Feature Article

Differentiating Among High-Achieving Learners: A Comparison of Classical Test Theory and Item Response Theory on Above-Level Testing

Gifted Child Quarterly
2020, Vol. 64(3) 219–237

DOI: 10.1177/0016986220924050
journals.sagepub.com/home/gcq
SAGE

Brandon LeBeau[1], Susan G. Assouline[1], Duhita Mahatmya[1], and Ann Lupkowski-Shoplik[1]

Abstract

This study investigated the application of item response theory (IRT) to expand the range of ability estimates for gifted (hereinafter referred to as high-achieving) students' performance on an above-level test. Using a sample of fourth- to sixth-grade high-achieving students ($N = 1{,}893$), we conducted a study to compare estimates from two measurement theories, classical test theory (CTT) and IRT. CTT and IRT make different assumptions about the analysis that impact the reliability and validity of the scores obtained from the test. IRT can also differentiate students based on the student's grade or within a grade by using the unique string of correct and incorrect answers the student makes while taking the test. This differentiation may have implications for identifying or classifying students who are ready for advanced coursework. An exploration of the differentiation for Math, Reading, and Science tests and the impact the different measurement frameworks can have on classification of students are explored. Implications for academic talent identification with the talent search model and development of academic talent are discussed.

Keywords

item response theory, talent search, talent identification

The federal definition of giftedness was first presented in the Marland (1972) report, the initial federal report to address the needs of gifted students. Giftedness referred to individuals with high intellectual ability or aptitude and who represent a minimum of 3% to 5% of the school-aged population. This definition would set the stage for identification processes and programming in schools that are highly variable across states and largely unchanged today (Callahan et al., 2017; No Child Left Behind [NCLB], 2003). The identification process for gifted education in school-based programs often is considered controversial because (a) the decisions for inclusion in school-based programs are dichotomous (i.e., yes/no), despite the fact that the data on which a programming decision is made is continuous (McBee et al., 2016); (b) the school-based programs are exclusive because they commonly serve 3% to 5% of the student population and selection is typically based on a combination of composite scores from various measures (McBee et al., 2014); and (c) there is little alignment between the definition of giftedness and schools' gifted program identification systems and the gifted program curriculum (Callahan et al., 2017). Outside of school-based gifted education identification and programming, as implemented in the typical school setting, there is a broadly used system of academic talent identification, referred to as the talent search model (Assouline & Lupkowski-Shoplik, 2012; Calvert, 2018; Stanley, 2005; Warne, 2012) that offers a different approach to identifying giftedness and creating programs based on the identification system.

The talent search model, developed in the early 1970s (Stanley, 2005), operationalized the definition of domain-specific talent through the process of large-scale above-level testing with high-achieving students. The talent search model begins by determining which students qualify as high achievers on grade-level achievement tests (see Assouline & Lupkowski-Shoplik, 2012; Lee et al., 2008; Olszewski-Kubilius, 2015, for detailed descriptions of the process), with high achievers

[1]University of Iowa, Iowa City, IA, USA

Corresponding Author:
Brandon LeBeau, University of Iowa, 600 Blank Honors Center, Iowa City, IA 52242, USA.
Email: brandon-lebeau@uiowa.edu

traditionally operationally defined as students at or above the 95th percentile on a grade-level achievement test (Lupkowski-Shoplik & Swiatek, 1999) when compared with grade-level-peers. Achievement, as measured by students' performance on a standardized grade-level achievement test, indicates what a student has already learned; the assumption is that students at the 95th percentile or higher have mastered the grade-level content as measured by the standardized achievement test. The model has proven effective as an approach for identifying high-achieving students, who have potential for higher achievement in advanced academic opportunities in specific content areas such as math (Brody & Mills, 2005; Lee et al., 2008; Lupkowski-Shoplik & Assouline, 1993; Olszewski-Kubilius, 2015; Olszewski-Kubilius & Lee, 2005; Stanley, 2000; Warne, 2012).

Both gifted education programming in some schools and the talent search model, used largely outside of the school setting, use grade-level achievement tests as the initial step in the identification process. However, the two systems diverge beyond that point. The demonstration of mastery notwithstanding, there are measurement issues that need to be considered when interpreting the scores earned by high achievers on grade-level tests (Warne, 2012). A fixed form grade-level standardized test is not designed to have a sufficient number of difficult items for high achievers, that is, those students whose scores are at the 95th percentile compared with their grade-level peers. The relative lack of difficult items is an artifact of test design where a grade-level test is typically designed to measure the middle range of academic performance across a normative sample, which creates ceiling effects for high scorers. The ceiling effect restricts the range of scores, which, in turn, reduces the reliability of the scores for high achievers. Grade-level scores earned by high achievers cannot adequately indicate a student's readiness for advanced content or how much advanced material the student requires for appropriate challenge and engagement in the content. Although the grade-level achievement results represent what have been learned (achieved), the grade-level test does not have enough ceiling to determine high-achieving students' readiness for future learning of new content. These nuanced distinctions between the concepts of past achievement and readiness for advanced material are important with respect to understanding alignment of identification with programming. When high achievers have hit the ceiling on an achievement test, one solution, typically employed by talent search programs, is to administer an above-level test to improve the differentiation of performance among high achievers and create greater alignment between assessment and programming.

Above-Level Testing

Above-level testing means administering a test developed for older students to younger students in order to have sufficiently difficult items to measure aptitude (i.e., readiness for advanced content). Measuring aptitude, used synonymously with potential, in a content area is the primary purpose of the talent search model (Brody & Mills, 2005; Stanley, 2005). Because of the assumption that aptitude represents a latent ability trait within an individual, the performance on the above-level test becomes a proxy of readiness, that is, potential to learn advanced content.

Above-level testing serves multiple purposes (Assouline & Lupkowski-Shoplik, 2012). First, it provides a brief, and early, intervention for high-achieving students, who often are not challenged in the regular school setting (Peters et al., 2017). Second, it creates a new distribution of high-achieving students (Assouline & Lupkowski-Shoplik, 2012). In a normal distribution of standardized grade-level achievement test scores of typically developing students, high-achieving students typically cluster at the right tail of the bell curve (greater than or equal to the 95th percentile). When those high-achieving students take an above-level test, the above-level test scores show greater variability among the students (Warne, 2012) and offer insights into their aptitude in the specific academic domains represented by the tested items. By using the new distribution of scores from above-level testing, educators can better differentiate among students who are more likely to benefit from enrichment and extension of the coursework and those students who are highly capable and would benefit from accelerative approaches. Brody and Mills (2005) indicate that the above-level performance "discriminates well within the group tested so that students with exceptionally advanced reasoning abilities can be identified and their educational programs adjusted to include more advanced content" (p. 98).

Additionally, Callahan et al.'s (2017) evaluation study offered an extensive review of relevant research that connects quality of curriculum with level of challenge, which increases achievement of high-ability students. Their study revealed that very few schools use the talent search model despite the potential for (a) discriminating among high-achieving students to identify students with high aptitude for advanced content and (b) aligning advanced curriculum with the learning needs of advanced learners. Of the 389 elementary districts in the study (Callahan et al., 2017), only 1 (0.3%) used the talent search model for their programs; a slightly higher percentage of middle school gifted programs (4 out of 286 or 1.4%) used the talent search model for programming. Callahan et al. (2017) corroborated the findings of Olszewski-Kubilius and Lee (2005) who

determined that most school educators recognize the value of the results of above-level testing; however, the purpose for a model that is based on above-level testing is perceived by educators to be primarily for out-of-school programming, not for use in determining readiness for advanced curriculum in schools. This may be one reason why, despite the logic inherent in this process of finding students with potential in a particular content area and then providing advanced curriculum, there is a dearth of research concerning the validity of interpretations of above-level testing results (Warne, 2012), especially in school-based settings.

The goal for the current study is to add to the above-level testing literature by investigating an extension of the statistical procedures used to interpret the above-level scores by high-achieving students who participate in the Talent Search Model through the calculation and interpretation of above-level test scores using item response theory (IRT; de Ayala, 2009; Lord, 1980). We investigate how using IRT to calculate estimates of aptitude can better differentiate the range of aptitude scores for high-achieving students. By demonstrating the feasibility of applying IRT to above-level testing in schools, findings from this research offer an approach to broaden the identification of highly capable students and interpretation of the above-level test scores for educational programming. Below, our discussion of data and testing reference above-level testing unless specified otherwise.

Using Classical Test Theory to Understand Above-Level Testing With Gifted Students

For decades (Lee et al., 2008; Olszewski-Kubilius & Lee, 2005; Stanley, 2005; Warne, 2012), the talent search model relied on results grounded in classical test theory (CTT; McDonald, 2011) to understand the performance of very bright students on above-level tests. Measures of central tendency, including scale scores and percentile rankings, offer educators a basic way of understanding student performance.

CTT assumes the observed score for an individual is a function of their true score and some error. More formally this representation is as follows:

$$X = T + E \tag{1}$$

where X represents the observed score, T represents the true score, and E represents error. Practically, only the observed score will be known; however, the true score is what is of interest to educators. The true score is a latent variable free from measurement error. In theory, this would be the individual's score if they responded to an infinite number of items. The observed score is based on that individual's responses to a limited number of items and is a combination of two latent variables: true score and measurement error. Because there are two latent or unobservable variables, statistical assumptions are made. The three primary CTT assumptions include the following: (a) both latent variables, true score and error score, are uncorrelated; (b) the average error across the population of individuals is zero; and (c) the error scores across parallel tests are uncorrelated. It is also common to assume that error scores follow a normal distribution (Hambleton & Jones, 1993). From these assumptions, the true score is operationalized as the difference between test score and error score or the expected test score over parallel tests (Hambleton & Jones, 1993).

The CTT framework has been successfully applied to create many high-quality standardized tests as well as to understand the performance of students on standardized tests. The focus through a CTT measurement framework has been at the test score level. Although the observed test score is a sum of individual items or questions answered correctly, CTT assumes that all of the items are equally discriminating and equally reliable and that the individual items are interchangeable. There are item-level statistics, such as proportion correct for each item, representing item difficulty, and biserial correlations, representing item discrimination, that have aided test developers in a CTT framework to create tests with desirable properties. However, the primary limitation of these statistics is that the true scores and item statistics are sample dependent, which can limit their usefulness, particularly when the specific sample obtained differs from the population for which the test was developed. This can have implications for using an above-level test as a way to identify aptitude in a specific subject, which is the traditional use of above-level testing within the Talent Search Model. High-achieving students, who are two or more grades younger than the normative sample, represent a different population from that for which the test was developed. Therefore, the inferences made from this framework do not automatically have the validity evidence desired for making decisions with a high-achieving sample. Furthermore, ceiling effects of fixed form grade-level achievement tests, which restrict variation of students' scores, further complicates the identification of high-achieving students who are ready for advanced content.

Using Item Response Theory to Understand Above-Level Testing With Gifted Students

The basic assumption with respect to high-achieving students is that they are ready for advanced content because they have already mastered the content at a particular grade level, as indicated by their achievement on a grade-level test. But how does an educator, psychologist, or parent know how advanced the content should be? Readiness for advanced material vis-à-vis above-level

testing connects to Vygotsky's Zone of Proximal Development, a fundamental principle of developmental psychology (Shabani et al., 2010). Because "both boredom and confusion can lead to distraction, frustration, and lack of motivation" (Shabani et al., 2010, p. 241), educators must be aware that if content is too advanced, the student may be unnecessarily frustrated and unable to learn. If content is too easy, the student is not learning something new.

Although overall performance on above-level tests, specifically the number of items answered correctly, can be one indicator of how advanced the material should be, important information about performance on specific items is lacking when overall performance is the only indicator.

Response strings from an individual student's answers to an above-level standardized test could provide further information regarding mastery of content. For example, a subset of the items could be used to identify subscales that students have mastered, such as mastery of algebra within the math domain. In addition, psychometricians and researchers use students' response strings to better understand the test's properties. Combining information about the number of items answered correctly with information about which items were answered correctly may provide additional information to better understand which students are ready for advanced content and what type of content.

IRT, which is sometimes referred to as modern measurement theory, models the likelihood that an individual with a given ability (aptitude) level, as revealed through performance on an above-level test in this case, will answer an item correctly. IRT has become a popular alternative to CTT for item analysis due to its flexibility and the ability to overcome some of the limitations found with CTT analyses. Most notably, the CTT limitations that can be overcome include invariance of the item parameters (e.g., discrimination and difficulty) and invariance of a latent trait (i.e., aptitude). These two related properties theoretically allow IRT to overcome the sample dependent nature of the CTT framework. Parameter invariance in IRT theory states that item parameters and latent trait scores (i.e., aptitude scores) are invariant up to a linear transformation for different samples taking the same test (de Ayala, 2009; Rupp & Zumbo, 2006). Practically, this assumption means that two samples from two different states could be placed on the same scale through a linear transformation and interpreted similarly. In addition, this assumption can be empirically explored by performing differential item functioning (DIF) or through measurement invariance testing.

A general IRT model for dichotomous responses (i.e., items scored correct or incorrect) is Birnbaum's three parameter logistic model (3pl; Birnbaum, 1968):

$$p(x_j = 1|\theta,\ a_j, b_j,\ c_j) = c_j + (1 - c_j)\frac{1}{1 + e^{-a_j(\theta - b_j)}} \quad (2)$$

in which the probability of answering the item correctly is based on four parameters, θ, a_j, b_j, and c_j, representing the latent variable, item discrimination, item difficulty, and a lower asymptote (also referred to as the guessing item parameter), respectively. The j subscript on the item discrimination, difficulty, and guessing parameters indicate that these are estimated uniquely for each item. Simpler models can be obtained by fixing parameter values. For example, a two parameter logistic (2pl) IRT model can be obtained by fixing the c_j parameter to 0, and a one parameter logistic (1pl) IRT model (sometimes called the Rasch model) can be obtained by fixing both the c_j and a_j values to 0 and 1, respectively. The model depicted above is unidimensional, meaning that a single latent trait is assumed to underlie the test (e.g., math aptitude for a math test). However, if this assumption is violated, multidimensional models could be explored as well, but these models have greater data requirements (see Reckase [2009] for more information about multidimensional IRT).

Comparing CTT and IRT Approaches for Interpreting Results From Above-Level Testing

CTT provides a framework to interpret the above-level test scores relative to the normed sample originally obtained by the test developer. In the above-level framework, the normative sample represents a group of students from a single grade that is at least two grade levels higher than the current grade of the above-level test-takers. Scores from CTT in this context would not consider the above-level test taker's current grade level or how far the student's age is from that of the target population. For example, a fourth grader and a sixth grader take the same above-level test (e.g., a test developed with eighth-grade content), and both students receive a score at the 50th percentile compared with the eighth-grade normative group. The fourth-grade student's aptitude is considered the same as the sixth-grade student's aptitude because both are at the 50th percentile based on the eighth-grade normative group. CTT would not differentiate explicitly among these students' aptitude. Users interpreting these percentiles may implicitly adjust scores at the same percentile for different grade levels; however, being more explicit and intentional about this difference would aid in the interpretation of these scores.

In contrast, IRT would be able to distinguish between these students through the questions answered correctly or incorrectly (i.e., their unique response string). For example, suppose a simple three-item exam was given

and the response strings for two students were as follows: Student 1: C I C and Student 2: I C C; where I represents incorrect and C represents correct. In this scenario, these two students would receive different aptitude estimates if IRT pattern scoring were used. The items that have larger discrimination will have more influence on the aptitude score, and items that have higher difficulty can greatly increase an aptitude estimate. IRT can also be implemented where grade-level average performance differences can be taken into account to create a developmental aptitude scale by estimation of the average latent aptitude means (and standard deviations) for each grade-level. Another advantage of IRT is through linking methods (Kolen & Brennan, 2014), which allows for the IRT characteristics of the questions and the aptitude scale to be placed on the same scale regardless of the questions used on the test. Grade-level differences and linking methodology can allow for growth to be explored more easily if this assessment is given more than one time. For example, grade-level specific distributions can be used as a reference point that can be linked together to ensure they are on the same scale.

Classifying High-Achieving Students for Gifted Programs

Classifying or identifying high-achieving students ready for additional challenge(s) has been a large part of research about identification in the gifted education literature (Lakin, 2018; Lohman, 2005; McBee et al., 2016; Peters et al., 2019). The inference here is that identification procedures should result in a decision regarding which students could benefit from additional challenge compared with the typical classroom environment while limiting false positives and negatives in identification (McBee et al., 2016). Lohman (2005) also argues for the identification to be based on the context the additional challenge would represent. More specifically, if a student is considered for advanced instruction or acceleration in math, understanding more about the student's math-related skills would be the best predictor of whether a student would succeed. The above-level test can help approximate aptitude in a particular subject through content for which the student may not have been exposed. The aptitude scores from the above-level test can be obtained using either CTT or IRT methods, and currently we are aware of no research study that has explored differences in these approaches in an above-level context. Thus, our exploratory study compares and contrasts obtaining scores using CTT and IRT in an above-level context through three research questions:

- Research Question 1: How do aptitude estimates obtained through CTT compare with estimates obtained from IRT in the above-level testing environment?
- Research Question 2: How does the item-level analysis with IRT impact aptitude estimates?
- Research Question 3: How do CTT and IRT methods provide similar classifications of high aptitude for students initially identified as high achieving on grade-level tests?

Method

Participants and Procedures

Participants were recruited using a university-based regional talent search approach (Lee et al., 2008). Educators received registration information from a university-based talent search program and notified parents that their school would nominate qualified students to participate in above-level testing. Fourth through sixth-grade students who scored at the 95th percentile or above on at least one section of a standardized grade-level test such as the *Iowa Assessments* were recommended to participate. Educators received the guidelines for recommendations for participation and were encouraged to use local norms to identify students. In some cases, educators determined that they wanted to include more students in the above-level testing step, so they invited more students than those who had qualified based on the 95th percentile criterion. For the current sample, and slightly different from traditional talent search models, the educator at a local school selected a test date and time that was convenient for the school and students and subsequently invited parents or guardians to register their children for testing based on the 95th percentile, or moderately lower (i.e., 85th to 90th percentile), guideline. Registration was completed online, and teachers could register students directly if families did not have access to a computer. Testing was also conducted online, using desktop or laptop computers provided by the school. Local teachers served as test proctors.

The study draws from students who participated in above-level testing between March 2016 and July 2017. Students who cheated, became ill, experienced technical difficulties, or attempted less than 20% of all the items were eliminated from analysis ($n = 2$) for a total analytic sample of 1,893 students across 92 schools. Students were in the fourth (260 students; 13.7%), fifth (532 students, 28.1%), or sixth (1,101 students, 58.2%) grades; 49% were female, and 51% were male overall; and about 90% of the sample identified as White. The percentages of male and White students were similar across grades. A majority of the students are from three Midwestern states, with nearly all (96%) being from one of the states.

Participants (percentages of 0.1% each) came from six other states across the country. The sample for this study is based on students with valid scores on all of the subtests (English, Math, Reading, and Science) administered.

Instrument

I-Excel is a computer-based test that uses an online platform to deliver an above-level assessment to high-achieving fourth- through sixth-grade students. I-Excel uses content developed by and licensed from ACT(2013) that was designed to measure the academic progress of eighth-grade students. The content of I-Excel consists of four multiple-choice subtests that cover: English (40 items), Math (30 items), Reading (30 items), and Science (28 items). The content is given in the same order to all examinees for each of the subtests. Because the content is at an eighth-grade level yet administered to high-achieving fourth to sixth graders, it is being used as an above-level test. Similar content was previously available as the Explore test (ACT, 2013) and was administered from 1993 to 2014 in a paper and pencil format through ACT test administration as part of the traditional model of students taking the test as part of a regional talent search on a Saturday morning (Assouline & Lupkowski-Shoplik, 2012; Lee et al., 2008; Olszewski-Kubilius, 2015). The multiple-choice format of I-Excel also is similar to Explore; however, the I-Excel platform is online and administered in the student's classroom. Classroom teachers are proctors and follow a detailed set of directions to ensure the security of the test.

Raw scores are converted to scale scores with a range of 1 to 25, and parents and teachers receive an interpretive report based on the scale scores for each of the four subtests. Current understanding of the results of above-level testing for elementary students is based largely on the results for Explore (Lee et al., 2008) and understood within the context of the normative group's (i.e., eighth graders) performance. Average scale scores and standard deviations for eighth graders are: English, 15.20 (4.3); Math, 15.9 (3.6); Reading, 15.00 (4.1); and Science, 16.80 (3.4). ACT(2013) reports the Explore reliability coefficients and average standard of errors of measurement using weighted frequency distributions. Kuder–Richardson 20 (KR-20) internal consistency reliability coefficients for Form A, Grade 8 Explore scale scores are English, .84; Math, 0.76; Reading, 0.86; and Science, 0.79. In addition, the standard error of measurement (SEM) for Form A, Grade 8 are English, 1.66; Math, 1.71; Reading, 1.44; and Science 1.53. Although the traditional approach to reporting scores has been to convert the raw scores to scale scores, the current analysis relied on the raw scores or percentile ranks for I-Excel. Using the raw scores allows us to compute reliability and SEM for this sample.

Data Analysis

CTT Methods. The number correct for each student and percentile ranks across the four subject areas are the primary statistics of interest in this study. Using the number correct metric, information about the reliability of the assessment for the above-level population was explored using coefficient alpha and a group-level reliability discussed by Raju et al. (2007). This calculation was done as follows:

$$\rho_{xx} = \frac{\sigma_x^2 - E\left(SEM_s^2\right)}{\sigma_x^2} \tag{3}$$

In Equation 3, ρ_{xx} is the population reliability, σ_x^2 is the variability in the observed raw scores, and $E\left(SEM_s^2\right)$ is the average of the individual SEM. The average SEMs were also used to explore differences in measurement precision with this above-level population.

IRT Methods. Data were not sufficient to estimate the lower asymptote found in the 3pl IRT model (de Ayala, 2009) due to smaller sample sizes in individual grade levels. Therefore, a two-parameter logistic (2pl) IRT model was fitted to the item-level data obtained from students who tested above-grade level. The 2pl IRT model takes the following form:

$$p\left(x_j = 1|\theta, a_j, b_j\right) = \frac{1}{1 + e^{-a_j(\theta - b_j)}} \tag{4}$$

where the probability of answering the item correctly is based on three parameters, θ, a_j, and b_j representing the latent variable, item discrimination, and item difficulty, respectively. The j subscript on the item discrimination and difficulty indicate that these are estimated uniquely for each item. A separate model was fitted to each subtest, English, Math, Reading, and Science. The mirt R package (version 1.31) was used for all IRT model fitting (Chalmers, 2012). The 2pl IRT model allowed for differential discrimination of each item.

The IRT model was fitted in a multigroup framework where the groups represented the different grade levels. In the current analysis, fourth grade was chosen as the reference group, and the latent aptitude scale was fixed to have a mean of 0 and a standard deviation of 1. The other two groups representing fifth and sixth grade were allowed to have their means and standard deviations vary to accommodate differences in performance and variation across the grade levels. It was expected that on average the students in fifth and sixth grade would have higher performance compared with fourth-grade students. Default estimation using the expectation-maximization algorithm first defined by Bock and Aitkin (1981) was used to

estimate item parameters. Evaluation of model fit was done with the M_2 statistic ($M_2 \approx \chi^2$ and was evaluated with $\alpha = .01$; Maydeu-Olivares & Joe, 2006), the root mean square error of approximation (RMSEA, if $<.05$), and the comparative fit index (CFI, if $>.93$). These thresholds are commonly used in the structural equation modeling literature (Hu & Bentler, 1999).

Initially, item parameters were constrained to be equal across the grade levels. On estimation of the constrained multigroup IRT models, DIF was explored to see if items were invariant across the grade levels. If evidence of DIF was found for a specific item, item parameters were allowed to vary for that item across the grades. This yielded different estimates for the items across grades that showed evidence of DIF, but other items that did not show evidence of DIF were held constant across grades. A highly constrained model was specified first (i.e., all the item parameters were constrained to be equal), then an iterative procedure was used to relax the constrained item parameters across the groups one by one. Likelihood ratio tests and model fit indices (i.e., AICc) were used to determine if an item showed evidence of DIF. The IRT model that allows items to vary across grades is more general compared with the IRT model with item parameters fixed across grades; therefore, these model results are presented below. The number of items that showed evidence of DIF for each subtest were: 9 out of 40 English items (22.5%), 12 out of 30 Math items (40%), 6 out of 30 Reading items (20%), and 3 out of 28 Science items (11%).

Reliability and average SEM were also estimated based on the IRT model. The group-level IRT reliability follows the discussion found in Raju et al. (2007). This approach calculates the SEM for each individual as the inverse of the test information at the individual's estimated aptitude and compares this with the variation in the estimated abilities. More formally, the group-level IRT reliability is calculated as follows:

$$\rho_{\hat{\theta}_s\hat{\theta}_s} = \frac{\hat{\sigma}^2_{\hat{\theta}} - E\left(SEM_s^2\right)}{\hat{\sigma}^2_{\hat{\theta}}} \tag{5}$$

In Equation 5, $\rho_{\hat{\theta}_s\hat{\theta}_s}$ is the population reliability of the estimated latent construct, $\hat{\sigma}^2_{\hat{\theta}}$ represents the variance of the estimated latent constructs (i.e., aptitude estimates in this study), and $E\left(SEM_s^2\right)$ is the average of the individual SEM, which is the inverse of the test information at the individual's estimated aptitude. The average SEMs were also used to explore differences in measurement precision with this above-level population.

A single-group unidimensional 2pl IRT model was also fitted to each content area. This model takes the same form as that in Equation 4 but ignores grade level. This approach would be more similar to the CTT approach, which uses the eighth-grade normative sample when reporting scale scores and percentiles. Using the single-group IRT model, a single cut score, similar to CTT, was used to determine student readiness for more advanced coursework in the areas tested. The same fit statistics, criteria, reliability, and average SEM measures were used as described for the multiple-group IRT model.

Classification. Classification of high-achieving students as having high-aptitude for advanced curriculum under CTT and IRT methods were compared. In the CTT framework, the 50th percentile using the scaled scores on the norm-referenced eighth-grade population was used (Lupkowski-Shoplik et al., 2003) to determine readiness for advanced or accelerated curriculum. Using IRT, aptitude scores for the students were estimated using Expected A Posteriori (EAP) scoring (Bock & Aitkin, 1981). EAP scoring is a Bayesian pattern-scoring procedure that uses the item parameters and the individual response strings to generate an estimated aptitude score. This means that for the same number correct, answering different items correctly will provide different aptitude estimates. For example, if one student were to answer a hypothetical three question exam: I C C and another student were to answer the same three question exam: C I C; where C means the student answered the question correctly, I incorrectly. In this scenario, these two students would have slightly different estimated aptitude scores when using pattern-scoring procedures and the estimated aptitude scores would depend on the item parameters for those three items. EAP scoring also implements a prior distribution that helps reduce variability in the estimates compared with maximum likelihood and provides estimates for individuals who may answer all the items correctly and incorrectly (Kolen & Tong, 2010). Similar to the CTT approach, the 50th percentile of the aptitude scores was used as the criterion to classify a student as ready for advanced or accelerated curriculum. However, unlike CTT, which used the 50th percentile of the norm-referenced eighth-grade population, the 50th percentile was calculated within grade levels because there are expected increases in performance as grade level increases. More specifically, average performance for a sample of high-achieving students on the above-level test is higher for sixth-grade students, and the multigroup IRT allows the average performance to vary across grades. Therefore, high-achieving students in sixth grade will need a higher above-level score to be classified as high aptitude compared with their fourth- and fifth-grade high-achieving peers and fifth-grade students will need a higher above-level score compared with those high achievers in fourth grade. A single cut score was used across grades representing the 50th percentile of the

above-level scores for each subject with the single-group IRT model, in contrast to the multiple-group IRT model where a cut score was established for each grade.

Results

IRT Model Fit and Reliability

Prior to exploring the aptitude estimates from the IRT model, IRT model fit and reliability of the scores were explored. Table 1 shows the model fit for the final IRT models fitted in a multiple-group and single-group framework. The table shows a significant M_2 statistic for the four subjects, however, the RMSEA and CFI in many cases met the criteria for adequate fit of less than 0.05 and larger than 0.93. The primary exception to this was for the IRT model fitted to the English test data with a CFI of .89 and .90, for multigroup and single-group IRT, respectively, suggesting less evidence of relative fit. The RMSEA for this model was within the boundaries; therefore, some caution needs to be taken when interpreting the IRT results from the English test data. With larger sample sizes, a three-parameter model or a multidimensional model may be worth exploring. The CFI was generally higher for the single-group IRT than for the multigroup IRT; however, the RMSEA was much lower in the multigroup IRT framework. We put more emphasis on the RMSEA than the CFI, therefore prefer the multigroup series of models.

Reliability of the scores obtained from the above-level tests are reported in Table 2. This table contains coefficient alpha, CTT group level reliability, and IRT group level reliability for the multigroup and single-group IRT models. Coefficient alpha was reported because it is a commonly reported reliability statistic; the CTT and IRT group level reliabilities (see Equations 3 and 5) are calculated similarly and provide a more direct comparison between CTT and IRT. In general, coefficient alpha provided estimates of the score reliability that are larger than the other two estimates and were consistently slightly smaller than .8. The CTT and IRT group level reliability follow similar trends and are similar in magnitude, but in many instances the IRT reliability was higher compared with the CTT reliability. Exceptions to this trend occurred for the English test across all grade levels and the fourth-grade Math test. Much lower reliability estimates were found for the Math assessment, particularly for fourth grade. The sample size for this grade was much smaller than the other two grades, which may affect the reliability estimate. The single-group IRT reliability estimates were higher than the multigroup IRT reliability in most cases, but the differences were modest. The one exception was the reliability for the fourth-grade Math test which was .35 compared with .46 for the multigroup IRT and single-group reliability, respectively.

Further information about the reliability and information of the assessments can be gained by exploring the average SEM estimates under the CTT and IRT framework (Table 2). The average SEM for the IRT framework is on a different scale and scale differences were adjusted by multiplying the IRT average SEM by the standard deviation of the raw scores to place the IRT and CTT average SEM on the same scale. Table 2 shows that the average SEM tends to be smallest when using coefficient alpha for its computation; this is not surprising given that the coefficient alpha reliability estimates tended to be larger. The average SEM also tended to be smaller in the IRT framework compared with CTT. Exceptions to this pattern occurred for three of the four sixth-grade tests (only the average SEM was smaller for IRT in the Reading test) and for the fourth-grade math test. Differences between the CTT and IRT average SEM are larger than half a score point for fourth-grade English, Reading, and fifth-grade Reading test with the IRT average SEM showing more accurate scores. The fourth-grade Math test was on average about a third of a score point more accurate in the CTT framework compared with the IRT framework. All other tests across the grade levels represented smaller differences, less than a third of a score point.

Regarding magnitude of differences, the metric of the SEM is interpreted on the number correct metric and is reported as whole scores. For example, suppose a fourth-grade student correctly answered 10 out of 30 math items correctly. The average SEM statistics give a sense of the degree of precision around that score of

Table 1. Fit Statistics for Two-Parameter IRT Multigroup Models by Subject.

	Multigroup IRT			Single-group IRT		
Subject	M_2	RMSEA [CI]	CFI	M_2	RMSEA [CI]	CFI
English	3930 (2340), $p < .01$	0.019 [0.018, 0.020]	0.893	2320 (740), $p < .01$	0.034 [0.032, 0.035]	0.906
Math	1630 (1283), $p < .01$	0.012 [0.010, 0.014]	0.949	805 (405), $p < .01$	0.023 [0.021, 0.025]	0.966
Reading	1895 (1307), $p < .01$	0.015 [0.014, 0.017]	0.965	902 (405), $p < .01$	0.025 [0.023, 0.028]	0.974
Science	1818 (1146), $p < .01$	0.018 [0.016, 0.019]	0.936	948 (350), $p < .01$	0.030 [0.028, 0.032]	0.952

Note. IRT = item response theory; RMSEA = root mean square error of approximation; CI = confidence interval; CFI = confirmatory factor index.

Table 2. Reliability and Average Standard Error of Measurement (SEM) Estimates for the Four Subject Tests by Grade Level, Calculated From CTT and IRT Frameworks.

	Grade 4				Grade 5				Grade 6			
	Alpha	CTT	MG-IRT	SG-IRT	Alpha	CTT	MG-IRT	SG-IRT	Alpha	CTT	MG-IRT	SG-IRT
Reliability												
English	0.83	0.79	0.79	0.77	0.77	0.73	0.70	0.72	0.79	0.76	0.68	0.69
Math	0.63	0.42	0.35	0.46	0.67	0.49	0.52	0.62	0.75	0.55	0.62	0.64
Reading	0.82	0.65	0.72	0.74	0.81	0.61	0.68	0.70	0.80	0.49	0.54	0.55
Science	0.71	0.55	0.59	0.67	0.76	0.59	0.67	0.69	0.77	0.55	0.64	0.64
Average SEM												
English	2.68	2.93	2.48	2.59	2.68	2.90	2.60	2.63	2.55	2.74	2.79	2.82
Math	2.28	2.85	3.21	2.31	2.35	2.96	2.94	2.25	2.31	3.09	3.28	2.38
Reading	2.24	3.09	2.24	2.03	2.15	3.09	2.50	2.23	1.95	3.06	3.02	2.59
Science	2.37	2.96	2.66	2.12	2.31	3.03	2.94	2.28	2.22	3.10	3.24	2.47

Note. Alpha is coefficient alpha; CTT is the group-level reliability shown in Equation 3; IRT is the group-level reliability found in Equation 5. For average standard error of measurement calculations, the SEM was calculated using alpha and CTT group-level reliability, whereas IRT was test information based. IRT = item response theory; CTT = classical test theory; MG-IRT = multigroup IRT; SG-IRT = single-group IRT.

10. With the average SEM for fourth graders on the Math test being 2.3 (coefficient alpha), 2.9 (CTT), 3.2 (multigroup IRT), and 2.3 (single-group IRT), confidence intervals can be created around the reported score to give ranges of likely values where the student's true number correct would be found. For example, an approximate 68% confidence interval (assuming a normal distribution) could be created around the score of 10 for each method as: coefficient alpha = 7.7 to 12.3; CTT = 7.1 to 12.9; multiple-group IRT = 6.8 to 13.2; and single-group IRT = 7.7 to 12.3. These scores would then be rounded to the nearest number correct for reporting and, for the most part, the methods would provide similar ranges for plausible scores with 68% confidence, with the range for CTT and multiple-group IRT extending an additional one score point on either side of the reported score range.

The reliability values reported in Table 2 were, on average, smaller; and, therefore, the average SEM statistics were larger compared with those published in the Explore technical manual (ACT, 2013; see the instrument section above for the values reported in the technical manual). This finding is not surprising given that the students taking these tests are taking the test above-level and have likely not been exposed to the content found on the tests. The statistics reported for the above-level performance from our sample of high-achieving students are also smaller and more homogeneous (i.e., represent a population of students in the top 5% to 15% of their grade) compared with the statistics reported for the Explore technical manual (i.e., a population of eighth-grade students spanning the eighth-grade achievement spectrum). A larger high-achieving sample of students younger than the normative sample would be desirable to obtain more stable reliability and average SEM statistics for the high-achieving population.

Comparison of CTT and IRT Aptitude Estimates

Table 3 presents a comparison of the average questions answered correctly (i.e., number correct) and the average aptitude scores under the CTT and IRT frameworks, respectively. The table shows a performance increase across the grade levels and also shows better performance in the average number correct for English and Reading compared with Math and Science.

CTT scores show a nearly uniform increase—about two more items answered correctly—across the grade levels. The Math test had a larger increase between fifth and sixth grade, of about four more items answered correctly.

Multigroup IRT aptitude scores had a much larger increase in aptitude between fifth and sixth grades compared with moving from fourth to fifth grade. For example, aptitude estimates increased by less than a third of a standard deviation in English and Reading between fourth and fifth grade but increased by over half a standard deviation between fifth and sixth grades. The difference was even larger for the Math test; however, a similar growth across the three grade levels was seen in the Science test. Similar changes in the aptitude estimates for the single-group IRT model were found compared with the multigroup IRT model. The largest difference is in the location of the estimates because now the single-group IRT model assumes an average aptitude of zero when all grades are pooled together. As such, the fourth- and fifth-grade average aptitude scores tended to be negative followed by a positive sixth-grade average aptitude score. A Spearman rank correlation between

Table 3. Average Number Correct (i.e., Number of Items Answered Correctly or Raw Scores) for CTT and Average Aptitude Estimates for Both IRT Models, MG-IRT and SG-IRT.

	English			Math			Reading			Science		
	CTT	MG-IRT	SG-IRT	CTT	MG-IRT	SG-IRT	CTT	MG-IRT	SG-IRT	CTT	MG-IRT	SG-IRT
Grade 4	24.8 (6.46)	0.00 (0.92)	–0.41 (0.94)	11.6 (3.74)	0.00 (0.81)	–0.77 (0.64)	20.2 (5.22)	0.00 (0.91)	–0.46 (0.86)	13.7 (4.41)	0.00 (0.87)	–0.55 (0.77)
Grade 5	26.3 (5.62)	0.19 (0.81)	–0.19 (0.85)	13.5 (4.12)	0.55 (0.86)	–0.41 (0.74)	21.8 (4.95)	0.29 (0.96)	–0.20 (0.89)	15.7 (4.75)	0.60 (1.08)	–0.17 (0.86)
Grade 6	28.8 (5.55)	0.66 (0.86)	0.19 (0.87)	17.4 (4.61)	1.83 (1.09)	0.38 (0.82)	23.9 (4.31)	0.86 (0.98)	0.20 (0.85)	17.8 (4.59)	1.19 (1.14)	0.21 (0.87)

Note. English contains 40 items, Math and Reading contain 30 items, and Science contains 28 items. Standard deviations are reported in parentheses. IRT = item response theory; CTT = classical test theory; MG-IRT = multigroup IRT; SG-IRT = single-group IRT.

the average number correct in CTT and IRT aptitude scores ranged between .95 and .98 across subject areas tested and grade levels, suggesting a high level of agreement in the rank ordering of test-level scores between the two measurement frameworks.

Figure 1 displays the CTT eighth-grade percentile ranks on the above-level test for the high-achieving student sample (fourth-sixth graders) using violin plots. Violin plots show the distribution of scores through the density; areas that are wider reflect more data at that point compared with areas that are lower. The percentile ranks, which are based on the normative eighth-grade sample (ACT, 2013), are helpful for above-level purposes to understand if younger students are ready for more advanced or accelerated coursework. The percentile ranks in Figure 1 also highlight the difficulty of the I-Excel Math test, particularly for fourth and fifth graders. Overall, all students do well in Reading suggesting that reading proficiency, including comprehension, is not influencing performance on the other test areas.

Figure 2 depicts the distribution of aptitude estimates for each of the four subject areas tested, each grade level for the multigroup (left figure) IRT models. Compared with Figure 1 of percentile ranks referenced to the grade eight normative group, these distributions are more symmetric and the average increase in aptitude performance across the grades is evident. In addition to the symmetry for the IRT aptitude estimates, there was evidence in the Math test of a floor effect for fourth and fifth grade with CTT that was not present in the IRT aptitude estimates. A ceiling effect was also found for the sixth-grade Reading test with CTT, which was not as strong in the IRT aptitude estimates. The shifts in average performance are particularly large for the Math and Science subtests as was described and shown in Table 3. Variation was similar across the grade levels suggesting there are similar ranges of students within each grade; however, the sixth-grade Science scores have more variation with a longer lower tail of the distribution. This is further shown with the standard deviations reported for the aptitude scores in Table 3. The two different IRT models produce similar distributions of aptitude scores although the single-group IRT aptitude estimates had evidence of larger variation (single-group IRT aptitude estimates not shown graphically). The similarities are not surprising given the estimation used a normal prior distribution for both methods.

Differentiating Aptitude. Figures 3 and 4 show the distribution of aptitude scores for a single CTT raw score for Math and Science, respectively. An exploration of these figures shows that the IRT aptitude estimates can vary within a single CTT raw score and differs due to students answering different questions correctly/incorrectly (i.e.,

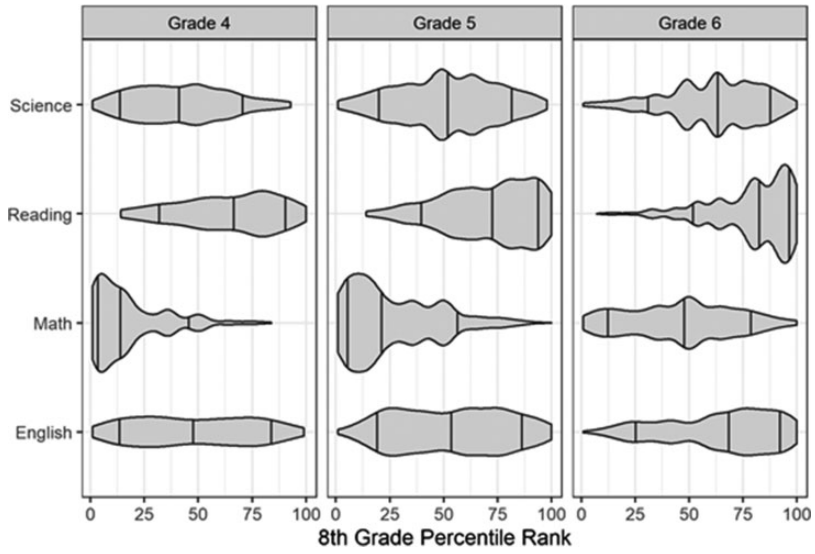

Figure 1. Violin plots of the eighth-grade percentile ranks for the four subject areas and grade level of above-level students with a CTT framework.
Note. The 10th, 50th, and 90th percentiles for the sample data are shown as vertical lines within the violin plots. CTT = classical test theory.

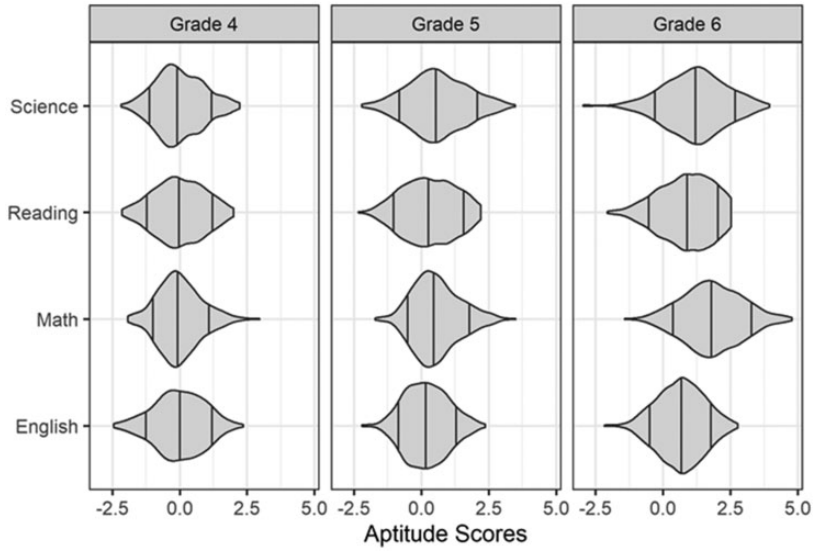

Figure 2. Violin plots of the aptitude scores for the four subject areas and grade level of above-level students estimated using multigroup IRT.
Note. The 10th, 50th, and 90th percentiles for the sample data are shown as vertical lines within the violin plots. Single-group IRT results are not shown as those were very similar to the multiple-group IRT results depicted. IRT = item response theory.

their unique response strings). The math aptitude estimates depicted are for students who answered half (15 of 30) of the items correctly on the Math test. There is more variation in fifth grade compared with the other two grades and, on average, students improve in their aptitude estimates as grade level increases. In other words, aptitude estimates improve, on average, when the grade level of the students approaches the grade level of the normative group. A similar figure showing science aptitude estimates is shown in Figure 4 for students who answered about 61% (17 out of 28) of the items correctly. Compared with the math aptitude estimates, the science estimates are more homogenous with slight increases as grade level increases.

These two figures also highlight differences in the multigroup (left figure) compared with single-group (right figure) IRT models. The multigroup IRT models were able to better differentiate across grade levels

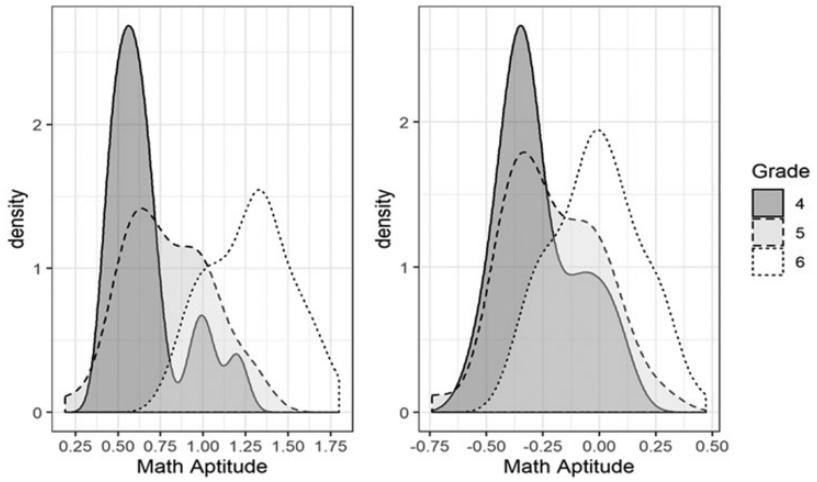

Figure 3. Distribution of math aptitude (i.e., theta or factor scores) scores by grade level for the same math raw score using multigroup IRT (left figure) compared with single-group IRT (right figure).
Note. Each *student* depicted correctly answered 15 of 30 math items. IRT = item response theory.

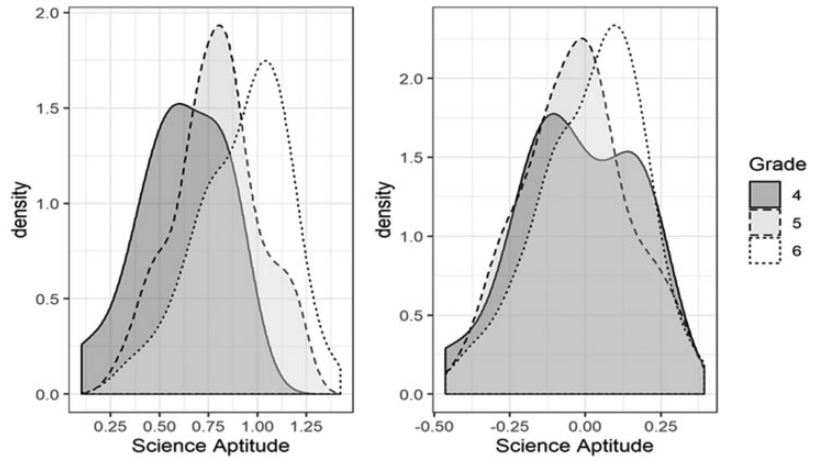

Figure 4. Distribution of science aptitude (i.e., theta or factor scores) by grade level for the same science raw score using multigroup IRT (left figure) compared with single-group IRT (right figure).
Note. Each *student* depicted correctly answered 17 of 28 science items. IRT = item response theory.

compared with the single-group IRT approach for both math and science (see Figures 3 and 4). However, the shape of each grade level is similar between the two approaches, which is not surprising given that the aptitude estimates are Bayesian and incorporate a normal prior distribution in their computation.

To more fully explore why there is such variability in the math aptitude estimates, response strings of two students in fifth grade with the same number correct scores are shown in Figure 5. In this figure, one student represents the highest IRT aptitude score and the other represents the lowest IRT aptitude scores for those fifth-grade students who answered 15 of 30 questions correctly. The aptitude estimates were around 0.2 and 1.3 for the lowest and highest student, respectively. Figure 5 shows their response strings, triangle shapes indicate correct responses and circles represent incorrect responses across the 30-item Math test. This figure then shows the aptitude estimate after the response for every item on the test. The figure illustrates that both students start at an ability of 0 and both answer the first question correctly, increasing their aptitude estimate to about 0.6. Item 2 is where the first divergence occurs. Student 1 answers the item incorrectly, whereas Student 2 answers the item correctly. Item two is highly discriminating and sufficiently difficult, resulting in Student 2's aptitude estimate increasing greatly to about 1.2, whereas Student 1's aptitude estimate decreases to about 0.2. Item 6 is another divergent point between these two students and increased the aptitude estimate of Student 2 but only slightly decreased the aptitude estimate of Student 1.

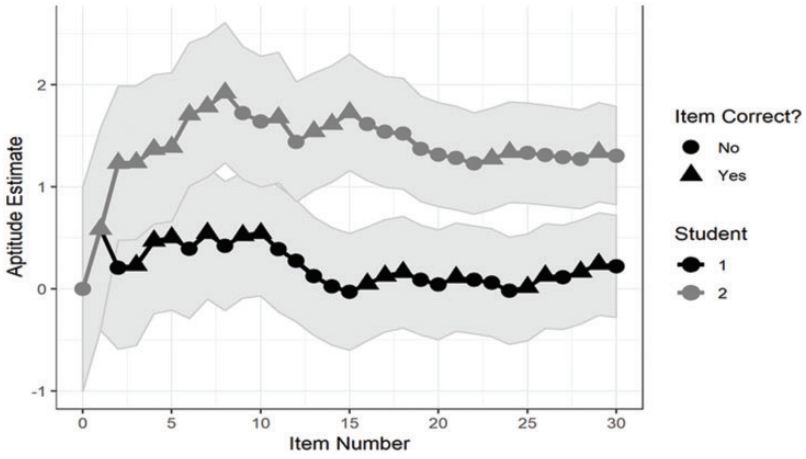

Figure 5. Estimated aptitude (i.e., theta or factor scores) for two fifth-grade students with the same number correct scores after answering each item using multigroup IRT.
Note. IRT = item response theory.

Following the path of the math aptitude estimates throughout, the 30-item test shows smaller adjustments to the aptitude estimate later in the test. These items are less discriminating at the end of the test and also represent less unique information given the previous items on the test. These items do help lower the standard error of the aptitude estimate shown in the figure with the grey shaded ribbon around each student. Early in the Math test, these regions overlap, suggesting that we cannot statistically differentiate the aptitude estimates between the two students. Later in the test, the ribbons do not overlap, suggesting that these two aptitude estimates are diverging from one another, even though the two students answered the same number of items correctly. This example shows that understanding which items the student answers correctly is highly relevant within an IRT framework and can provide different aptitude estimates for students who answered the same number of items correctly.

Classification Consistency

The classification consistency is shown in Table 4 for tests of Math, Reading, and Science. The English test was not explored due to not meeting the model fit criteria established earlier. Table 4 shows a two-by-two classification table within each grade level where the rows represent the classification using CTT methods (i.e., above the eighth-grade normative percentile) and the columns represent the IRT classification using the multigroup and single-group IRT methods. The consistency row of the table reflects the percentage of students in the No/No and Yes/Yes row/column combination in each grade and subject area. The More ID row of the table represents the percentage of students who are classified as potentially ready for above-level coursework in the IRT framework who were not classified as ready in the CTT framework (i.e., the No/Yes row/column combination). Finally, the Less ID row of the table represents those who were classified using CTT but subsequently not classified using IRT methods (i.e., the Yes/No row/column combination).

Multigroup IRT Classification Results. The Math test results show that classification consistency increases as grade level increases ranging from only 53% classified similarly in fourth grade up to 83% in sixth grade. Using IRT aptitude scores did classify additional students as above-level compared with CTT methods. This is most pronounced in fourth grade where 47% more students were classified as ready for above-level content. In sixth grade, an additional 17% of students were classified as ready for above-level content. Part of the decrease in additional students classified as potentially ready for above-level coursework is due to the increasing cut scores used for classification in the IRT context. The IRT framework builds in natural increases in student aptitude across grade levels; therefore, higher levels of aptitude are needed for a sixth-grade student to be classified as ready for above-level compared with a fourth-grade student.

In the CTT framework, the student scores are all being compared with the eighth-grade normative group; therefore, the same number correct, regardless of which items were answered correctly, will give students the same percentile. The Math test was the most difficult subject test; therefore, students in fourth and fifth grade were much less likely to meet the 50th percentile based on eighth-grade norms; however, when comparing their performance with their fourth- or fifth-grade high-achieving peers they were more likely to be identified. Finally, no students who were classified

Table 4. Comparison of Consistency of Identification of High-Achieving Students for Advanced Programming under CTT Compared With Multigroup and Single-Group IRT by Grade Level for Math, Reading, and Science.

	Multigroup IRT Classification						Single-group IRT Classification					
	Grade 4		Grade 5		Grade 6		Grade 4		Grade 5		Grade 6	
CTT	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
Math												
No	130	122	266	222	551	186	229	23	386	102	332	405
Yes	0	8	0	44	0	364	0	8	0	44	0	364
Consistency	53%		58%		83%		91%		81%		63%	
More ready (IRT > CTT)	47%		42%		17%		9%		19%		37%	
Less ready (IRT < CTT)	0%		0%		0%		0%		0%		0%	
Reading												
No	67	0	86	0	83	0	67	0	86	0	83	0
Yes	63	130	180	266	468	550	122	71	233	213	357	661
Consistency	76%		66%		58%		53%		56%		68%	
More ready (IRT > CTT)	0%		0%		0%		0%		0%		0%	
Less ready (IRT < CTT)	24%		34%		43%		47%		44%		32%	
Science												
No	130	62	259	42	398	1	190	2	294	7	380	19
Yes	0	68	7	224	153	549	10	58	30	201	43	659
Consistency	76%		91%		86%		95%		93%		94%	
More ready (IRT > CTT)	24%		8%		<1%		<1%		1%		2%	
Less ready (IRT < CTT)	0%		1%		14%		4%		6%		4%	

Note. Consistency refers to the percentage of cases where the same decision was made between CTT and IRT. More ready refers to the percentage of students that IRT classifies as ready for advanced content that CTT did not, whereas Less ready refers to the percentage of students that IRT classifies as not ready for advanced content that CTT did. CTT decision is shown in the first column, and IRT decision is shown in subsequent columns. IRT = item response theory; CTT = classical test theory.

as above-level with CTT methods were subsequently classified as not above-level using IRT aptitude scores.

Reading test classification results can also be seen in Table 4. The results for the Reading test were the opposite compared with the Math test classification results. The Reading test had the greatest classification consistency in fourth grade at 76% and decreased as grade level increased to a low of 58% for sixth-grade students. Students who were classified as ready for above-level content using CTT methods were subsequently not ready for above-level content using IRT aptitude estimates as shown by the Less ID row in Table 4. This "misclassification" was largest for sixth grade at 43% and decreased to 24% in fourth grade.

The Science test results have a higher level of classification consistency across the three grade levels, ranging from 76% to 91%. In many instances, these rates are higher than the math and reading subjects, particularly for fifth and sixth grade. Fourth grade had a large percentage of students (24%) who were classified as being ready for advanced science content using IRT but not CTT and had no students not classified using IRT who were classified with CTT. Fifth and sixth grade had fewer students classified using IRT compared with CTT (8% and <1%, respectively) but had increasing percentages of students classified with CTT methods but not IRT methods (1% and 14%, respectively). Science was the only subject where there were cases of students being in both the "More ID" and "Less ID" portions of the table.

Single-Group IRT Classification Results. The results for the single-group IRT classification followed somewhat different trends compared with the multigroup IRT results. The classification consistency for math was highest in fourth grade where most of the students were classified as not high aptitude. The classification consistency then decreased across the grades, with increasing percentages of students additionally identified compared with CTT. The classification consistency compared with CTT was higher in fourth and fifth grade compared with the multigroup IRT; however, fewer students in these grades were identified as ready for above-level content.

The Reading test had lower classification consistency for fourth and fifth grade compared with the multigroup IRT results. However, more students were not identified as ready for above-level content with the single-group IRT compared with the multigroup approach. This trend was the opposite for sixth grade as more students were classified similarly with fewer not identified as

ready for above-level reading content. Finally, the Science test results for the single-group IRT approach had high classification accuracy, suggesting a high degree of overlap in classification with CTT.

Discussion

The current study explored the aptitude estimates obtained from above-level testing of high-achieving fourth- to sixth-grade students by contrasting the CTT and IRT frameworks. Both frameworks provide useful information in the identification of students who may be ready for advanced or accelerated content as part of their curricula. However, the results from this study demonstrate how, compared with CTT estimates, using IRT can improve the differentiation of high-achieving students, especially when the students answer the same number of items, although different items, correctly (Research Question 1). The aptitude estimates in the IRT framework use the response strings from students rather than a singular focus on the number of items answered correctly (Research Question 2). As a result, when a student answers a difficult question correctly (i.e., one in which many students are unable to answer correctly), they receive a much larger boost to their aptitude score compared with answering a question that many students answer correctly.

The different aptitude scores for the same number of correct items may improve the classification accuracy of students as high aptitude or could, dependent on the content area, identify additional students who are likely ready for advanced content (Research Question 3). This was shown through the classification with the Math test (see Table 4). Expanding the talent search model through use of IRT has advantages for the identification of more students who are ready for more advanced or accelerated curricula in some content areas. However, educators are appropriately concerned with not frustrating students with test items that are too difficult (Lupkowski-Shoplik & Swiatek, 1999). Additional research is needed to validate this approach to relate it back to the decisions made regarding advanced content and provide further evidence to the appropriate percentile used for classification.

Moreover, the classification results and the impact of using IRT showed evidence of differences based on the subject. For example, the multigroup IRT methods identified more students as being potentially ready for advanced content for the Math test, particularly in fourth and fifth grade, compared with CTT and single-group IRT methods. Reading had the opposite results where the multigroup IRT model identified fewer students as ready for advanced content, particularly in sixth grade, compared with the other two methods. Some of these results may be due to the Math test being much more difficult for fourth and fifth graders and the Reading test being easier for sixth graders, which was shown to have floor (math) and ceiling (reading) effects in the CTT results (see Figure 1). The multigroup IRT model was able to differentiate these students better due to the unique response strings and also allowing for different cut scores to be identified for advanced content compared with the CTT and single-group IRT methods. Different cut scores would reflect the level of the student better than comparing all students with a single eighth-grade criterion.

It may also be the case that the different subjects have different development trajectories, which have implications for both IRT and CTT models of identifying latent aptitude among high achievers. For example, math content seems to build on subsequent topics, and it may be more difficult for students to complete items with content they have never seen before, hence the difficulty of the Math test for this sample of high-achieving fourth- and fifth-grade students. However, science may be somewhat different in that common foundational laws may be able to be applied in different contexts or the interpretation of scientific evidence (i.e., through tables or figures) may be a skill that can be more easily performed out of context. Future research could explore these topics, particularly how advanced content can or cannot be successfully completed without content exposure. Regardless, these results suggest that it is important to be mindful, in general, of the specific content area of identification and subsequent programming.

Related to the topic of classification, IRT also provides a developmental scale in which the latent aptitude estimates are on the same scale across grade levels, which allows for increased aptitude scores as grade level increases. This was done in the current study by using a multigroup IRT framework. This approach increases the aptitude cut score used to identify whether a high-achieving student has high aptitude as grade level increases. For example, a sixth-grade student would need a higher aptitude score (i.e., estimated latent theta score) compared with a fourth-grade student to be classified as high aptitude. Aligning the test items with the aptitude of a student at a specific grade level could improve efficiency of the test and allow for better discrimination of classification categories and ultimately better alignment with advanced content.

A single-group IRT approach, which would mimic the more traditional CTT approach with a single cutoff value across grades, was also considered. The classification consistency for this method was more similar to the CTT approach, but differences still arose. Much of these differences are likely attributed to CTT using a normative eighth-grade group in contrast to single-group IRT using the specific high-achieving sample to generate the cut score. As mentioned, single-group IRT

uses a prior distribution in the estimation of the latent aptitude. Similar to the discussion about the multigroup IRT results, differences in classification consistency mirror test difficulty; more students were identified with the single-group IRT model in math and fewer in reading. A high-achieving sample used for standard setting and evaluating the placement of an appropriate cut score to identify students as ready for above-level content in specific subjects would help provide validity evidence for that score. As described in the Limitations, the current study did not have data beyond those reported through the test itself to evaluate the validity of the cut scores chosen.

Limitations

The current study did not have information on decisions made regarding gifted identification and/or acceleration before or after the above-level testing was completed. Future research should measure the predictive validity of the IRT-based classifications to evaluate the impact of using different percentiles for classification on actual decisions made in the school regarding identification and acceleration. Future research should also assess the differential impact of the classification decisions on students' later achievement. These improved validation efforts would assist in understanding the best cut score to promote correct classification. Validation of the use of IRT could also confirm that the students who were identified or not identified as high-aptitude using IRT methods were accurately classified as ready or not ready for advanced or accelerated content and were matched with the appropriate learning opportunities. On the one hand, it is important to guard against incorrectly classifying students as ready for advanced or accelerated curricula because of the risk of frustrating students who are not prepared. On the other hand, it is important to identify students who would benefit from advanced content, especially students who are at risk of being underidentified because they attend underresourced schools.

Our sample was homogenous with respect to racial/ethnic identity and did not have individual-level data to measure socioeconomic status; thus, future research should replicate this work with more representative samples that could disaggregate outcomes by social identity groups. In other words, the performance on individual items could also be used to help predict the classification or to explore the contribution of individual items to the identification of students as high aptitude who might not typically be identified as high aptitude.

There was also evidence that the English subtest did not have sufficient model fit. Care needs to be taken when interpreting the aptitude scores from this model and additional structures such as a bifactor model could be explored.

This study used a 2pl IRT model fitted to the student response data. As discussed in the methods, this form was chosen due to sample size constraints (i.e., small sample sizes in fourth grade) and the desire to have a more accurate estimation of discrimination and difficulty item parameters. A 3pl IRT model can account for guessing for the questions and may offer better fit. When looking at the number of correct items for each of the tests, there was no strong evidence of guessing within the Math and Science tests, with some students answering only 5% of the questions correctly in each subject. The lowest raw score for English and Reading was 20% of questions answered correctly, and guessing may be an interpretation for this and may have improved model fit. Exploring the use of a three-parameter model would be another avenue for additional research.

Finally, the data were obtained from students nested within schools. We ignored this data structure in the current study due to small sample sizes for some of the schools. Future research could also explore the use of mixed IRT models that contain random effect parameters to adjust for school effects and the nested data.

Conclusions

The 1972 Marland Report intended to bring to the public's attention the need to identify and provide programming for gifted individuals in schools. By positing that gifted individuals would represent a minimum of 3% to 5% of the population, the precedent was cast for very exclusive programming in most schools, which is still the case today. While schools were establishing gifted programs, an out-of-school talent discovery and development model, the talent search model (Stanley, 2005), was forming. The talent search model relies on above-level tests in particular content areas such as math to determine readiness for advanced curriculum in specific subjects. The Talent Search Model has relied on CTT for understanding and interpretation of scores earned by high-achieving students when they take an above-level test. Although CTT has been a valued approach to understanding the performance of high-achieving students on above-level tests, applying IRT refines the score interpretation by differentiating high-achieving students across the aptitude score scale and accounting explicitly for average grade level differences.

A major purpose of utilizing the talent search model in schools is to be more inclusive of bright students who may be underrepresented in traditional gifted education programs (Plucker et al., 2010; Plucker et al., 2018) and for whom talent search programs (Lee et al., 2008) and the associated programming may be inaccessible. By extending the above-level testing model into schools and improving the information from above-level testing, in particular latent aptitude estimates and classifications,

greater numbers of students may participate in the identification process. Inclusiveness on these two levels has the potential to enhance the overall academic experience by increasing access to advanced opportunities (Assouline et al., 2017; Plucker & Harris, 2015).

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Brandon LeBeau https://orcid.org/0000-0002-1265-8761
Ann Lupkowski-Shoplik https://orcid.org/0000-0001-7512-3679

References

ACT. (2013). *ACT explore* (Assessment instrument). Author.

Assouline, S. G., Ihrig, L. M., & Mahatmya, D. (2017). Closing the excellence gap: Investigation of an expanded talent search model for student selection into an extracurricular STEM program in rural middle schools. *Gifted Child Quarterly*, *61*(3), 250–261. https://doi.org/10.1177/0016986217701833

Assouline, S. G., & Lupkowski-Shoplik, A. E. (2012). The talent search model of gifted identification. *Journal of Psychoeducational Assessment*, *30*(1), 45–59. https://doi.org/10.1177/0734282911433946

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, *46*(4), 443–459. https://doi.org/10.1007/BF02293801

Brody, L. E., & Mills, C. J. (2005). Talent search research: What have we learned? *High Ability Studies*, *16*(1), 97–111. https://doi.org/10.1080/13598130500115320

Callahan, C. M., Moon, T. R., & Oh, S. (2017). Describing the status of programs for the gifted: A call for action. *Journal for the Education of the Gifted*, *40*(1), 20–49. https://doi.org/10.1177/0162353216686215

Calvert, E. (2018). Identification and assessment in a K-12 talent development framework. In P. Olszewski-Kubilius, R. F. Subotnik, & F. C. Worrell (Eds.), *Talent development as a framework for gifted education: Implications for best practices and applications in schools* (pp. 25-42). Prufrock Press.

Chalmers, P. R. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*(6), 1–29. https://doi.org/10.18637/jss.v048.i06

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

Hambleton, R. K., & Jones, R. W. (1993). An NCME instructional module on: Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, *12*(3), 38–47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Hu, L.-t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. http://doi.org/10.1080/10705519909540118

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling, and linking: Methods and practices*. Springer.

Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice*, *29*(3), 8–14. https://doi.org/10.1111/j.1745-3992.2010.00179.x

Lakin, J. M. (2018). Making the cut in gifted selection: Score combination rules and their impact on program diversity. *Gifted Child Quarterly*, *62*(2), 210–219. https://doi.org/10.1177/0016986217752099

Lee, S., Matthews, M. S., & Olszewski-Kubilius, P. (2008). A national picture of talent search and talent search educational programs. *Gifted Child Quarterly*, *52*(1), 55–69. https://doi.org/10.1177/0016986207311152

Lohman, D. F. (2005). An aptitude perspective on talent: Implications for identification of academically gifted minority students. *Journal for the Education of the Gifted*, *28*(3-4), 333–360. https://doi.org/10.4219/jeg-2005-341

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Routledge.

Lupkowski-Shoplik, A. E., & Assouline, S. G. (1993). Identifying mathematically talented elementary students: Using the lower level of the SSAT. *Gifted Child Quarterly*, *37*(3), 118–123. https://doi.org/10.1177/001698629303700304

Lupkowski-Shoplik, A. E., Benbow, C., Assouline, S., & Brody, L. (2003). Talent searches: Meeting the needs of academically talented youth. In N. Colangelo & G. Davis (Eds.), *Handbook of gifted education* (pp. 204-218). Allyn & Bacon.

Lupkowski-Shoplik, A. E., & Swiatek, M. A. (1999). Elementary student talent searches: Establishing appropriate guidelines for qualifying test scores. *Gifted Child Quarterly*, *43*(4), 265–272. https://doi.org/10.1177/001698629904300405

Marland, S. P., Jr. (1972). Education of the gifted and talented: Report to the Congress of the United States by the U.S. Commissioner of Education and background papers submitted to the U.S. Office of Education (Government Documents, Y4.L 11/2: G36, 2 vols.). U.S. Government Printing Office.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

McBee, M. T., Peters, S. J., & Miller, E. N. (2016). The impact of the nomination stage on gifted program identification: A comprehensive psychometric analysis. *Gifted Child Quarterly*, *60*(4), 258–278. https://doi.org/10.1177/0016986216656256

McBee, M. T., Peters, S. J., & Waterman, C. (2014). Combining scores in multiple-criteria assessment systems: The impact of combination rule. *Gifted Child Quarterly*, *58*(1), 69–89. https://doi.org/10.1177%2F0016986213513794

McDonald, R. P. (2011). *Test theory: A unified treatment*. Routledge.

No Child Left Behind (NCLB). (2003). Act of 2001, 20 U.S.C. A. § 6301 et seq.

Olszewski-Kubilius, P. (2015). Talent searches and accelerated programming for gifted students. In S. G. Assouline, N. Colangelo, J. VanTassel-Baska, & A. Lupkowski-Shoplik (Eds.), *A nation empowered: Evidence trumps the excuses holding back America's brightest students* (Vol. 2, pp. 111-121). Belin-Blank Center for Gifted and Talented Education.

Olszewski-Kubilius, P., & Lee, S.-Y. (2005). How schools use talent search scores for gifted adolescents. *Roeper Review*, *27*(4), 233–240. https://doi.org/10.1080/02783190509554324

Peters, S. J., Rambo-Hernandez, K., Makel, M. C., Matthews, M. S., & Plucker, J. A. (2017). Should millions of students take a gap year? Large numbers of students start the school year above grade level. *Gifted Child Quarterly*, *61*(3), 229–238. https://doi.org/10.1177/0016986217701834

Peters, S. J., Rambo-Hernandez, K., Makel, M. C., Matthews, M. S., & Plucker, J. A. (2019). Effect of local norms on racial and ethnic representation in gifted education. *AERA Open*, *5*(2), Article 446. https://doi.org/10.1177/2332858419848446

Plucker, J. A., Burroughs, N., & Song, R. (2010). *Mind the (other) gap! The growing excellence gap in K-12 education*. https://files.eric.ed.gov/fulltext/ED531840.pdf

Plucker, J. A., Glynn, J., Healey, G., & Dettmer, A. (2018). *Equal talents, unequal opportunities, Second edition: A report card on state support for academically talented low-income students*. Jack Kent Cooke Foundation.

Plucker, J. A., & Harris, B. (2015). Acceleration and economically vulnerable children. In S. G. Assouline, N. Colangelo, J. Van Tassel-Baska, & A. E. Lupkowski-Shoplik (Eds.), *A nation empowered: Evidence trumps the excuses holding back America's brightest students* (pp. 181–188). The Belin-Blank Center for Gifted and Talented Education.

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement*, *31*(3), 169–180. https://doi.org/10.1177/0146621606291569

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory* (pp. 79–112). Springer.

Rupp, A. A., & Zumbo, B. D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement*, *66*(1), 63–84. https://doi.org/10.1177%2F0013164404273942

Shabani, K., Khatib, M., & Ebadi, S. (2010). Vgotsky's zone of proximal development: Instructional implications and teachers' professional development. *English Language Teaching*, *3*(4), 237–248. https://doi.org/10.5539/elt.v3n4p237

Stanley, J. C. (2000). Helping students learn only what they don't already know. *Psychology, Public Policy, and Law*, *6*(1), 216–222. https://doi.org/10.1037/1076-8971.6.1.216

Stanley, J. C. (2005). A quiet revolution: Finding boys and girls who reason exceptionally well mathematically and/or verbally and helping them get the supplemental educational opportunities they need. *High Ability Studies*, *16*(1), 5–14. https://doi.org/10.1080/13598130500115114

Warne, R. T. (2012). History and development of above-level testing of the gifted. *Roeper Review*, *34*(3), 183–193. https://doi.org/10.1080/02783193.2012.686425

Author Biographies

Brandon LeBeau is an assistant professor of educational measurement and statistics. His research is at the intersection of research software development, statistical methodology, and applied statistical analyses focusing on evaluation of program effectiveness. In particular, he has expertise in longitudinal data analysis with mixed models and using item response theory to create developmental scales to more accurately explore growth across a lengthy developmental span. His software development includes R packages for extracting text data from PDF documents and simulating data within a linear mixed model framework (simglm).

Susan G. Assouline is professor of School Psychology and Director of the Belin-Blank Center and holds the Myron and Jacqueline N. Blank Endowed Chair in Gifted Education. Throughout her career, she has been interested in twice-exceptionality, acceleration, and identification of academic talent in elementary students. In 2015, she coedited with Nicholas Colangelo, Joyce VanTassel-Baska, and Ann Lupkowski-Shoplik in *A Nation Empowered: Evidence Trumps the Excuses Holding Back America's Brightest Students*. In 2016, she received the National Association for Gifted Children 2016 Distinguished Scholar Award; in 2018, she was recognized by the UI with the Award for Faculty Excellence.

Duhita Mahatmya, PhD, is a research methodologist for the College of Education. She completed her doctorate in Human Development and Family Studies from Iowa State University, with an emphasis on adolescent development and research methods. She received her bachelor's degrees in Psychology and English from Drake University. In her research, she utilizes a systems approach to examine family, school, and neighborhood factors that shape inequities in students' academic and psychosocial development. She has also led and managed multiple projects assessing K-16 student and faculty development, with particular attention to the experiences of individuals from marginalized social identity groups.

Ann Lupkowski-Shoplik, PhD, is the administrator for the Acceleration Institute and Research at the

University of Iowa Belin-Blank Center. She founded and directed the Carnegie Mellon Institute for Talented Elementary Students (C-MITES) at Carnegie Mellon University for 22 years. She coauthored *Developing Math Talent: A Comprehensive Guide to Math Education for Gifted Students in Elementary and Middle School* (2nd ed.), and the Iowa Acceleration Scale, and coedited *A Nation Empowered: Evidence Trumps the Excuses Holding Back America's Brightest Students*. She recently coauthored *Developing Academic Acceleration Policies: Whole Grade, Early Entrance, and Single Subject* with Wendy A. Behrens and Susan G. Assouline. Her professional interests include identifying exceptionally mathematically talented students and devising appropriately challenging opportunities for them as well as assisting educators in understanding the talent search model and how it can be utilized in schools.

Manuscript received: December 6, 2019; Final revision received: April 13, 2020; Accepted: April 14, 2020