8	.62	.53	.63	.56
9	.49	.49	.49	.52
10	.27	.40	.34	.43
11	.84	.72	.92	.74
12	.58	.76	.82	.78
13	.71	.67	.83	.70
14	.53	.59	.68	.63
15	.29	.33	.48	.36
16	.80	.68	.88	.71
17	.42	.41	.50	.44
18	.36	.49	.45	.52
19	.49	.34	.56	.37
20	.20	.38	.22	.41
21	.42	.45	.45	.49
22	.13	.29	.24	.32
23	.44	.43	.37	.46
24	.18	.35	.33	.38
25	.36	.32	41	.35
26	.20	.32	.30	.35
27	.13	.32	.21	.35
28	.20	.37	.26	.40
29	.13	.30	.30	.35
30	.47	.42	.45	.46
31	.78	.51	.71	.55
32	.67	.38	.60	.41
33	.24	.31	.36	.34
34	.60	.41	.47	.45
35	.31	.49	.48	.53
36	.33	.35	.15	.42
37	.27	.56	.44	.62
38	.76	.71	.76	.78
39	.18	.30	.13	.37
40	.49	.64	.62	.71
41	.53	.53	.56	.67
42	.38	.38	.47	.53
43	.09	.09	.17	.48
44	.31	.31	.27	.52
45	.16	.16	.11	.50
46	.07	.07	.04	.39

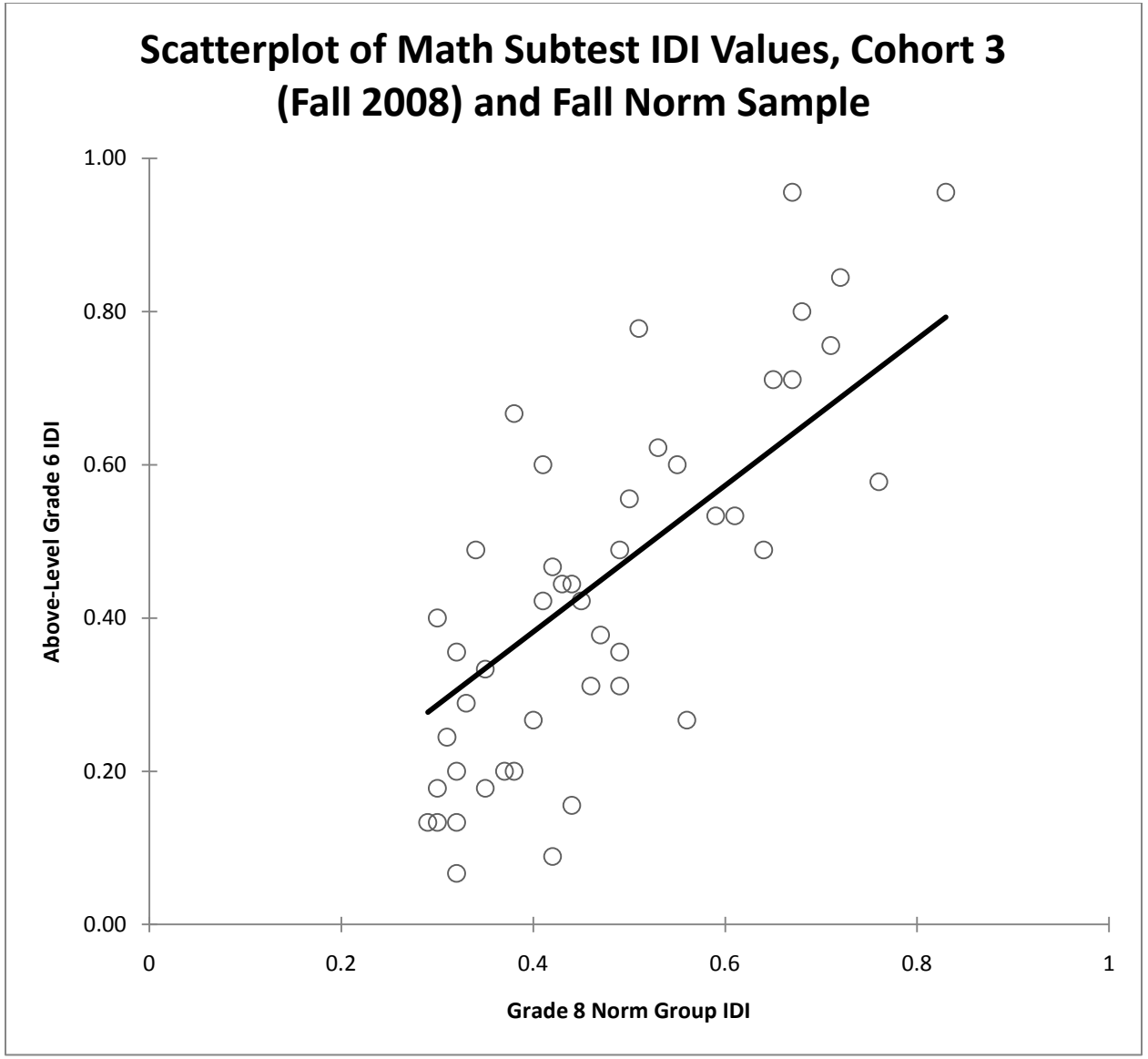

Figure A5 Scatterplot of math subtest IDI values, Cohort 3 (Fall 2008) and Fall Norm Sample. IDI values correlation is r = .786 (p < .001).

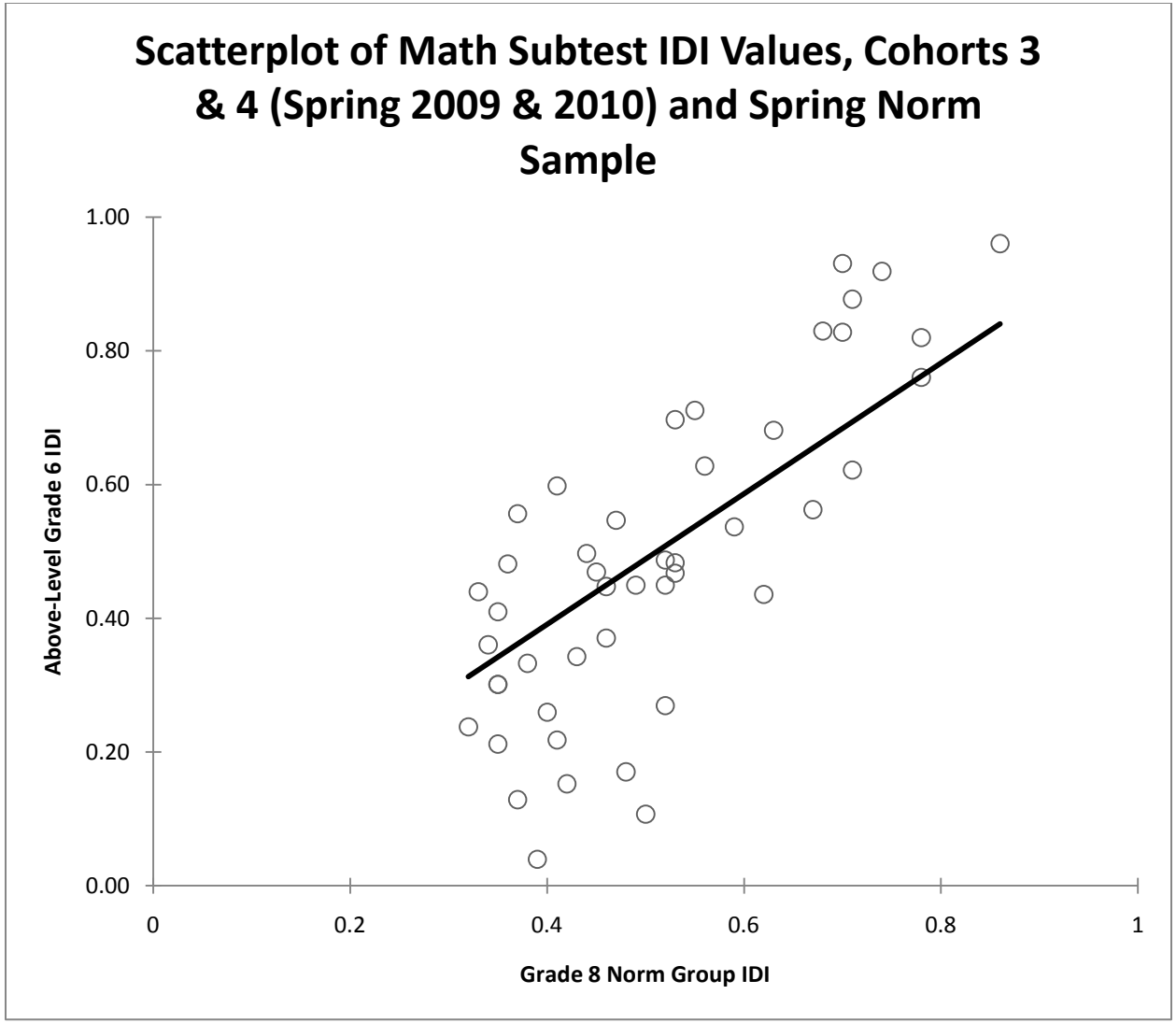

Figure A6 Scatterplot of math subtest IDI values, Cohorts 3 & 4 (Spring 2009 & 2010) and Spring Norm Sample. IDI values correlation is $r = .787$ ($p < .001$).

Table A4
ITED Vocabulary Subtest Item Difficulty Indexes, Gifted Grade 7 and National Grade 9 Norm Groups

Item #	Fall Gifted IDI	Fall Norm IDI	Spring Gifted IDI	Spring Norm IDI
1	.92	.79	.94	.83
2	.72	.72	.84	.46
3	.83	.53	.81	.58
4	.68	.51	.73	.56
5	.49	.55	.61	.60
6	.57	.70	.80	.75
7	.70	.56	.70	.61
8	.40	.44	.55	.49
9	.75	.68	.80	.73
10	.51	.57	.66	.62
11	.72	.66	.75	.71
12	.64	.46	.67	.51
13	.42	.49	.52	.65
14	.38	.41	.46	.46
15	.87	.65	.93	.70
16	.66	.63	.78	.68
17	.64	.3	.83	.68
18	.79	.77	.84	.81
19	.85	.63	.84	.68
20	.53	.50	.69	.55
21	.83	.73	.84	.77
22	.87	.75	.86	.79
23	.94	.75	.94	.79
24	.28	.44	.36	.48
25	.79	.57	.86	.62
26	.68	.57	.66	.62
27	.51	.57	.63	.62
28	.49	.50	.50	.55
29	.57	.43	.52	.47
30	.17	.44	.26	.48
31	.58	.46	.70	.51
32	.47	.48	.58	.53
33	.43	.49	.58	.54
34	.55	.58	.79	.62
35	.28	.38	.42	.42
36	.42	.31	.37	.35
37	.53	.34	.47	.38
38	.34	.37	.36	.41
39	.13	.32	.14	.36
40	.40	.44	.44	.49

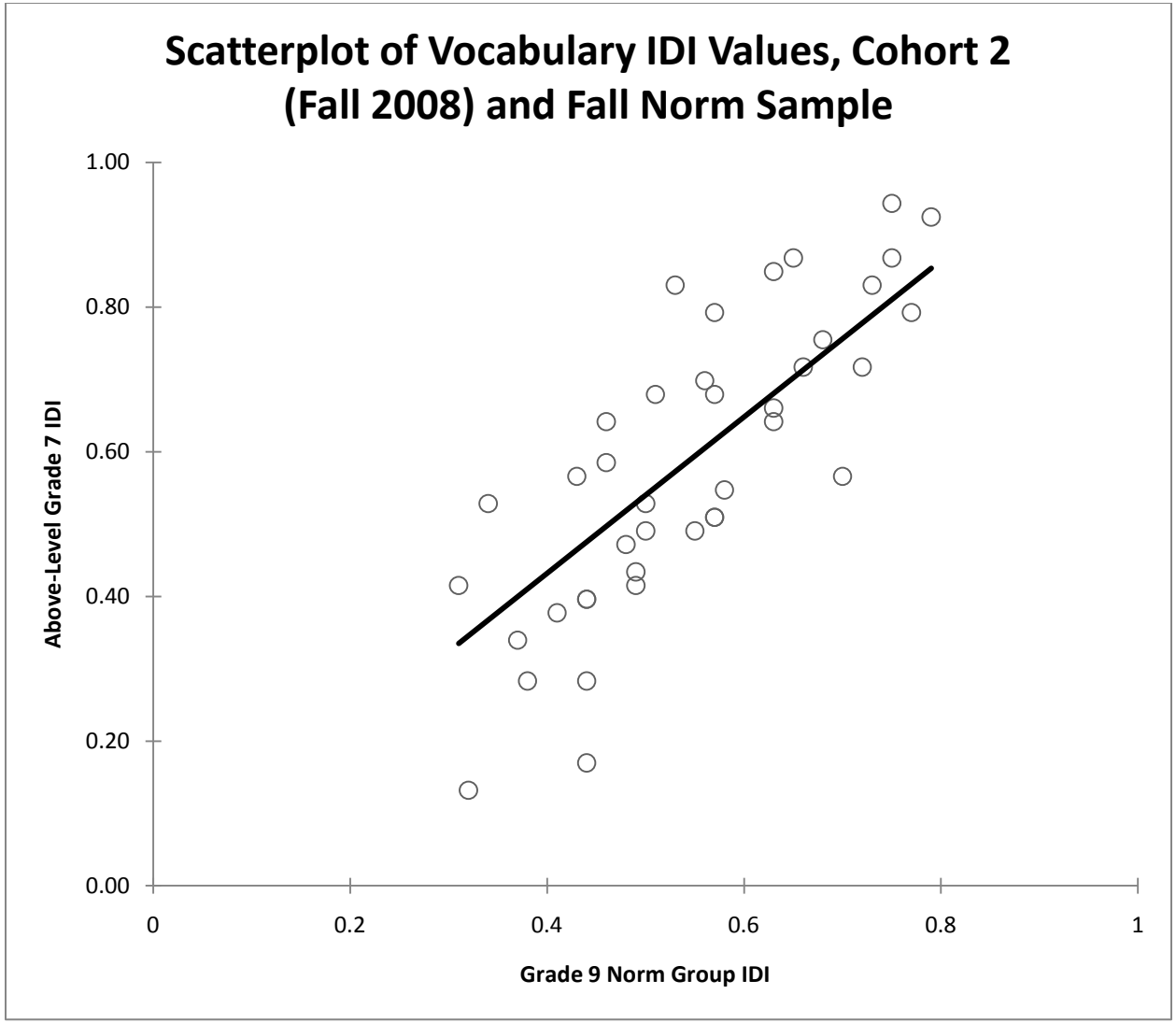

Figure A7 Scatterplot of vocabulary subtest IDI values, Cohort 2 (Fall 2008) and Fall Norm Sample. IDI values correlation is $r = .806$ ($p < .001$).

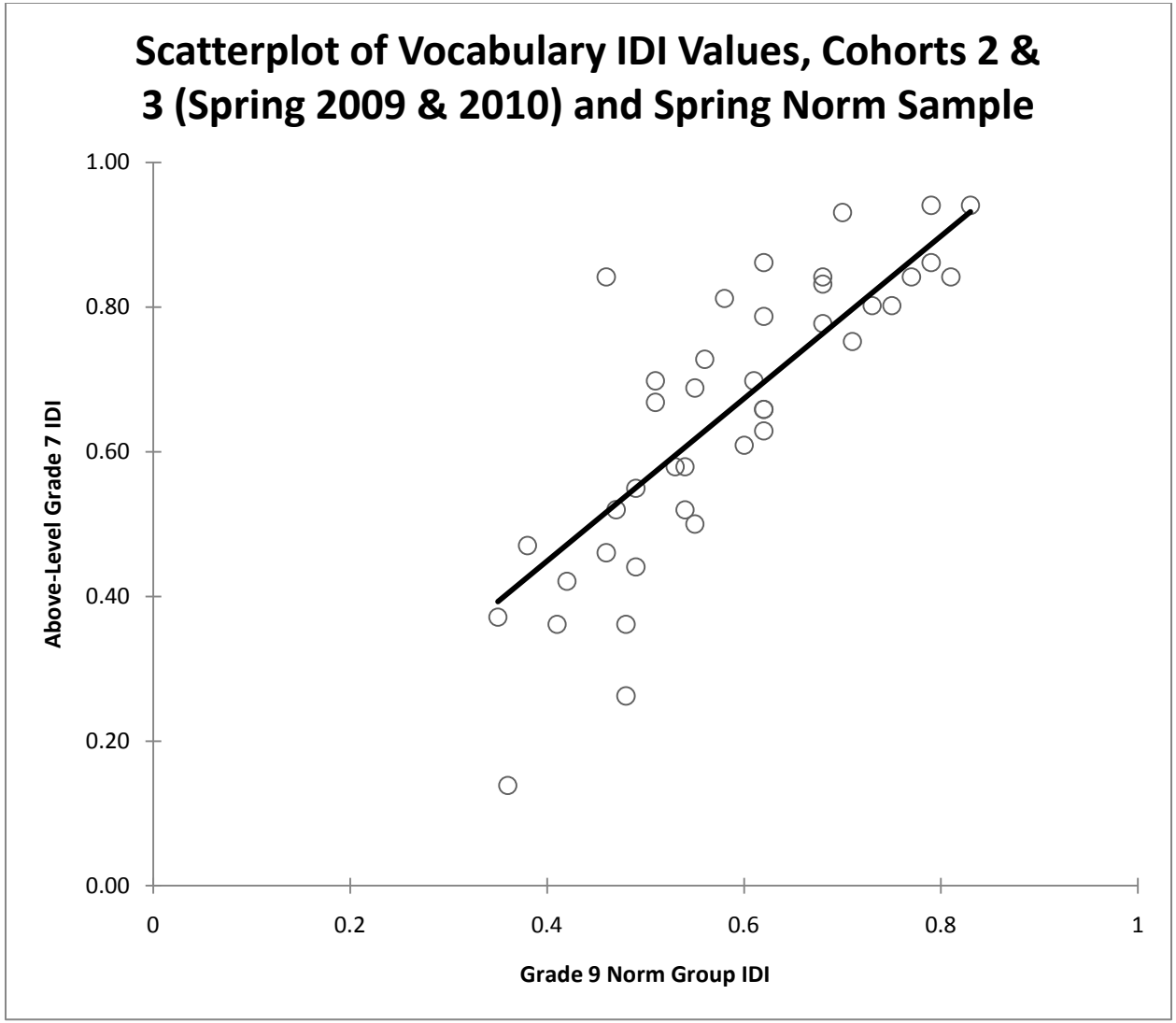

Figure A8 Scatterplot of vocabulary subtest IDI values, Cohorts 2 & 3 (Spring 2009 & 2010) and Spring Norm Sample. IDI values correlation is $r = .832$ ($p < .001$).

Table A5
ITED Reading Comprehension Subtest Item Difficulty Indexes, Gifted Grade 7 and National Grade 9 Norm Groups

Item #	Fall Gifted IDI	Fall Norm IDI	Spring Gifted IDI	Spring Norm IDI
1	.83	.69	.82	.72
2	.91	.69	.85	.72
3	.81	.67	.90	.70
4	.60	.58	.74	.61
5	.77	.66	.75	.69
6	.81	.65	.83	.68
7	.91	.71	.88	.74
8	.77	.63	.87	.67
9	.83	.66	.86	.69
10	.83	.71	.89	.74
11	.85	.63	.87	.66
12	.83	.72	.86	.75
13	.96	.76	.96	.79
14	.87	.70	.95	.73
15	.60	.52	.71	.56
16	.55	.51	.76	.54
17	.75	.67	.82	.70
18	.55	.60	.48	.63
19	.47	.47	.59	.50
20	.66	.60	.83	.63
21	.81	.68	.87	.71
22	.85	.64	.83	.67
23	.64	.52	.62	.55
24	.66	.47	.57	.50
25	.91	.73	.99	.76
26	.53	.41	.56	.44
27	.64	.51	.71	.54
28	.42	.46	.66	.49
29	.71	.61	.80	.64
30	.74	.56	.79	.59
31	.66	.49	.70	.52
32	.77	.61	.80	.64
33	.45	.50	.59	.53
34	.55	.52	.59	.55
35	.38	.34	.45	.37
36	.55	.49	.61	.3
37	.40	.43	.54	.46
38	.40	.35	.46	.38
39	.60	.51	.56	.55
40	.49	.42	.53	.45
41	.49	.40	.50	.43
42	.47	.45	.41	.48
43	.42	.45	.48	.48
44	.49	.35	.40	.38

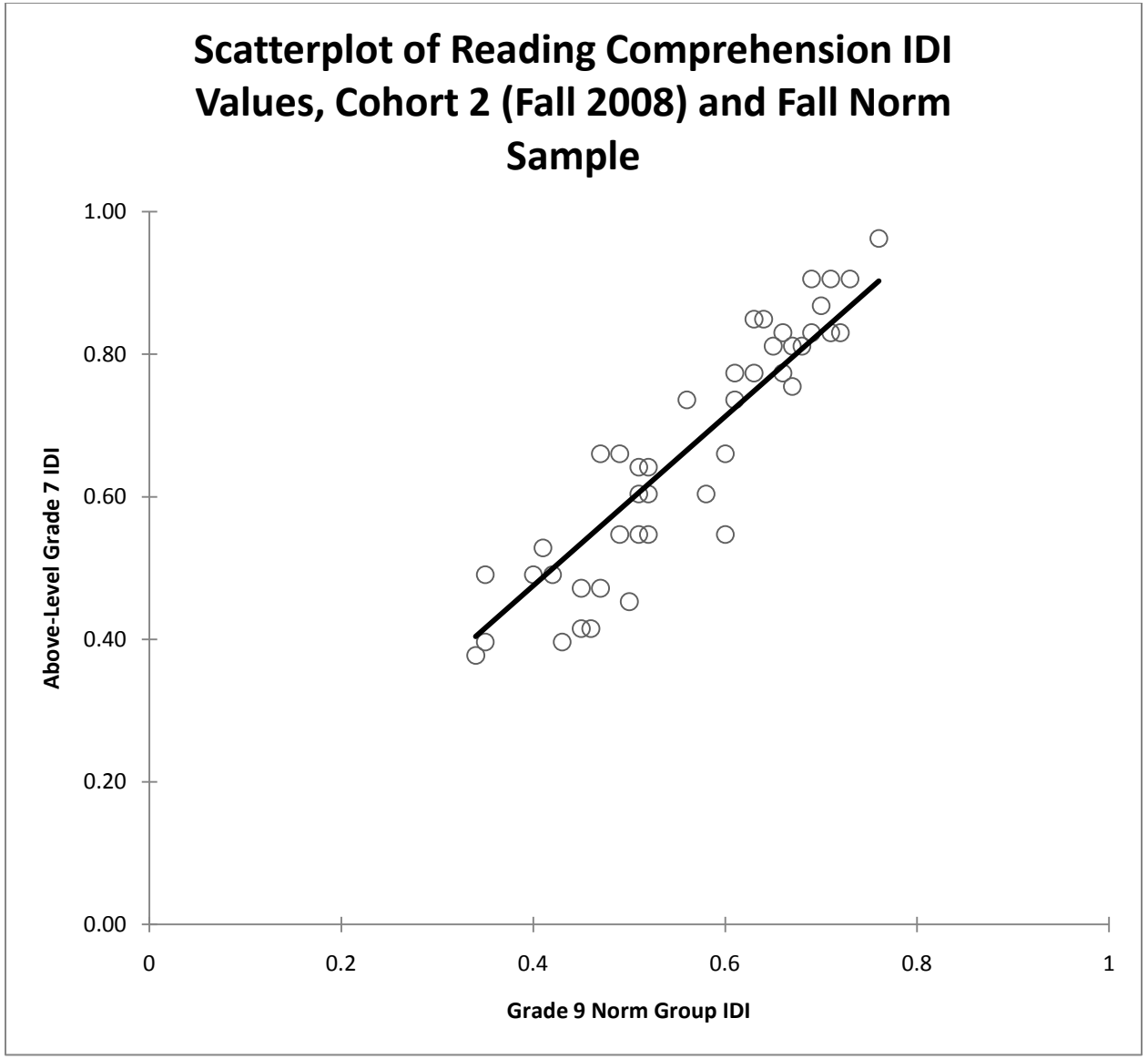

Figure A9 Scatterplot of reading comprehension subtest IDI values, Cohort 2 (Fall 2008) and Fall Norm Sample. IDI values correlation is $r = .923$ ($p < .001$).

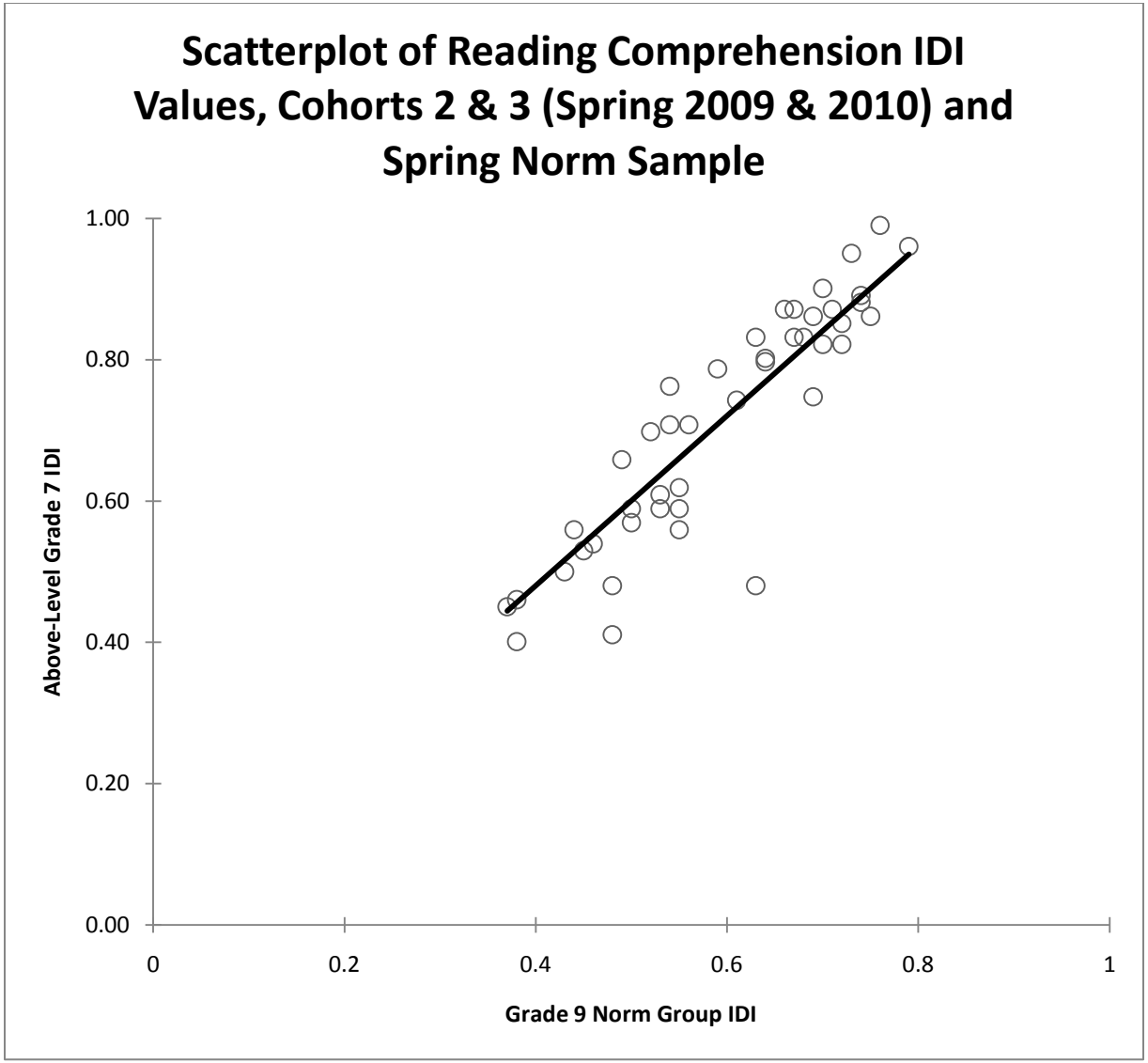

Figure A10 Scatterplot of reading comprehension subtest IDI values, Cohorts 2 & 3 (Spring 2002 & 2010) and Spring Norm Sample. IDI values correlation is $r = .906$ ($p < .001$).

Table A6
ITED Spelling Subtest Item Difficulty Indexes, Gifted Grade 7 and National Grade 9 Norm Groups

Item #	Fall Gifted IDI	Fall Norm IDI	Spring Gifted IDI	Spring Norm IDI
1	.81	.43	.84	.44
2	.85	.38	.79	.39
3	.70	.61	.77	.64
4	.49	.46	.61	.49
5	.55	.67	.66	.70
6	.58	.66	.58	.69
7	.81	.53	.85	.56
8	.64	.58	.56	.61
9	.83	.54	.87	.57
10	.64	.60	.59	.63
11	.64	.57	.81	.59
12	.47	.56	.45	.59
13	.36	.55	.33	.59
14	.57	.39	.68	.42
15	.74	.41	.67	.44
16	.36	.68	.35	.71
17	.15	.38	.29	.40
18	.42	.34	.47	.36
19	.45	.58	.56	.61
20	.28	.40	.38	.42
21	.28	.58	.27	.61
22	.42	.69	.57	.71
23	.49	.25	.44	.23
24	.23	.33	.25	.35
25	.49	.40	.48	.42
26	.62	.49	.60	.52
27	.57	.64	.57	.66
28	.34	.55	.40	.57
29	.38	.49	.32	.51
30	.13	.43	.16	.45

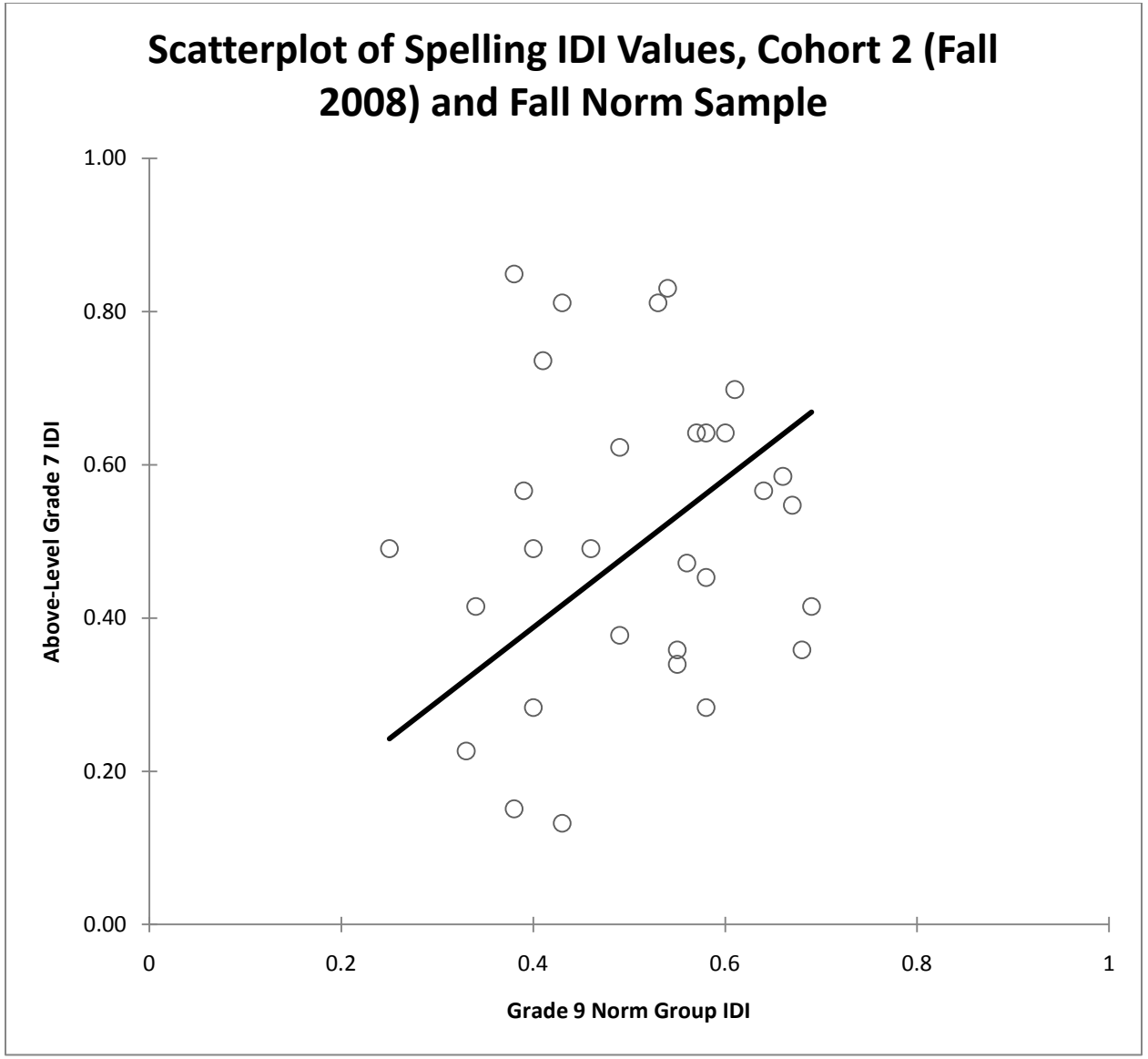

Figure A11 Scatterplot of spelling subtest IDI values, Cohort 2 (Fall 2008) and Fall Norm Sample. IDI values correlation is r = .134 (p = .480).

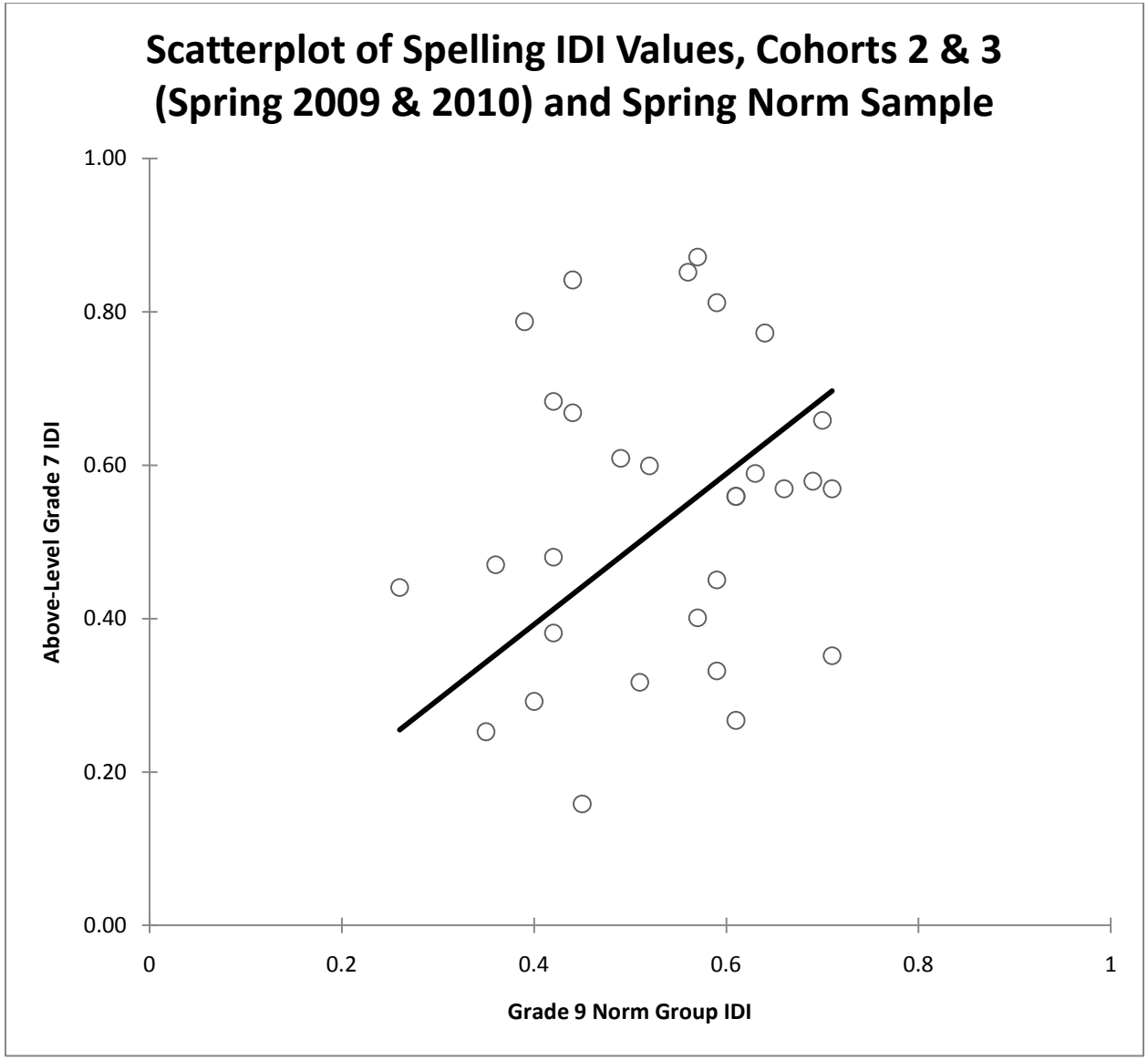

Figure A12 Scatterplot of spelling subtest IDI values, Cohorts 2 & 3 (Spring 2009 & 2010) and Spring Norm Sample. IDI values correlation is r = .171 (p = .366).

Table A7
ITED Revising Written Materials Subtest Item Difficulty Indexes, Gifted Grade 7 and National Grade 9 Norm Groups

Item #	Fall Gifted IDI	Fall Norm IDI	Spring Gifted IDI	Spring Norm IDI
1	.75	.74	.76	.77
2	.30	.75	.47	.78
3	.64	.62	.62	.65
4	.60	.45	.76	.48
5	.72	.48	.70	.51
6	.51	.56	.47	.59
7	.32	.70	.48	.73
8	.87	.50	.82	.53
9	.25	.77	.35	.80
10	.68	.50	.70	.53
11	.62	.54	.54	.57
12	.70	.41	.84	.44
13	.70	.40	.79	.43
14	.75	.42	.66	.45
15	.57	.65	.63	.68
16	.64	.38	.84	.41
17	.74	.34	.71	.37
18	.49	.43	.51	.46
19	.66	.56	.75	.59
20	.57	.40	.69	.43
21	.62	.27	.70	.30
22	.62	.57	65	.60
23	.57	.44	.59	.47
24	.47	.33	.49	.36
25	.68	.51	.77	.54
26	.36	.51	.57	.54
27	.77	.54	.80	.57
28	.49	.43	.43	.46
29	.66	.31	.72	.34
30	.51	.32	.63	.35
31	.96	.60	.90	.63
32	.87	.43	.87	.45
33	.53	.44	.69	.46
34	.77	.51	.86	.54
35	.70	.53	.63	.56
36	.77	.61	.72	.64
37	.74	.40	.76	.41
38	.70	.72	.77	.75
39	.77	.61	.75	.64
40	.43	.63	.53	.66
41	.51	.43	.55	.46
42	.85	.33	.86	.34
43	.60	.51	.44	.54
44	.40	.60	.29	.63
45	.70	.48	.71	.50
46	.57	.64	.44	.67
47	.66	.57	.69	.60
48	.75	.39	.72	.47
49	.26	.54	.31	.58
50	.32	.57	.47	.59
51	.30	.54	.44	.56
52	.55	.51	.60	.54

53	.34	.35	.39	.36
54	.51	.43	.47	.44
55	.58	.53	.50	.56
56	.25	.44	.33	.47

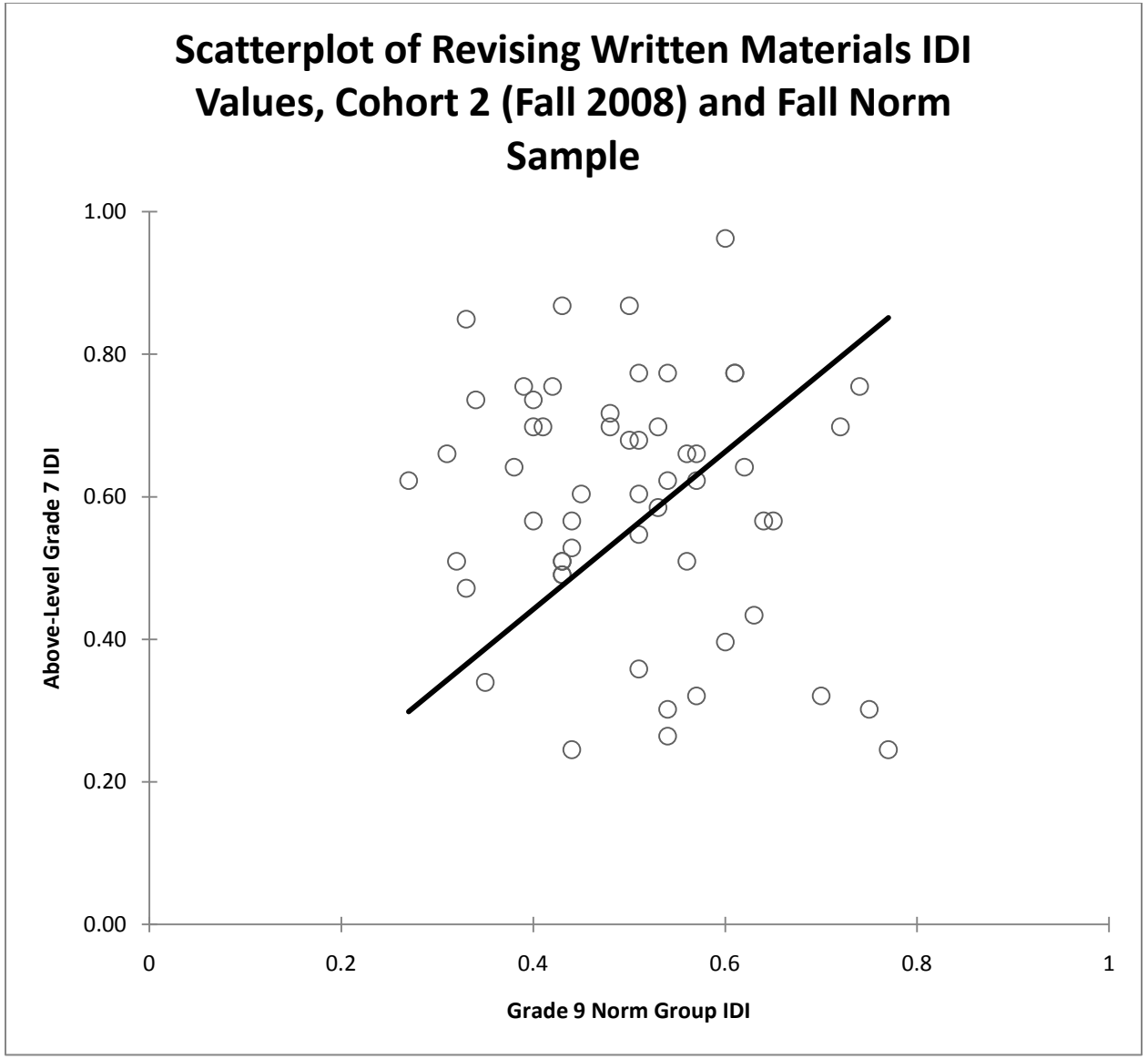

Figure A13 Scatterplot of revising written materials subtest IDI values, Cohort 2 (Fall 2008) and Fall Norm Sample. IDI values correlation is r = -.182 (p = .179).

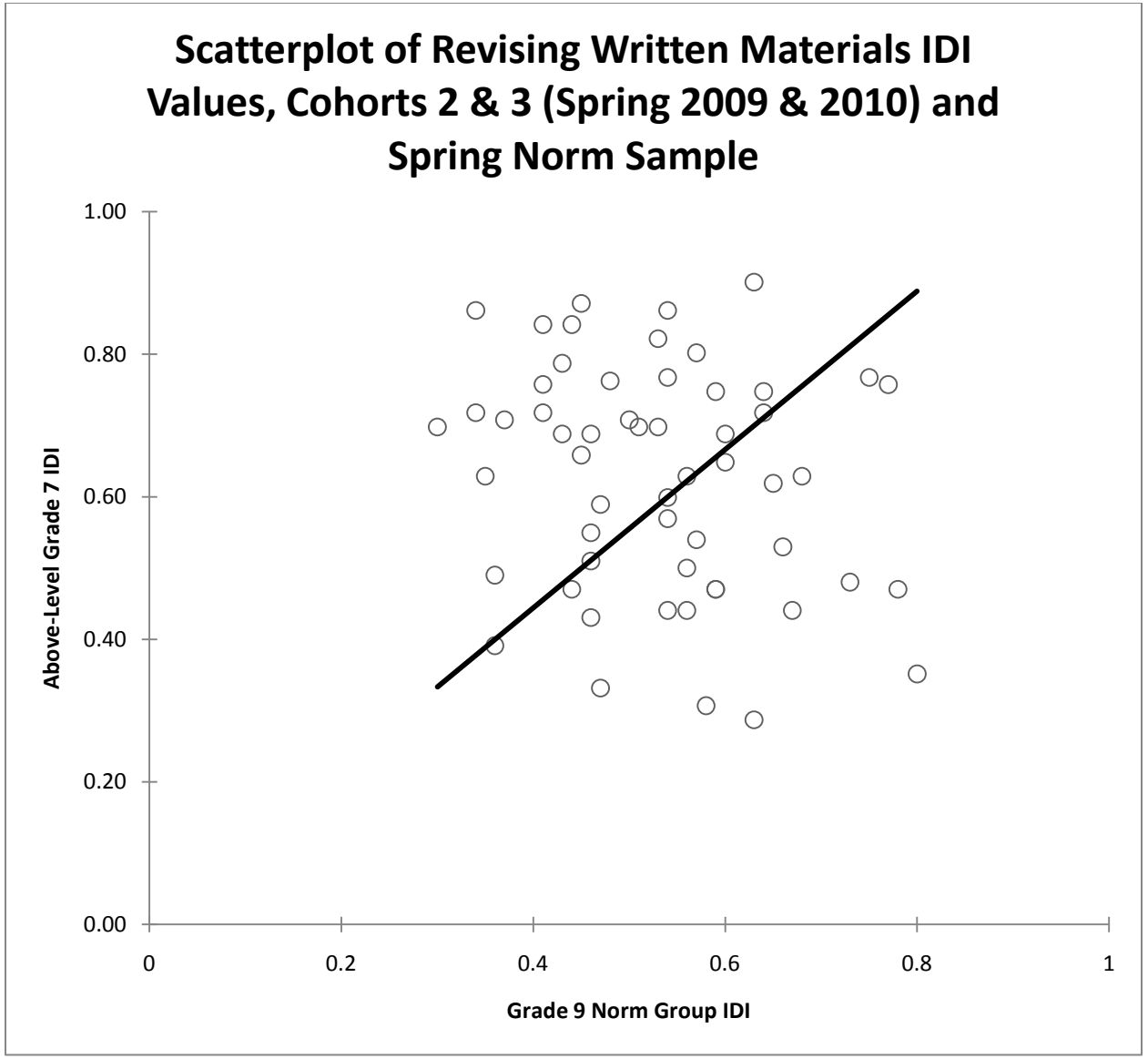

Figure A14 Scatterplot of revising written materials subtest IDI values, Cohorts 2 & 3 (Spring 2009 & 2010) and Spring Norm Sample. IDI values correlation is r = -.216 (p = .101).

Table A8
ITED Mathematics Concepts & Problem Solving Subtest Item Difficulty Indexes, Gifted Grade 7 and National Grade 9 Norm Groups

Item #	Fall Gifted IDI	Fall Norm IDI	Spring Gifted IDI	Spring Norm IDI
1	.83	.66	.90	.69
2	.87	.66	.95	.69
3	.72	.48	.80	.52
4	.55	.53	.78	.57
5	.72	.34	.61	.38
6	.21	.29	.31	.33
7	.81	.53	.87	.57
8	.70	.54	.78	.58
9	.66	.56	.71	.60
10	.38	.41	.52	.45
11	.40	.41	.44	.45
12	.55	.42	.69	.46
13	.58	.56	.69	.60
14	.25	.30	.34	.34
15	.28	.22	.38	.23
16	.83	.63	.79	.67
17	.83	.51	.89	.55
18	.55	.52	.69	.56
19	.38	.36	.57	.40
20	.11	.41	.46	.45
21	.43	.51	.64	.55
22	.43	.26	.48	.29
23	.64	.49	.81	.53
24	.68	.63	.70	.66
25	.42	.37	.49	.40
26	.51	.56	.59	.59
27	.23	.34	.42	.37
28	.47	.49	.57	.53
29	.43	.49	.72	.52
30	.43	.55	.70	.58
31	.28	.34	.40	.37
32	.30	.39	.35	.42
33	.15	.22	.11	.24
34	.23	.22	.20	.23
35	.15	.22	.17	.23
36	.25	.36	.33	.39
37	.43	.48	.51	.51
38	.36	.38	.37	.41
39	.42	.42	.60	.45
40	.26	.23	.28	.25

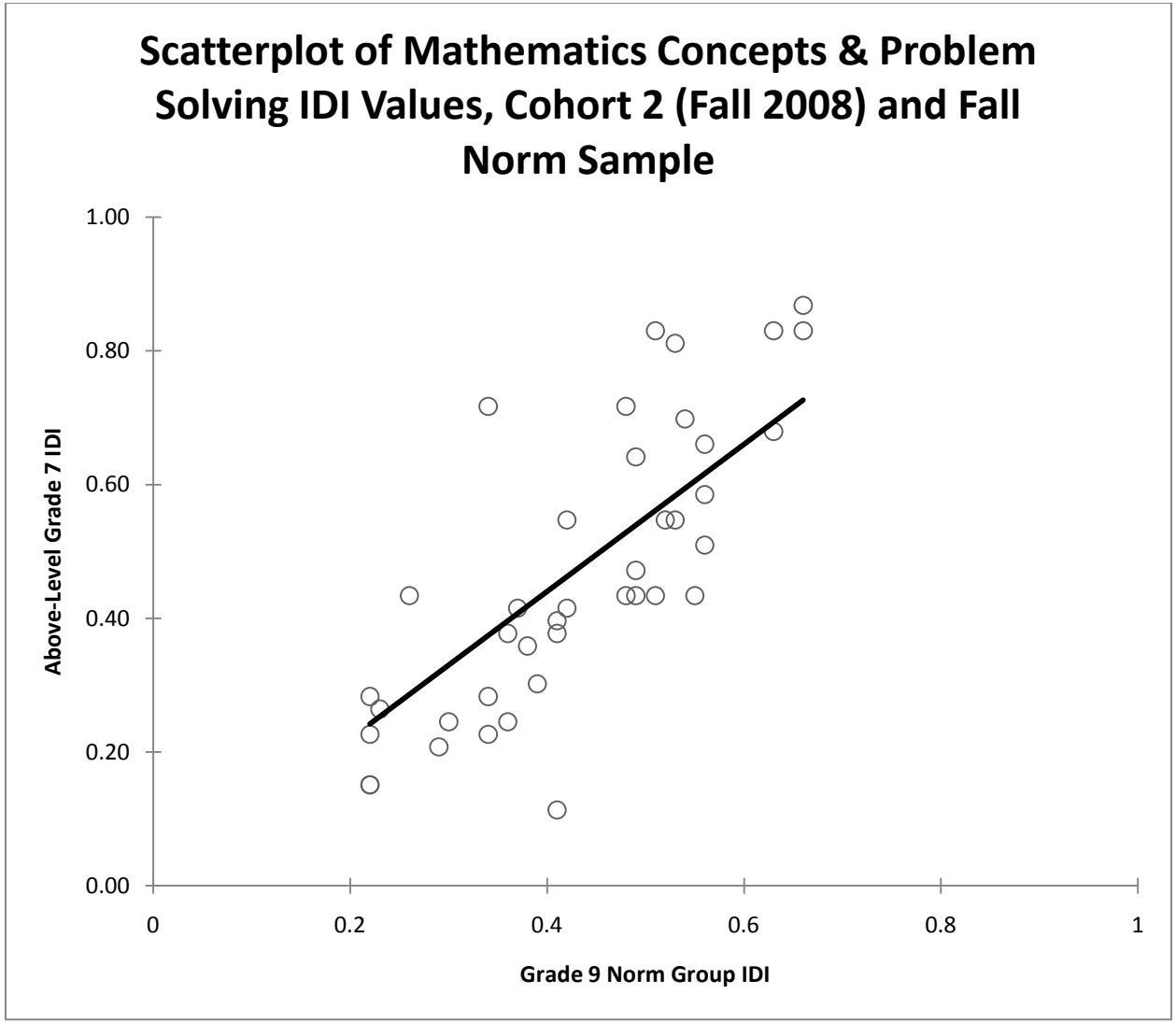

Figure A15 Scatterplot of mathematics concepts & problem solving subtest IDI values, Cohort 2 (Fall 2008) and Fall Norm Sample. IDI values correlation is $r = .792$ ($p < .001$).

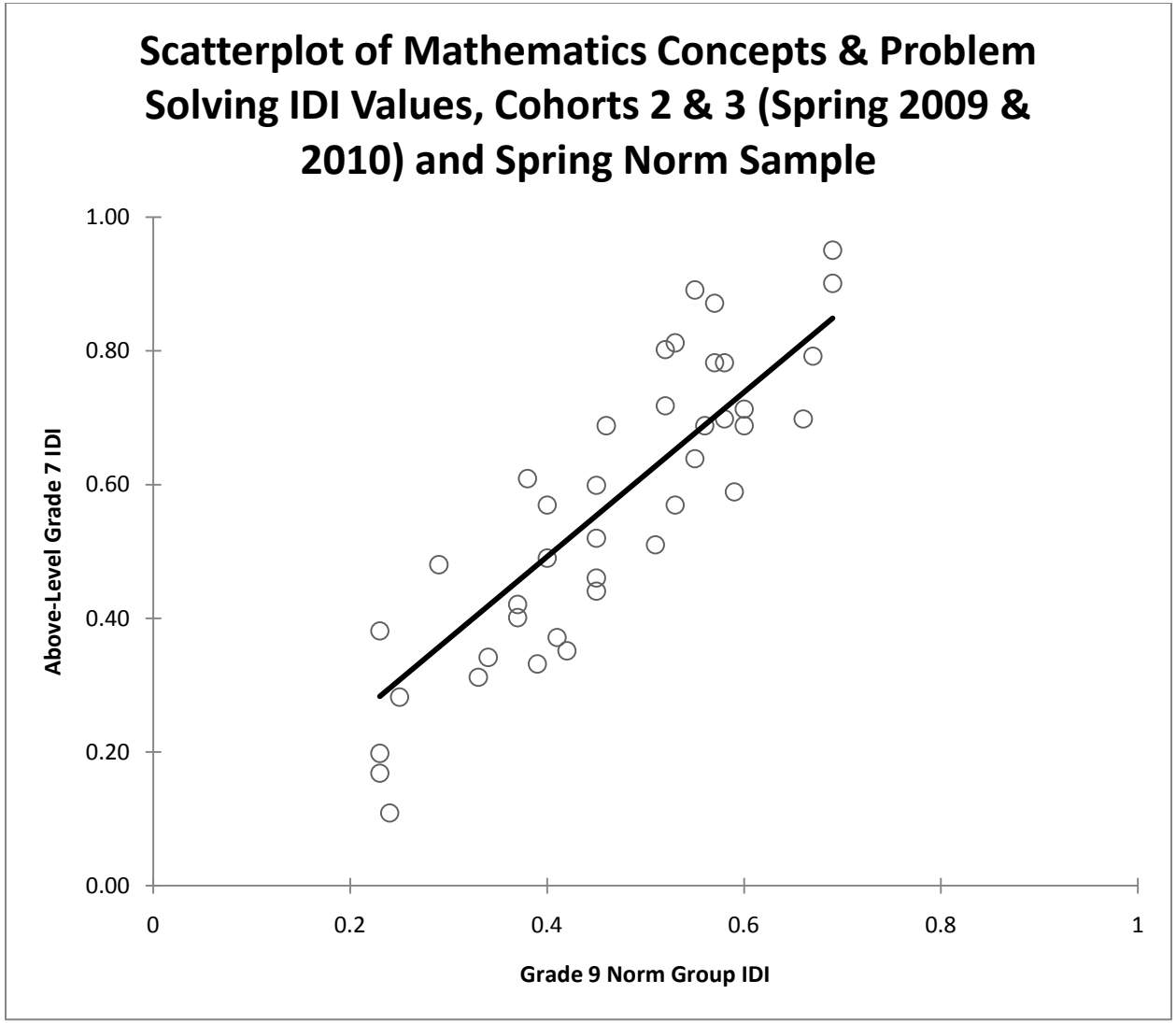

Figure A16 Scatterplot of mathematics concepts & problem solving subtest IDI values, Cohorts 2 & 3 (Spring 2009 & 2010) and Spring Norm Sample. IDI values correlation is $r = .877$ ($p < .001$).

Table A9
ITED Mathematics Computation Subtest Item Difficulty Indexes, Gifted Grade 7 and National Grade 9 Norm Groups

Item #	Fall Gifted IDI	Fall Norm IDI	Spring Gifted IDI	Spring Norm IDI
1	.92	.88	.89	.92
2	.18	.75	.83	.78
3	.60	.65	.67	.68
4	.81	.66	.76	.69
5	.68	.65	.58	.69
6	.62	.50	.73	.53
7	.79	.53	.74	.56
8	.58	.45	.50	.48
9	.34	.45	.44	.48
10	.26	.39	.34	.42
11	.26	.41	.34	.44
12	.36	.64	.57	.68
13	.60	.65	.66	.68
14	.15	.25	.09	.25
15	.49	.60	.32	.63
16	.42	.57	.45	.60
17	.13	.26	.20	.29
18	.36	.53	.31	.56
19	.19	.31	.16	.34
20	.23	.32	.24	.35
21	.30	.36	.36	.39
22	.15	.40	.27	.43
23	.11	.32	.16	.35
24	.11	.26	.09	.29
25	.28	.35	.20	.38
26	.28	.34	.36	.37
27	.06	.19	.09	.19
28	.06	.20	.12	.20
29	.06	.26	.07	.29
30	.04	.18	.05	.15

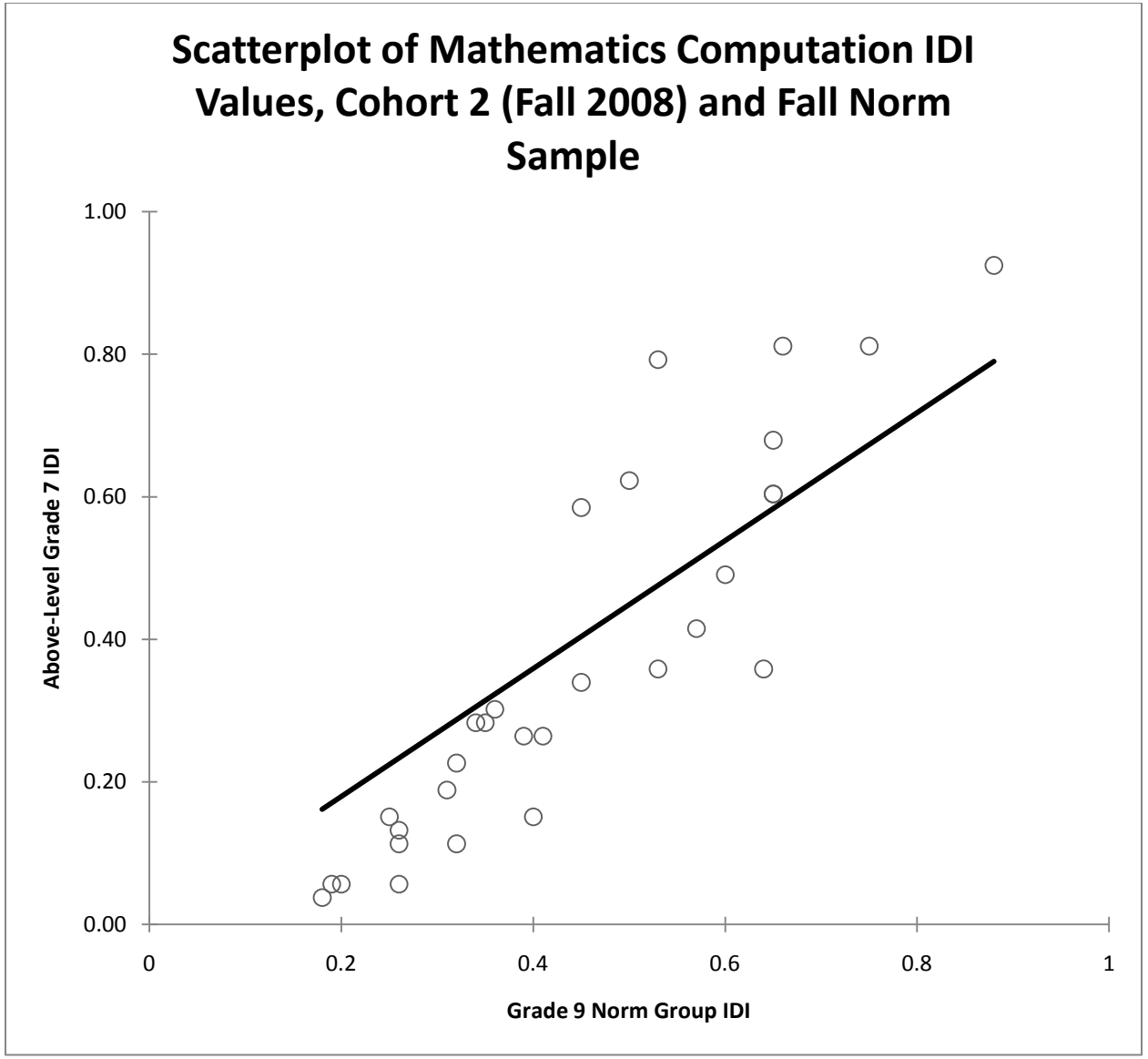

Figure A17 Scatterplot of mathematics computation subtest IDI values, Cohort 2 (Fall 2008) and Fall Norm Sample. IDI values correlation is $r = .904$ ($p < .001$).

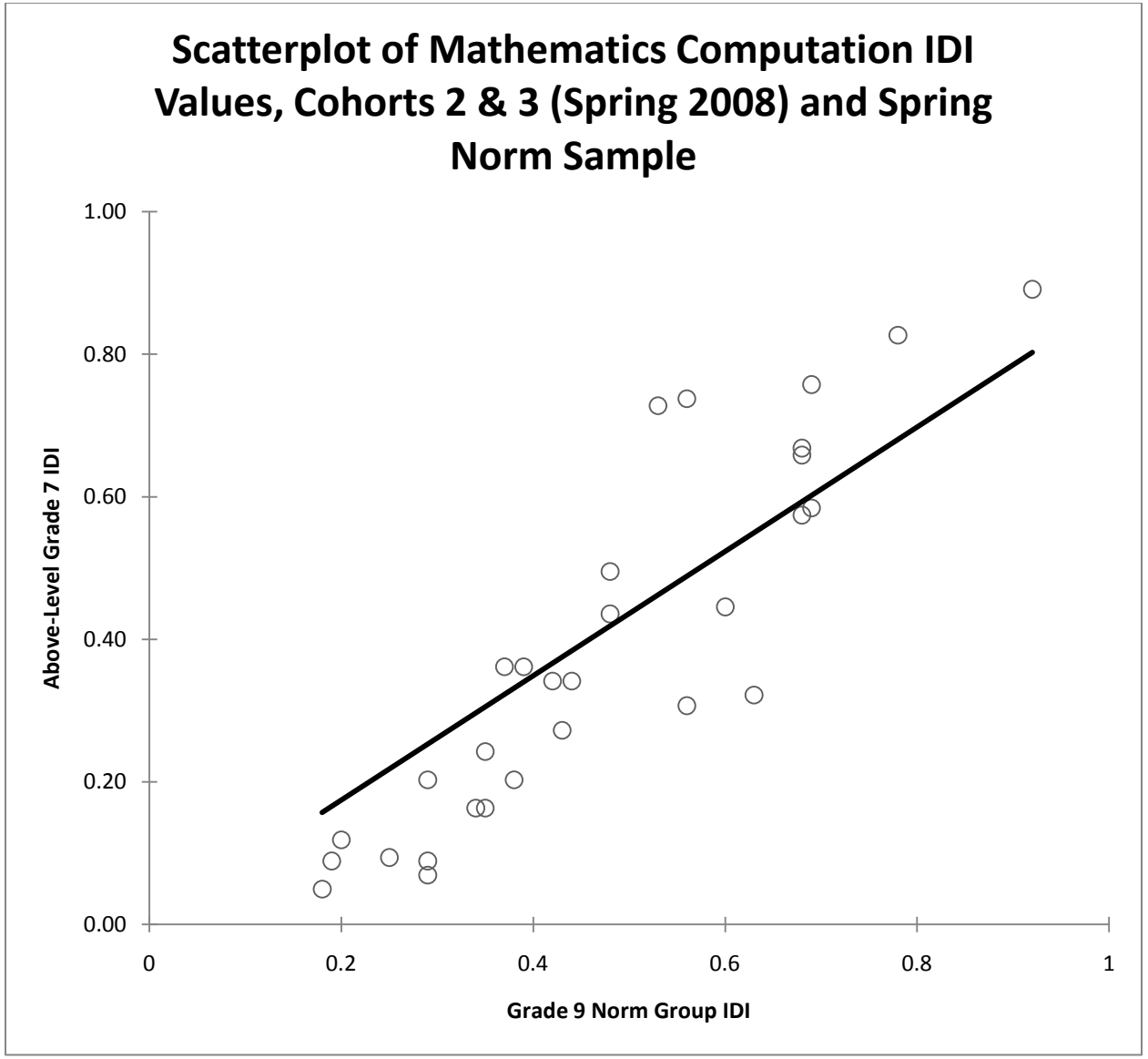

Figure A18 Scatterplot of mathematics computation subtest IDI values, Cohorts 2 & 3 (Spring 2009 & 2010) and Spring Norm Sample. IDI values correlation is $r = .903$ ($p < .001$).

Table A10
ITED Vocabulary Subtest Item Difficulty Indexes, Gifted Grade 8 and National Grade 10 Norm Groups

Item #	Fall Gifted IDI	Fall Norm IDI	Spring Gifted IDI	Spring Norm IDI
1	.90	.78	.87	.81
2	.92	.80	.95	.82
3	.95	.80	.95	.82
4	.38	.50	.44	.54
5	.85	.63	.87	.65
6	.72	.66	.69	.72
7	.69	.67	.79	.74
8	.74	.57	.69	.60
9	.62	.59	.76	.67
10	.85	.69	.85	.78
11	.59	.57	.73	.66
12	.56	.54	.61	.55
13	.46	.51	.52	.59
14	.44	.49	.42	.52
15	.38	.45	.45	.51
16	.36	.36	.54	.38
17	.36	.39	.53	.41
18	.49	.47	.47	.56
19	.23	.37	.23	.40
20	.56	.52	.57	.58
21	.82	.80	.88	.82
22	.59	.67	.66	.69
23	.74	.60	.80	.62
24	.77	.71	.86	.73
25	.56	.55	.57	.57
26	.62	.58	.63	.60
27	.77	.63	.73	.65
28	.74	.59	.66	.61
29	.67	.65	.66	.67
30	.72	.58	.67	.60
31	.69	.48	.65	.50
32	.46	.46	.50	.48
33	.62	.63	.70	.65
34	.49	.48	.62	.50
35	.62	.57	.64	.59
36	.44	.54	.58	.56
37	.41	.54	.53	.56
38	.51	.59	.47	.61
39	.54	.59	.57	.61
40	.41	.54	.51	.56

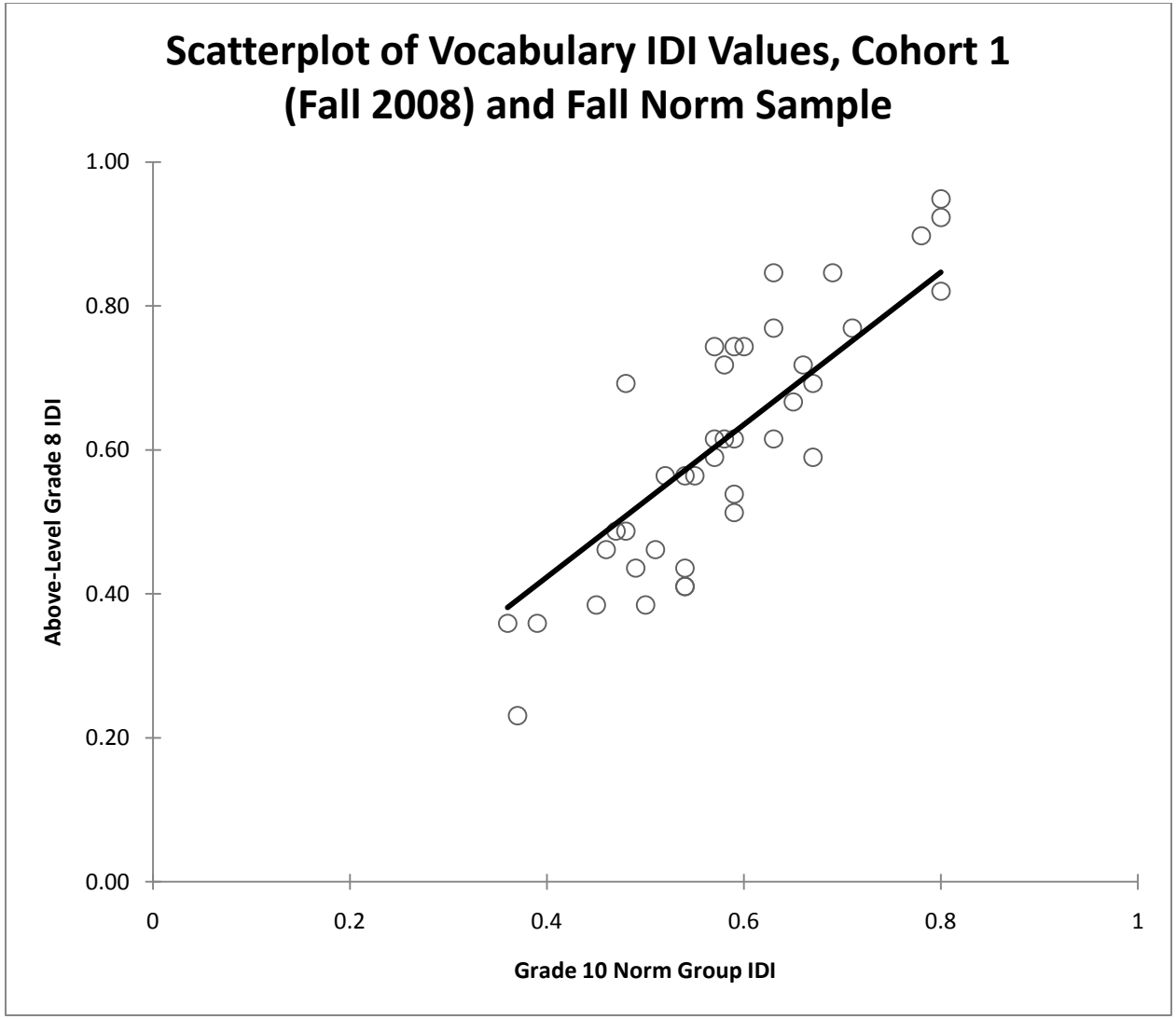

Figure A19 Scatterplot of vocabulary subtest IDI values, Cohort 1 (Fall 2008) and Fall Norm Sample. IDI values correlation is $r = .867$ ($p < .001$).

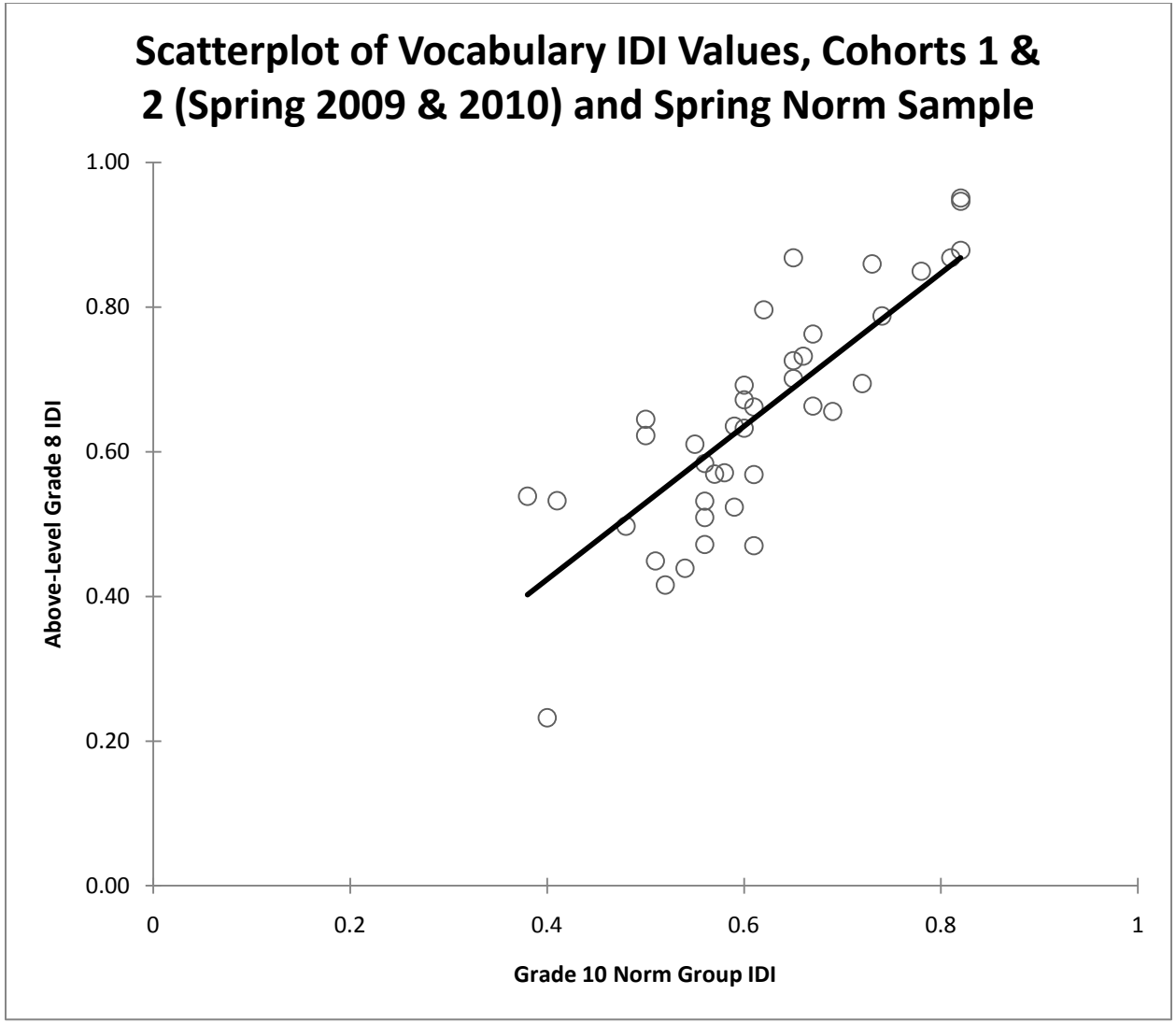

Figure A20 Scatterplot of vocabulary subtest IDI values, Cohorts 1 & 2 (Spring 2009 & 2010) and Spring Norm Sample. IDI values correlation is $r = .838$ ($p < .001$).

Table A11
ITED Reading Comprehension Subtest Item Difficulty Indexes, Gifted Grade 8 and National Grade 10 Norm Groups

Item #	Fall Gifted IDI	Fall Norm IDI	Spring Gifted IDI	Spring Norm IDI
1	.82	.59	.61	.84
2	.79	.64	.66	.76
3	.87	.63	.65	.92
4	.90	.64	.66	.96
5	.64	.51	.53	.83
6	.54	.40	.42	.55
7	.77	.49	.51	.85
8	.82	.57	.59	.84
9	.72	.48	.50	.70
10	.87	.76	.77	.88
11	1.00	.82	.83	.92
12	.79	.74	.75	.80
13	.77	.63	.65	.76
14	.74	.73	.74	.85
15	.72	.56	.58	.63
16	.69	.68	.69	.65
17	.82	.75	.76	.81
18	.72	.64	.66	.77
19	.64	.73	.74	.71
20	.77	.73	.74	.70
21	.82	.71	.72	.79
22	.74	.63	.66	.75
23	.62	.70	.71	.63
24	.62	.69	.70	.75
25	.74	.75	.76	.80
26	.69	.68	.69	.70
27	.74	.70	.71	.71
28	.77	.57	.63	.72
29	.51	.48	.52	.63
30	.44	.40	.43	.61
31	.67	.58	.62	.76
32	.69	.48	.55	.67
33	.69	.47	.53	.71
34	.41	.52	.57	.57
35	.51	.51	.67	.56
36	.51	.42	.56	.48
37	.46	.50	.45	.52
38	.46	.42	46	.44
39	.46	.48	.54	.50
40	.36	.41	.46	.43
41	.18	.34	.29	.36
42	.41	.46	.44	.48
43	.38	.48	.44	.50
44	.56	.53	.54	.55

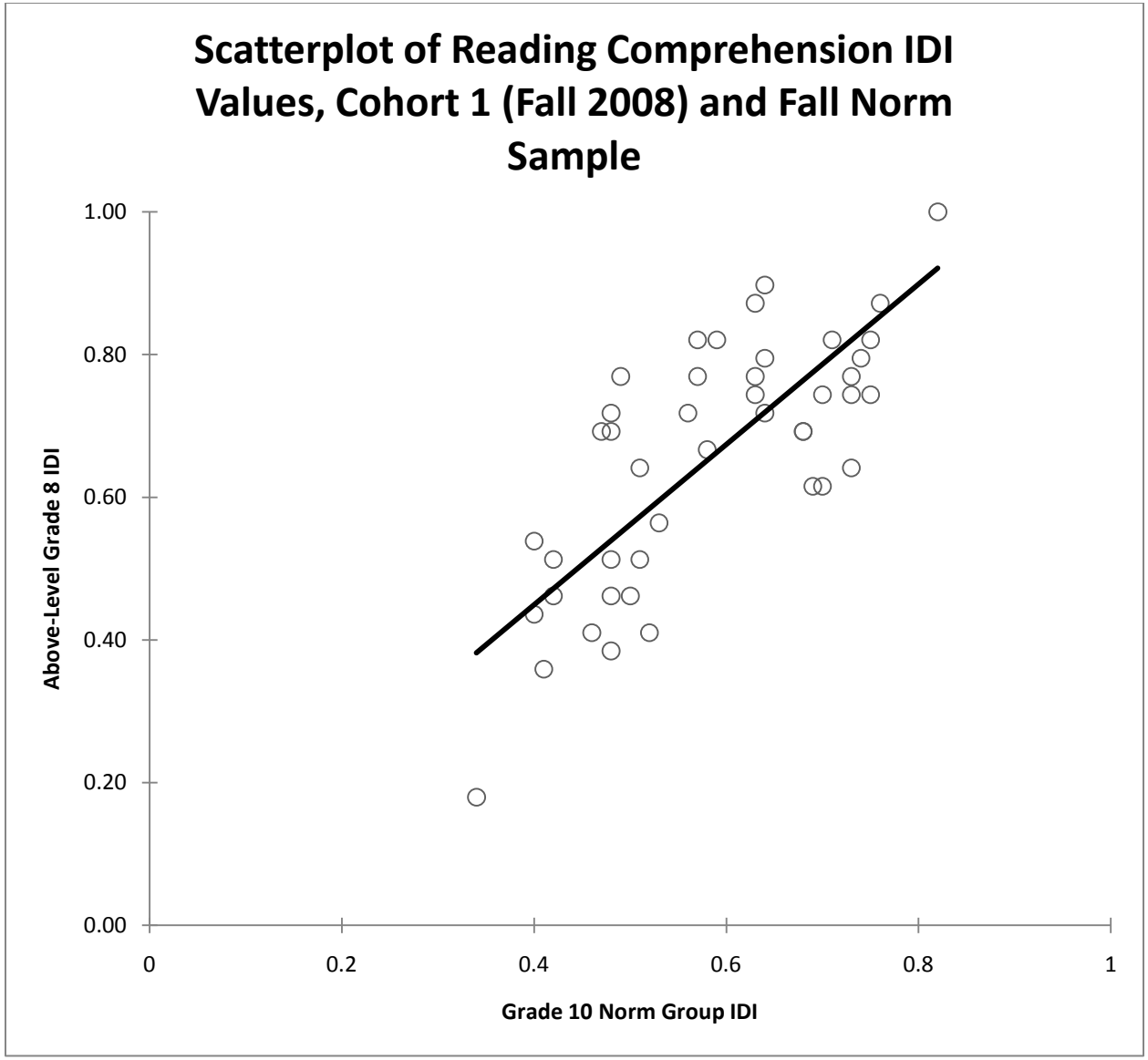

Figure A21 Scatterplot of reading comprehension subtest IDI values, Cohort 1 (Fall 2008) and Fall Norm Sample. IDI values correlation is $r = .750$ ($p < .001$).

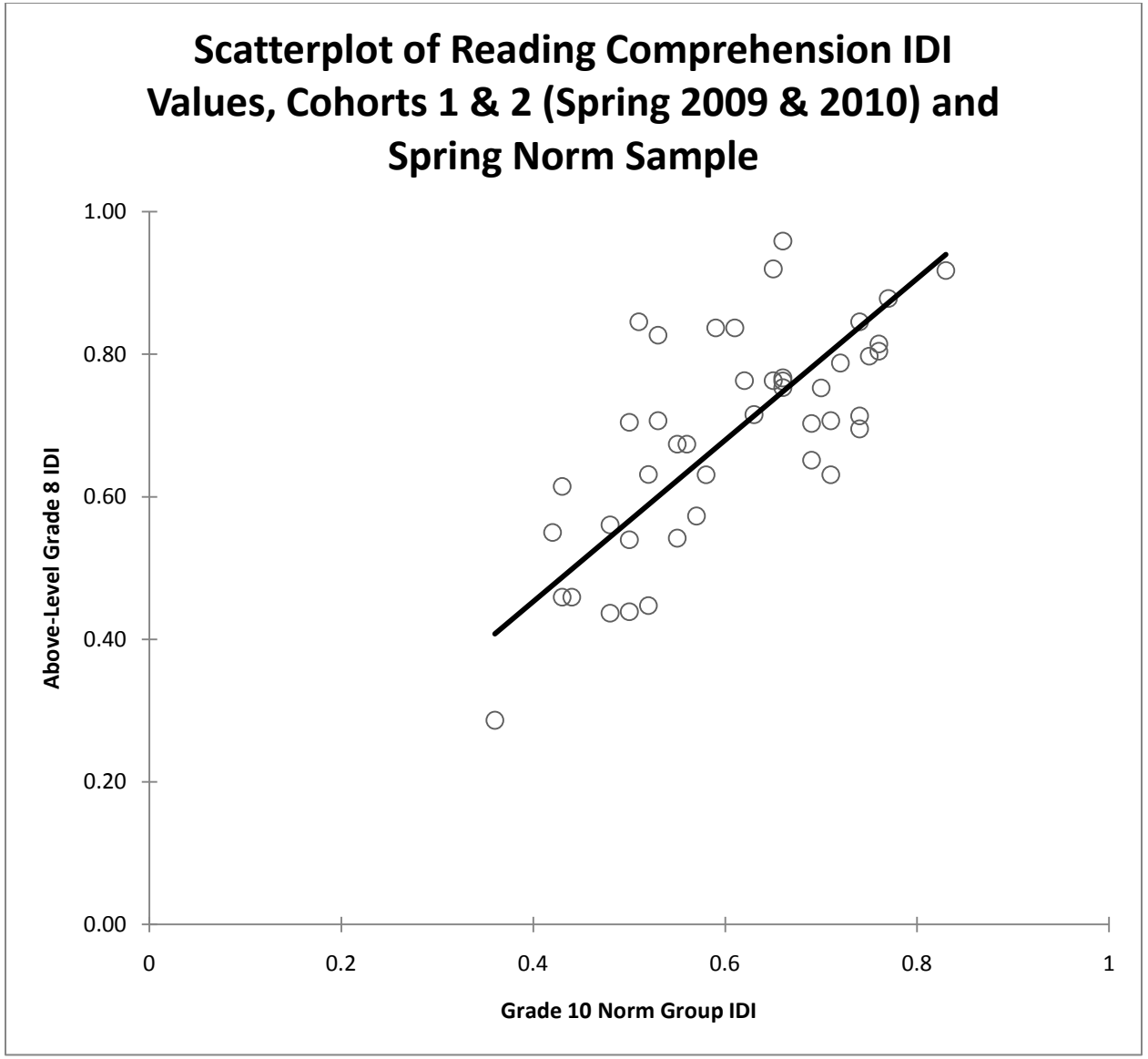

Figure A22 Scatterplot of reading comprehension subtest IDI values, Cohorts 1 & 2 (Spring 2009 & 2010) and Spring Norm Sample. IDI values correlation is $r = .713$ ($p < .001$).

Table A12
ITED Spelling Subtest Item Difficulty Indexes, Gifted Grade 8 and National Grade 10 Norm Groups

Item #	Fall Gifted IDI	Fall Norm IDI	Spring Gifted IDI	Spring Norm IDI
1	.38	.59	.50	.61
2	.38	.48	.43	.50
3	.41	.63	.46	.65
4	.56	.49	.63	.51
5	.36	.51	.45	.53
6	.44	.53	.42	.55
7	.59	.67	.68	.69
8	.56	.74	.52	.79
9	.44	.69	.37	.71
10	.49	.61	.59	.63
11	.54	.62	.64	.64
12	.62	.42	.71	.44
13	.36	.61	.63	.63
14	.44	.60	.48	.61
15	38	.48	.32	.50
16	.82	.43	.83	.44
17	.64	.65	.64	.66
18	.51	.51	.51	.55
19	.59	.72	.77	.75
20	.62	.71	.72	.74
21	.67	.58	.61	.61
22	.62	.63	.74	.67
23	.67	.59	.72	.62
24	.49	.65	.55	.68
25	.62	.60	.58	.62
26	.49	.61	.56	.64
27	.49	.60	.43	.61
28	.31	.43	.42	.45
29	.28	.46	.26	.49
30	.18	.72	.33	.73

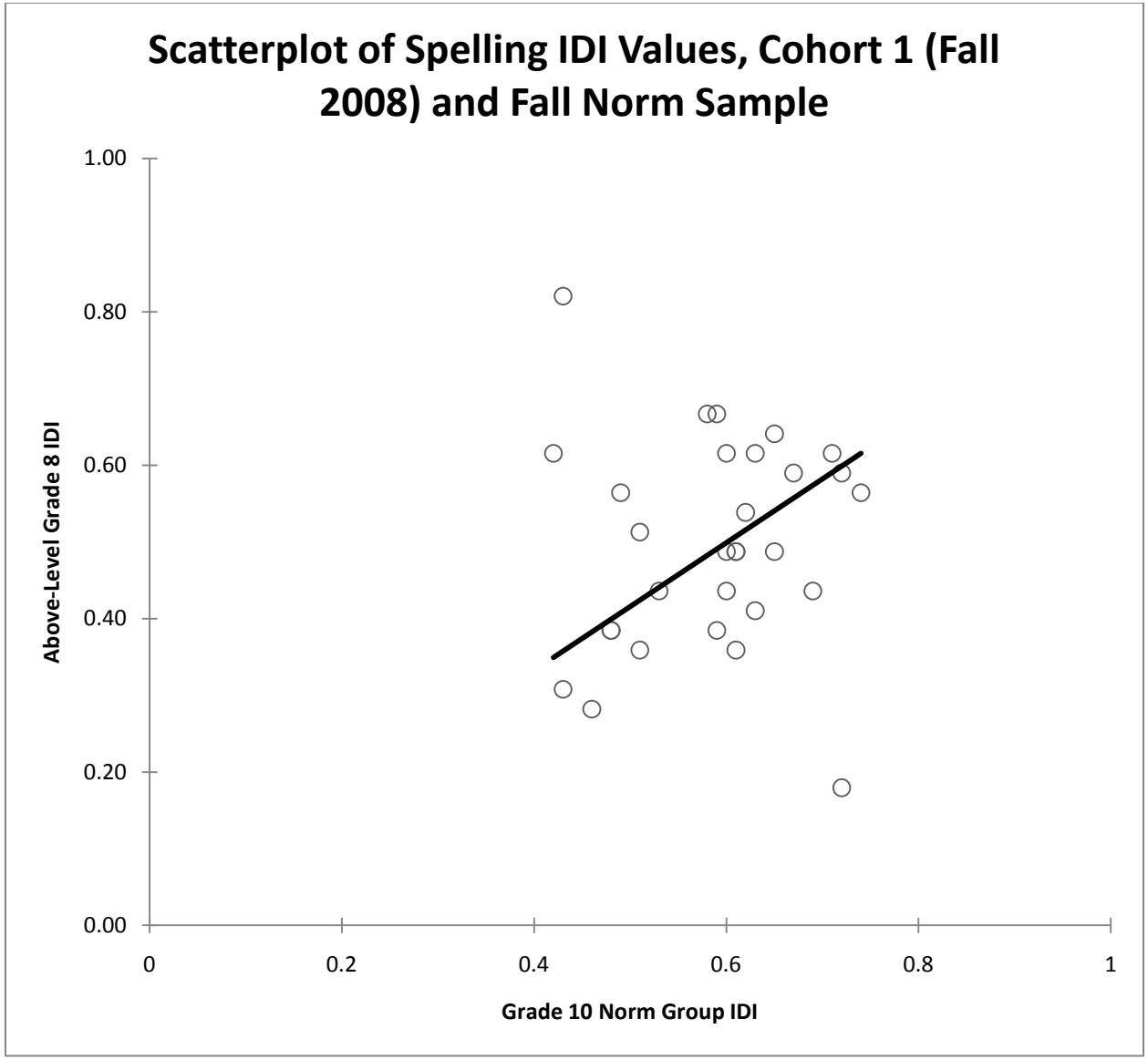

Figure A23 Scatterplot of spelling subtest IDI values, Cohort 1 (Fall 2008) and Fall Norm Sample. IDI values correlation is r = .056 (p = .769).

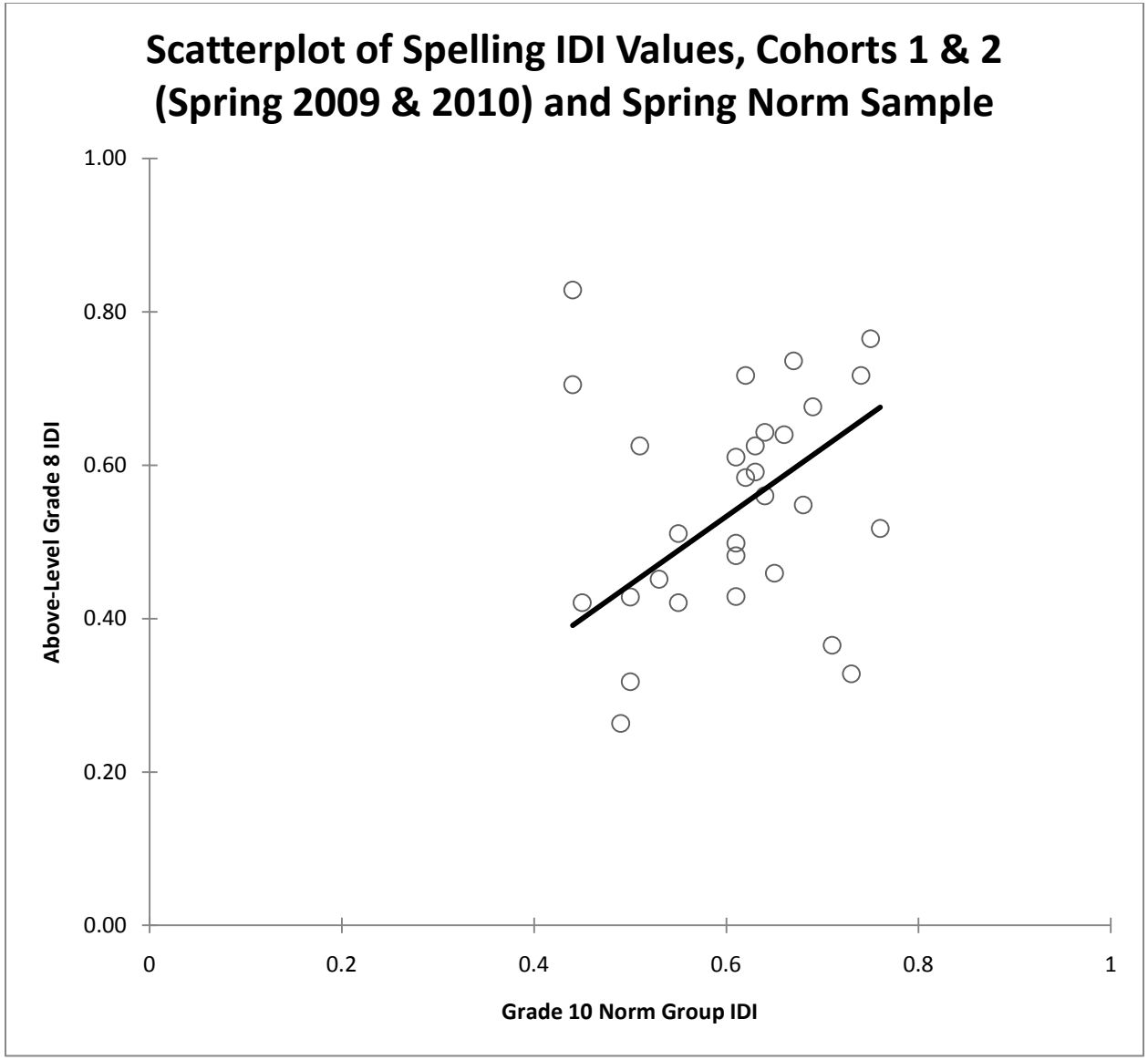

Figure A24 Scatterplot of spelling subtest IDI values, Cohorts 1 & 2 (Spring 2009 & 2010) and Spring Norm Sample. IDI values correlation is $r = .156$ ($p = .410$).

Table A13
ITED Revising Written Materials Subtest Item Difficulty Indexes, Gifted Grade 8 and National Grade 10 Norm Groups

Item #	Fall Gifted IDI	Fall Norm IDI	Spring Gifted IDI	Spring Norm IDI
1	.33	.47	.34	.48
2	.92	.41	.93	.46
3	.77	.43	.77	.47
4	.92	.61	.96	.64
5	.79	.44	.83	.45
6	.41	.31	.40	.33
7	.85	.62	.83	.65
8	.85	.48	.81	.49
9	.18	.37	.22	.38
10	.59	.55	.70	.57
11	.77	.55	.75	.56
12	.79	.58	.86	.61
13	.72	.48	.74	.51
14	.62	.35	.68	.36
15	.77	.37	.69	.41
16	.56	.72	.72	.75
17	.67	.55	.76	.58
18	.77	.53	.86	.56
19	.67	.70	.62	.73
20	.82	.57	.84	.60
21	.77	.57	.81	.60
22	.38	.59	.55	.62
23	.71	.63	.53	.66
24	.72	.47	.68	.50
25	.46	.46	.62	.50
26	.44	.58	.41	.61
27	.64	.39	.62	.42
28	.44	.31	.33	.34
29	.87	.32	.78	.35
30	.49	.37	.54	.40
31	.85	.37	.81	.39
32	.69	.73	.76	.75
33	.87	.60	.77	.62
34	.62	.76	.62	.78
35	.82	.38	.80	.70
36	.67	.29	.57	.30
37	.62	.63	.67	.65
38	.59	.68	.49	.70
39	.62	.28	73	.28
40	.64	.52	.61	.54
41	.54	.64	.63	.66
42	.51	.68	.63	.70
43	.67	.66	.75	.68
44	.59	.52	.60	.54
45	.69	.55	.82	.57
46	.77	.59	.79	.61
47	.49	.65	.60	.67
48	.69	.73	.74	.75
49	.59	.60	.65	.62
50	.54	.66	.75	.68
51	.67	.63	.73	.65
52	.62	.50	.78	.52

53	.59	.42	.67	.44
54	.49	.60	.50	.62
55	.38	.42	.59	.44
56	.64	.42	.73	.44

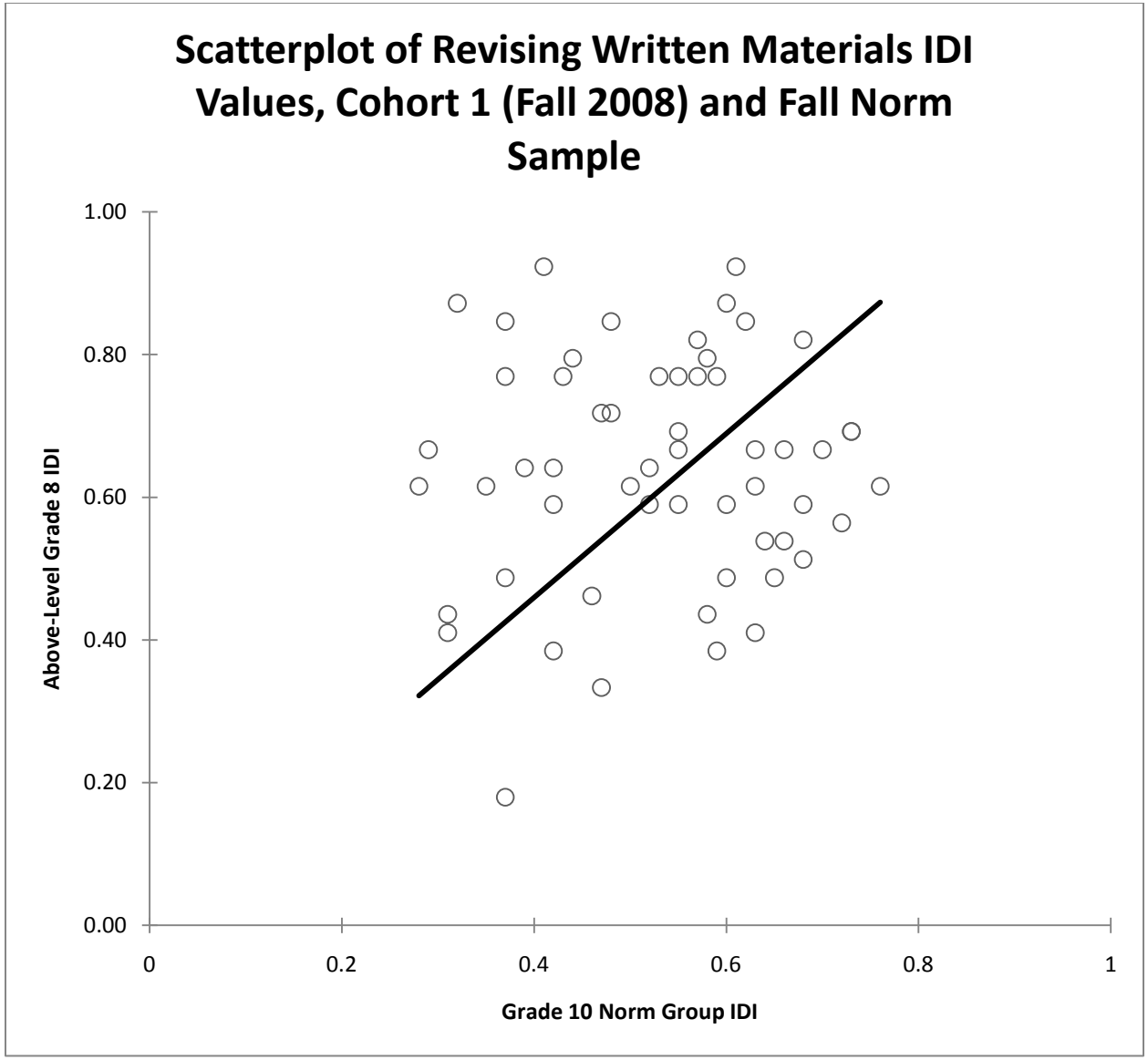

Figure A25 Scatterplot of revising written materials subtest IDI values, Cohort 1(Fall 2008) and Fall Norm Sample. IDI values correlation is r = .065 (p = .634).

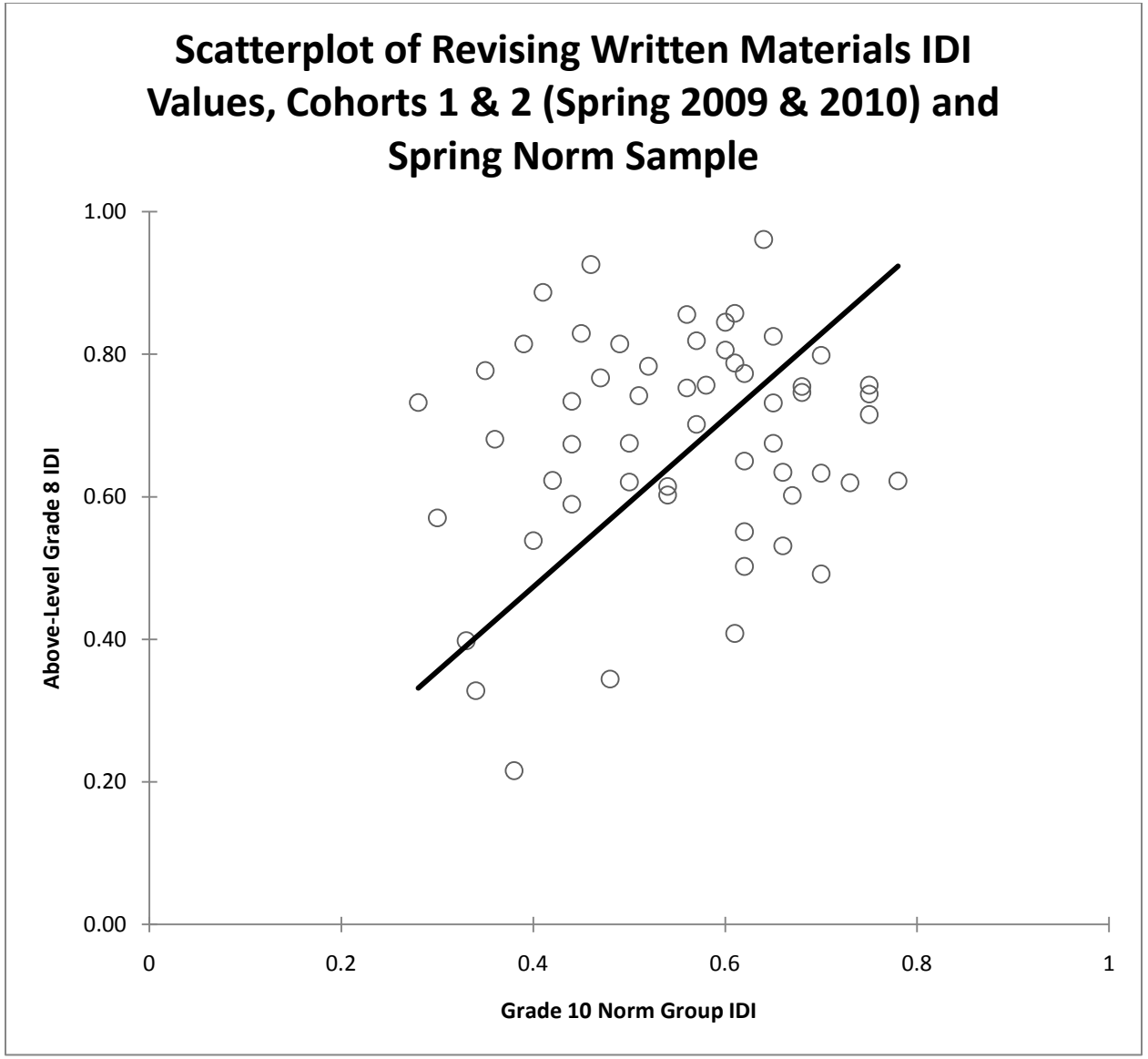

Figure A26 Scatterplot of revising written materials subtest IDI values, Cohorts 1 & 2 (Spring 2009 & 2010) and Spring Norm Sample. IDI values correlation is $r = .179$ ($p = .187$).

Table A14
ITED Mathematics Concepts & Problem Solving Subtest Item Difficulty Indexes, Gifted Grade 8 and National Grade 10 Norm Groups

Item #	Fall Gifted IDI	Fall Norm IDI	Spring Gifted IDI	Spring Norm IDI
1	.82	.70	.81	.74
2	.77	.44	.73	47
3	.67	.59	.67	.62
4	.74	.43	.74	.45
5	.62	.43	.67	.46
6	.67	.62	.80	.65
7	.31	.31	.31	.33
8	.64	.40	.62	.41
9	.67	.53	.76	.55
10	.74	.60	.84	.61
11	.46	.39	.62	.40
12	.87	.59	.74	.63
13	.54	.63	.73	.66
14	.85	.69	.79	.71
15	72	.61	.70	.63
16	49	.48	.60	.54
17	.15	.25	.14	.26
18	.38	.37	.41	.39
19	.72	.58	.75	.59
20	.51	.47	.57	.50
21	.56	.50	.57	.53
22	.49	.45	.62	.48
23	.62	.56	.62	.58
24	.28	.40	.31	.43
25	.54	.55	.66	.57
26	.54	.55	.66	.58
27	.36	.39	.36	.40
28	.67	.54	.69	.58
29	.33	.40	.38	.43
30	.31	.25	.64	.27
31	.46	.32	.33	.34
32	.54	.59	.65	.60
33	.33	.39	.40	.41
34	.38	.43	.49	.45
35	.38	.49	.56	.53
36	.21	.25	.41	.26
37	.56	.57	.65	.58
38	.31	.37	.40	.39
39	.38	.30	.29	.32
40	.23	.35	.30	.37

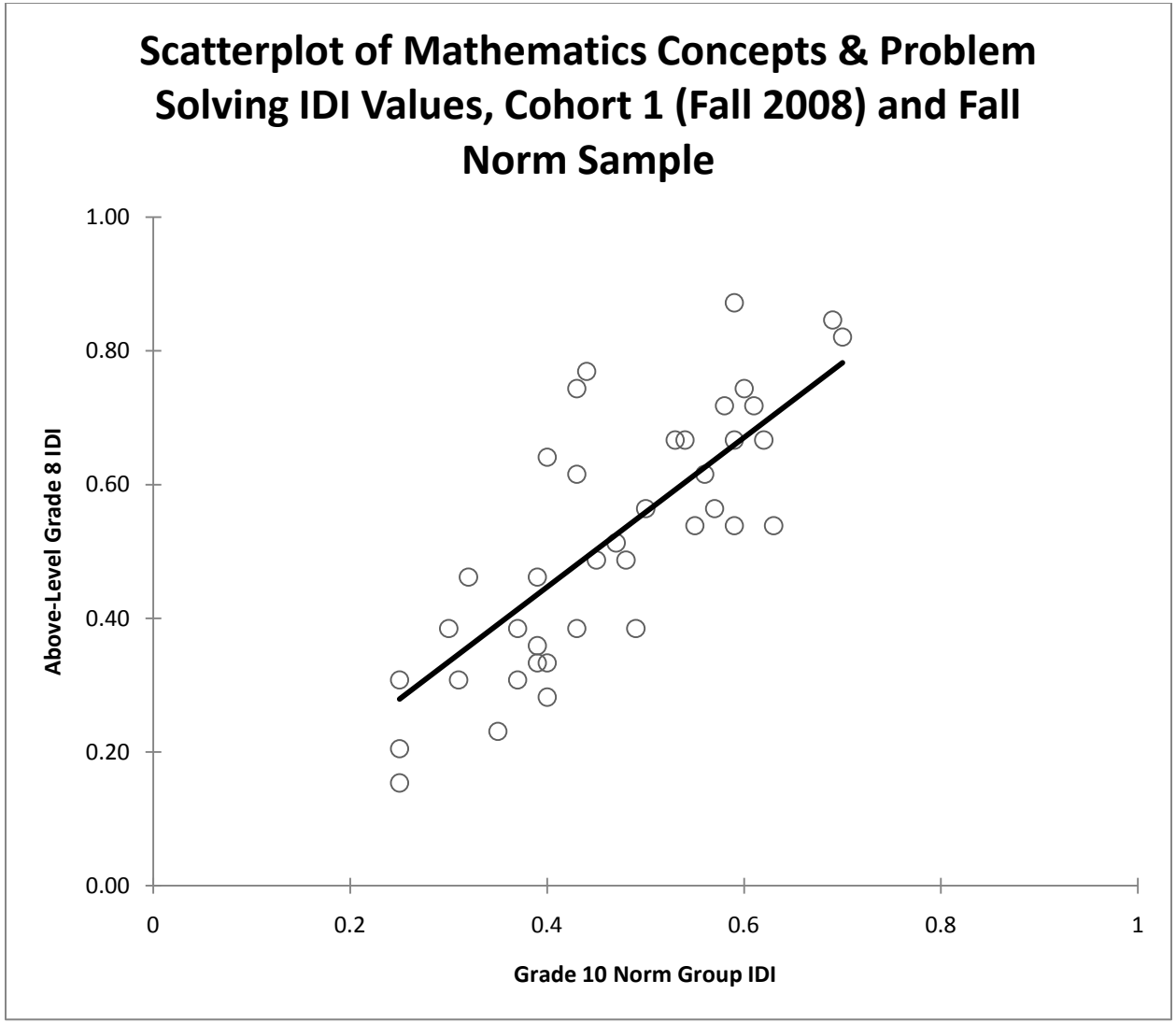

Figure A27 Scatterplot of mathematics concepts & problem solving subtest IDI values, Cohort 1 (Fall 2008) and Fall Norm Sample. IDI values correlation is $r = .798$ ($p < .001$).

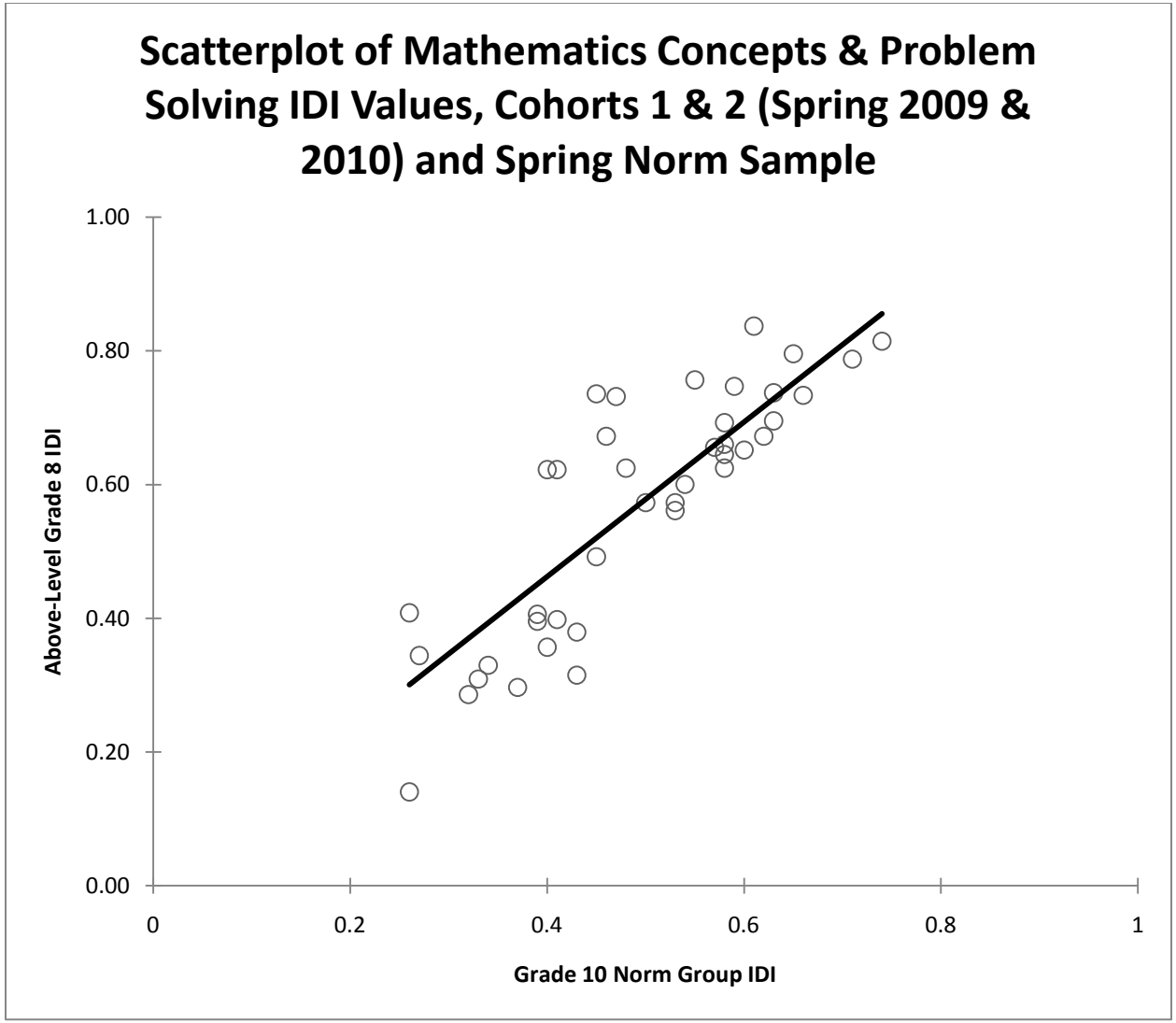

Figure A28 Scatterplot of mathematics concepts & problem solving subtest IDI values, Cohorts 1 & 2 (Spring 2009 & 2010) and Spring Norm Sample. IDI values correlation is $r = .854$ ($p < .001$).

Table A15
ITED Mathematics Computation Subtest Item Difficulty Indexes, Gifted Grade 8 and National Grade 10 Norm Groups

Item #	Fall Gifted IDI	Fall Norm IDI	Spring Gifted IDI	Spring Norm IDI
1	.87	.81	.86	.83
2	.72	.64	.73	.65
3	.85	.71	.83	.73
4	.51	.47	.50	.47
5	.72	.59	.73	.61
6	.69	.57	.79	.59
7	.41	.28	.40	.29
8	.59	.62	.62	.63
9	.56	.69	.64	.76
10	.77	.68	.74	.71
11	.10	.28	.12	.29
12	.15	.30	.19	.32
13	.54	.45	.61	.58
14	.23	.29	.27	.34
15	.21	.41	.26	.46
16	.74	.68	.67	.78
17	62	.60	.55	.72
18	.59	.65	.61	.67
19	.56	.60	.58	.68
20	.18	.37	.26	.40
21	.18	.27	.27	.28
22	.15	.37	.30	.43
23	.03	.25	.11	.27
24	.26	.43	.31	.46
25	.26	.46	.23	.47
26	.00	.25	.08	.26
27	.18	.42	.22	.44
28	.05	.40	.17	.43
29	.05	.29	.04	.30
30	.00	.24	.06	.25

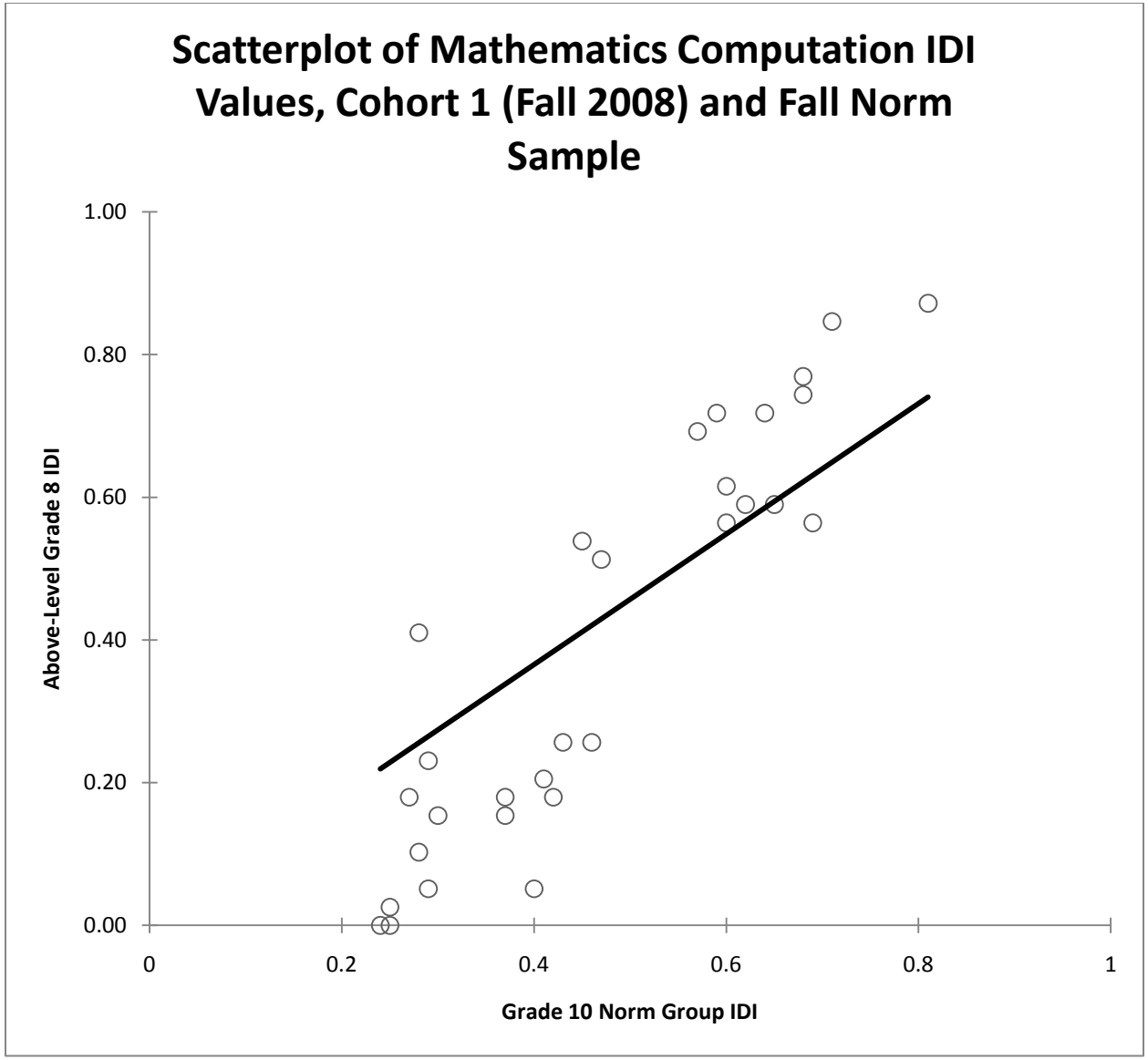

Figure A29 Scatterplot of mathematics computation subtest IDI values, Cohort 1 (Fall 2008) and Fall Norm Sample. IDI values correlation is $r = .798$ ($p < .001$).

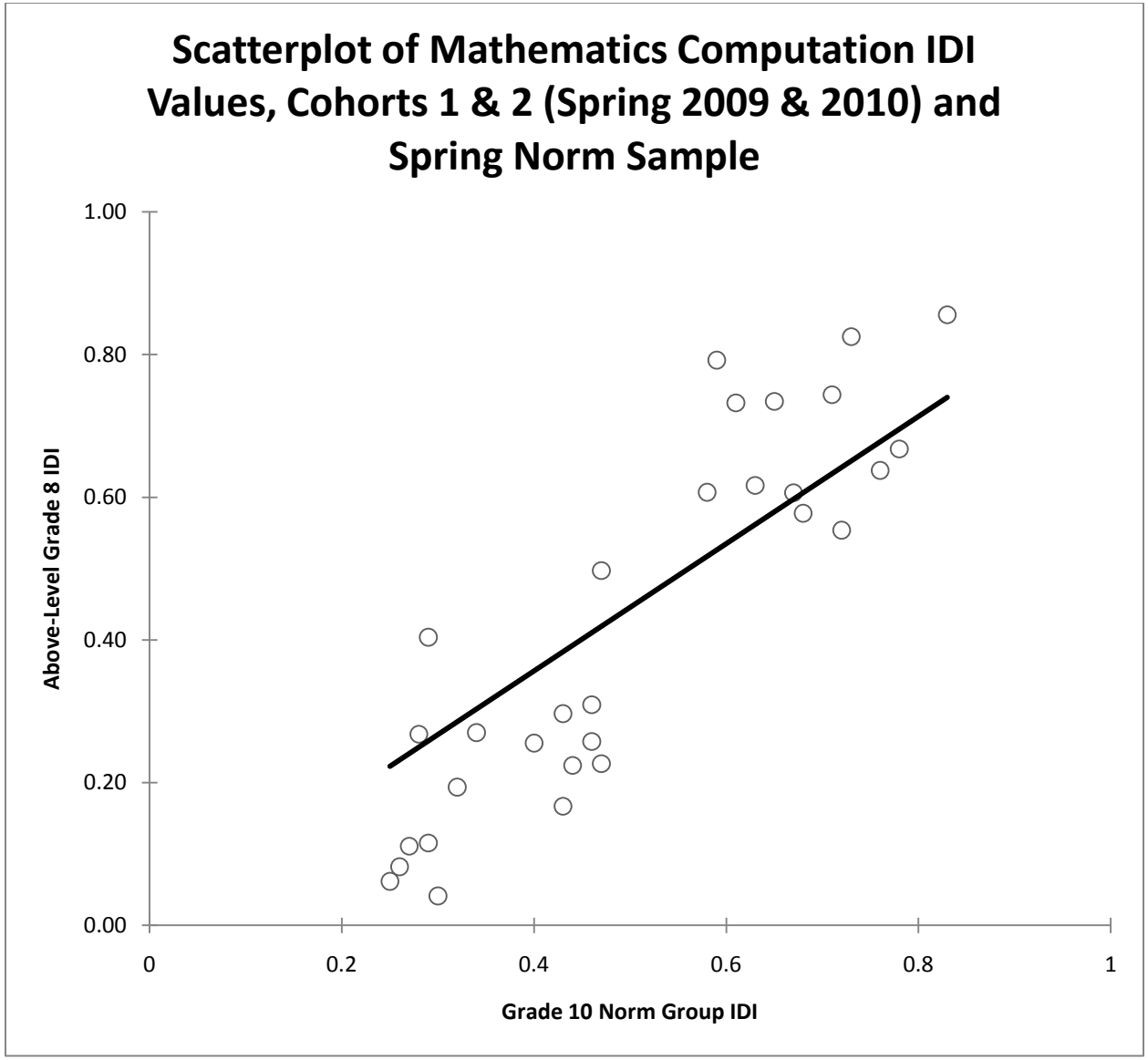

Figure A30 Scatterplot of mathematics computation subtest IDI values, Cohorts 1 & 2 (Spring 2009 & 2010) and Spring Norm Sample. IDI values correlation is $r = .854$ ($p < .001$).

Table A16
Item Difficulty Mean Comparisons Across Gifted and Norm Groups

	Fall					Spring				
Level and Subtest	Gifted Mean (SD)	Norms Mean (SD)	*t* (df)	*p*	*d*	Gifted Mean (SD)	Norms Mean (SD)	*t* (df)	*p*	*d*
Grade 8 Reading	.62 (.15)	.58 (.14)	1.19 (72)	.238	0.280	.69 (.16)	.62 (.13)	2.07 (72)	.042	0.487
Grade 8 Language	.54 (.18)	.53 (.15)	0.33 (116)	.742	0.061	.56 (.19)	.56 (.15)	0.00 (116)	1.000	0.000
Grade 8 Math	.44 (.24)	.47 (.14)	-0.73 (90)	.467	-0.154	.49 (.24)	.51 (.14)	-0.49 (90)	.625	-0.103
Grade 9 Vocabulary	.58 (.21)	.55 (.13)	0.77 (78)	.444	0.174	.58 (.13)	.65 (.20)	-1.86 (78)	.067	-0.420
Grade 9 Reading Comprehension	.66 (.17)	.56 (.12)	3.19 (86)	.002	0.688	.71 (.17)	.59 (.12)	3.83 (86)	< .001	0.825
Grade 9 Spelling	.51 (.20)	.51 (.12)	0.00 (58)	1.000	0.000	.54 (.20)	.53 (.12)	0.23 (58)	.819	0.062
Grade 9 Revising Written Materials	.59 (.17)	.50 (.12)	3.24 (110)	.002	0.617	.63 (.16)	.53 (.12)	3.74 (110)	< .001	0.714
Grade 9 Math Concepts & Problem Solving	.43 (.13)	.47 (.21)	-1.02 (78)	.311	-0.232	.57 (.22)	.47 (.13)	2.48 (78)	.015	0.560
Grade 9 Math Computation	.37 (.26)	.44 (.18)	-1.21 (58)	.231	-0.318	.39 (.25)	.47 (.19)	-1.40 (58)	.167	-0.366
Grade 10 Vocabulary	.61 (.18)	.58 (.11)	0.90 (78)	.371	0.204	.65 (.16)	.61 (.11)	1.30 (78)	.197	0.295
Grade 10 Reading Comprehension	.66 (.17)	.58 (.12)	2.55 (86)	.013	0.550	.69 (.15)	.61 (.12)	2.76 (86)	.007	0.545
Grade 10 Spelling	.50 (.14)	.59 (.09)	-2.96 (58)	.005	-0.778	.55 (.14)	.61 (.09)	-1.97 (58)	.054	-0.778
Grade 10 Revising Written Materials	.64 (.16)	.53 (.13)	3.99 (110)	< .001	0.761	.55 (.13)	.68 (.15)	-4.90 (110)	< .001	-0.761
Grade 10 Math Concepts & Problem Solving	.52 (.19)	.47 (.12)	1.41 (78)	.163	0.319	.57 (.18)	.49 (.13)	2.28 (78)	.025	0.319
Grade 10 Math Computation	.39 (.28)	.47 (.17)	-1.34 (58)	.186	-0.351	.42 (.26)	.50 (.18)	-1.39 (58)	.170	-0.351

VITA

Russell Thomas Warne
Behavioral Science Dept., Utah Valley University
800 W. University Parkway, Orem, UT 84058
rwarne@tamu.edu

EDUCATION

Texas A&M University, College Station, TX — May 2011
Ph.D., Educational Psychology
Emphasis: Research, Measurement, and Statistics

Brigham Young University, Provo, UT — April 2007
B.S., Psychology
Minor: Theatre Studies

PUBLICATIONS

Warne, R. T. (in press). An investigation of measurement invariance across genders on the Overexcitability Questionnaire-Two (OEQII). *Journal of Advanced Academics*.

Warne, R. T., McKyer, E. L. J., & Smith, M. L. (in press). An introduction to item response theory for health behavior researchers. *American Journal of Health Behavior*.

Larsen, R., & Warne, R. T. (2010). Estimating confidence intervals for eigenvalues in exploratory factor analysis. *Behavior Research Methods, 42*, 871-876. doi: 10.3758/BRM.42.3.871

Warne, R. T. (2009). Comparing tests used to identify ethnically diverse gifted children: A critical response to Lewis, DeCamp-Fritson, Ramage, McFarland, & Archwamety. *Multicultural Education, 17*(1), 48-53.

Warne, R. T., & Juntune, J. (2009). Recent trends in gifted identification in Texas. *Tempo, 29*(4), 22-26.

Yanchar, S. C., Slife, B. D., & Warne, R. (2009). Advancing disciplinary practice through critical thinking: A rejoinder to Bensley. *Review of General Psychology, 13*, 278-280. doi: 10.1037/a0017135

Yanchar, S. C., Slife, B. D., & Warne, R. (2008). Critical thinking as disciplinary practice. *Review of General Psychology, 12*, 265-281. doi: 10.1037/1089-2680.12.3.265