Roeper Review, 34:183–193, 2012

ISSN: 0278-3193 print / 1940-865X online
DOI: 10.1080/02783193.2012.686425

Routledge
Taylor & Francis Group

IDENTIFICATION ISSUES

History and Development of Above-Level Testing of the Gifted

Russell T. Warne

Above-level testing (also called *out-of-level testing*, *off-grade testing*, and *off-level testing*) is the practice of administering a test level that was designed for and normed on an older population to a gifted child. This comprehensive literature review traces the practice of above-level testing from the earliest days of gifted education through the present. It was found that there were five reasons frequently given for above-level testing: raising the test ceiling, increasing score variability and discrimination, improving reliability, the sound interpretations of above-level test data, and reducing regression toward the mean. Although all of these reasons were theoretically supported, the strength of the empirical evidence varied. The article concludes by suggesting future directions of psychometric and applied research in above-level testing.

Keywords: above-level testing, assessment, gifted, literature review, off-level testing, psychometrics, special populations, testing

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 article is to provide a comprehensive literature review that traces the genesis, development, and present status of above-level testing in gifted education. The author also critically evaluates the current state of the literature supporting above-level testing and gives recommendations for further research on the practice.

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. 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., Lee, Matthews, & Olszewski-Kubilius, 2008) and *off-level testing* (e.g., Gross, 2004).

Several procedures were used in the attempt to gather all relevant scholarly literature on above-level testing. First,

Accepted 17 August 2011.

Address correspondence to Russell T. Warne, Department of Behavioral Science, Utah Valley University, 800 W. University Parkway LA-012G, Orem, UT 84050. E-mail: rwarne@uvu.edu

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 author examined early case studies of high-ability children 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 14-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). Unfortunately, Terman and Fenton did not explain why they gave these above-level tests to the 7-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 (1926) would later administer the eighth-grade level of the Stanford Achievement Test (SAT) 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 did not 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.

In addition, 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 was 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 of 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, 1974, 1990). Stanley, 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 200 middle-school students in Maryland who scored above the mean of high-school seniors on the SAT-M (Stanley, 1976). 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 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 five general claims about the benefits of above-level testing:

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

The following section of this article examines the psychometric theory behind these claims and evaluates 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, Colangelo, Lupkowski-Shoplik, Lipscomb, & Forstadt, 2009; Feldhusen, Proctor, & Black, 2002; Olszewski-Kubilius & Kulieke, 2008; 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; 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 4 or more years older than them (Ebmeier & Schmulback, 1989; Lupkowski-Shoplik & Swiatek, 1999; Olszewski-Kubilius, Kulieke, Willis, & Krasney, 1989; VanTassell-Baska, 1986).[1] More research needs to be done to investigate the exact 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, "[O]ne 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 (Keating, 1975, 1976; Lupkowski-Shoplik, Benbow, Assouline, & Brody, 2003; Swiatek, 2007; VanTassel-Baska, 1996; Wendler, Ninneman, & Feigenbaum, 2001). 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; VanTassel-Baska, 1986; Wendler et al., 2001) and above-level test scores often form a distribution that is approximately normal (Keating & Stanley, 1972; Lupkowski-Shoplik & Swiatek, 1999; Wendler et al., 2001). 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).

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; Pollins, 1984; 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 0.01% 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 et al., 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. 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 reliability coefficients from above-level achievement tests administered to a gifted sample 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 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 that all researchers report the reliability of the data at hand (American Educational Research Association, 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. 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. Therefore, more research is needed to determine whether the assumptions on the reliability of above-level test scores are tenable.

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 multilevel tests, such as the Cognitive Abilities Test (CogAT; Lohman & Hagen, 2002), provide 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. Researchers could also report item and test information statistics, which are item response theory-based statistics that are analogous to reliability. Information statistics—like the conditional SEM—also vary by item difficulty. Readers should consult Embretson and Reise (2000) for an accessible introduction to this and other aspects of item response theory.

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). 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. A future study that investigates the factor structure and measurement invariance of above-level test items would fill one of the most important gaps in the current understanding of above-level testing.

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% 5 years later. Similarly, when Terman retested some children in his gifted sample about 8 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 (z_2) from the following equation:

$$z_2 = r_{xy} \cdot z_1$$

where r_{xy} 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

$$|z_2| - |z_1|,$$

which can easily be converted back to the units in which that the original scores measured.[2]

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 preadolescents' SAT scores are moderately good predictors of advanced placement 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 grade point average. 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). In addition, 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 $h = 0.16$), acquire a patent ($h = 0.18$), and obtain tenure at a university ($h = 0.28$; 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 2 and 5 years above the child's grade level and it is in the child's and the program administrator's 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 is not as intensive or selective as Talent Search (Olszewski-Kubilius et al., 1989).

Threlfall and Hargreaves (2008) conducted a study to see whether 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; Lee & Olszewski-Kubilius, 2006; Pollins, 1984; Stanley, 1977–1978; Wendler et al., 2001) 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 on above-level testing has provided a firm foundation for understanding how above-level tests function with gifted populations. The findings also have led to experimentation in above-level testing in nonacademic domains (Achter et al., 1996; Gross, 2004). However, there are still some issues that remain unresolved. Most important, 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; 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. In addition, more tests should be evaluated for their suitability for above-level testing.

Evidence for the 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; Wendler et al., 2001). 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., Anastasi, 1974; Merwin & Gardner, 1962; Schmeiser & Welch, 2006; Zwick, 2006). Modern theorists recognize aptitude as a product of interest, motivation, affect, the specific environment, intelligence, metacognitive abilities, and academic experiences (Corno et al., 2002). At most, above-level tests may measure the knowledge-based and reasoning aspects of academic aptitude (Pollins, 1984). 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: the test level, the age of the child, the opportunity to learn the more advanced material, test content, etc. 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) or through SMPY. Therefore, it is not clear whether above-level academic achievement tests outperform vertically aligned aptitude tests (like the CogAT) in predicting successful grade acceleration. 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 when and under what circumstances 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 processes 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, factor structure of above-level test items, measurement invariance across age groups, 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 achievement test scores and understanding what construct(s) they may be measuring may be difficult to determine with the SAT. A multilevel, vertically aligned academic achievement 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 in academic areas, program evaluation, and many other topics related to above-level testing.

In addition, 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 the 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.

Notes

1. It should be noted that 7.0% of Lupkowski-Shoplik and Swiatek's (1999) sample were only 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.
2. This paragraph is a simplified discussion of regression toward the mean. A more detailed and technical treatise on the relationship between reliability, high ability, and regression toward the mean can be found in Ziegler and Ziegler (2009).

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

Anastasi, A. (1974). Commentary on the precocity project. In J. C. Stanley, D. P. Keating, & L. H. Fox (Eds.), *Mathematical talent: Discovery, description, and development* (pp. 87–100). Baltimore, MD: Johns Hopkins University Press.

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

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 of 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*. Retrieved from Eric database. (ED449174)

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.

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.

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.

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

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

Feldhusen, J. F., Proctor, T. B., & Black, K. N. (2002). Guidelines for grade advancement of precocious children. *Roeper Review*, *24*, 169–171. doi:10.1080/02783198609553000

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

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.

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.

Johnsen, S. K., & Corn, A. L. (2001). *Screening assessment for gifted elementary and middle school student's 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, MD: 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. doi: 10.3102/0013189X001009003

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., 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

Lee, S., 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

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., & 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

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

Loyd, B. H. (1980). *Functional level testing and reliability: An empirical study* (Unpublished 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

Lupkowksi-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, MA: Allyn & 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.

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

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

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

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., & 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

Pollins, L. D. (1984). *The construct validity of the Scholastic Aptitude Test for young gifted students* (Unpublished doctoral dissertation). Duke University, Durham, NC.

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* (5th ed., Technical Manual). Itasca, IL: Riverside.

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

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

Stanley, J. C., Jr. (1951). On the adequacy of standardized test administered to extreme norm groups. *Peabody Journal of Education*, *29*, 145–153. doi: 10.1080/01619565109536335

Stanley, J. C. (1954). Identification of superior learners in Grades ten through fourteen. In *Promoting maximal reading growth among able learners* (Supplemental Educational Monograph no. 81, pp. 31–34). Chicago, IL: University of Chicago Press.

Stanley, J. C. (1974). Intellectual precocity. In J. C. Stanley, D. P. Keating, & L. H. Fox (Eds.), *Mathematical talent: Discovery, description, and development* (pp. 1–22). Baltimore, MD: Johns Hopkins University 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, MD: The Johns Hopkins University Press.

Stanley, J. C. (1977–1978). The predictive value of the SAT for brilliant seventh- and eighth-graders. *College Board Review*, *106*, 30–37.

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.

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

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 at mid-life: Thirty-five years' follow-up of the superior child*. Stanford, CA: Stanford University Press.

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

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

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.

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), 47–52.

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

Wendler, C., Ninneman, A., & Feigenbaum, M. (2001). *Evaluating the appropriateness of the SAT I: Reasoning Test for seventh and eighth graders* (College Board Notes RN-12). New York, NY: The College Board Office of Research.

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.

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

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.

AUTHOR BIO

Russell T. Warne is an assistant professor of psychology in the Department of Behavioral Science at Utah Valley University. He earned a bachelor's degree in psychology from Brigham Young University and a doctoral degree in educational psychology from Texas A&M University. His research interests are in psychometrics, assessment, quantitative methodology, intellectual giftedness, and intelligence. E-mail: rwarne@uvu.edu