


The Impact of the Nomination Stage on Gifted Program Identification: A Comprehensive Psychometric Analysis

Gifted Child Quarterly
1–21
© 2016 National Association for
Gifted Children
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0016986216656256
gcq.sagepub.com


Matthew T. McBee¹, Scott J. Peters², and Erin M. Miller³

Abstract

The use of the nomination stage as the first step in the identification process is pervasive across the field of gifted education. In many cases, nominations are used to limit the number of students who will need to be evaluated using costly and time-consuming assessments for the purpose of gifted program identification and placement. This study evaluated the effect of the nomination stage on the overall efficacy of a gifted identification system. Results showed that in nearly all conditions, identification systems that require a nomination before testing result in a large proportion of gifted students being missed. Under commonly implemented conditions, the nomination stage can cause the false negative rate to easily exceed 60%. Changes to identification practices are urgently needed in order to ensure that larger numbers of gifted students receive appropriate educational placement and to maintain the integrity of gifted education services.

Keywords

identification, nomination, psychometrics, sensitivity, ROC

Introduction

Assessment holds fundamental importance for many educational fields and gifted education is no exception. The information gained through assessment is vital to making effective decisions regarding a child's educational experience, including the decision of whether or not that child should receive gifted education services. Quality assessment is often expensive and time-consuming, requiring resources of money and time that schools almost always have in short supply. The conflict between the competing goals of achieving high-quality assessment data while simultaneously preserving scarce resources has led to the adoption of two-stage assessment systems both within and outside of education. Despite the fact that the qualities assessed during the gifted identification process exist on a continuous scale (such as level of readiness or need), the resulting decision is dichotomous. Regardless of what criteria are used, which assessments are administered, or which students are tested, in the end the decision comes down to which students receive a particular service and which do not. Even if a school offers several different types of gifted education services, some criteria must be set to decide which students are placed in each of those services and which will receive no services.

Multistage Diagnostic Systems

The idea of a two-stage diagnostic system is simple. All members of the population take a quick and inexpensive

Stage-I assessment or screening test. Those that test positive on the Stage-I screening assessment go on to take the Stage-II confirmatory assessment. The Stage-II assessment is generally more expensive, more time consuming, and more invasive but it also possesses much higher psychometric quality (meaning excellent sensitivity and specificity) than the Stage-I assessment. The final decision is made on the basis of the Stage-II assessment, hence it is often called the "confirmation assessment." Figure 1 shows a flowchart describing the two-stage screening process.

The two-stage process is often used in medicine. For example, it is very common for women to receive regular mammograms (the Stage-I assessment) to screen for possible breast cancer (National Cancer Institute, 2014). Those women with suspicious or problematic mammograms are referred for additional imaging or biopsy of the suspected tumor (the Stage-II assessment) in order to provide definitive evidence to begin a treatment plan.¹ Two-phase systems are common across many domains. In psychology, a common screener is the Mini-Mental State Examination (Folstein, Folstein, & McHugh, 1975), which is given to screen for

¹East Tennessee State University, Johnson City, TN, USA

²University of Wisconsin–Whitewater, WI, USA

³Bridgewater College, Bridgewater, VA, USA

Corresponding Author:

Matthew T. McBee, East Tennessee State University, 413 Rogers Stout, Johnson City, TN 37614, USA.

Email: mcbeem@etsu.edu

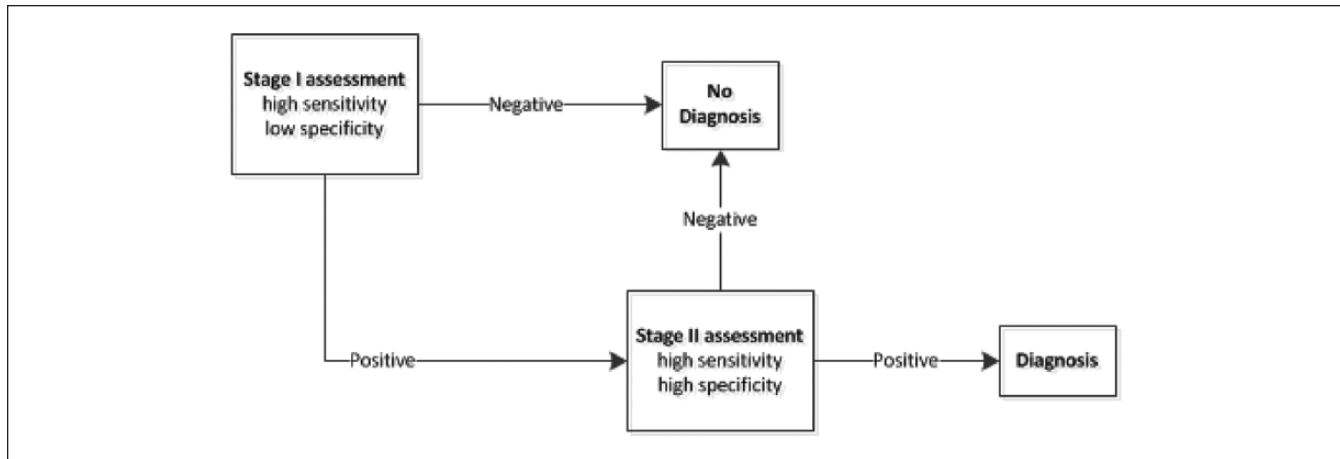


Figure 1. The medical model for screening tests.

dementia in mental health evaluations. During the identification of specific learning disabilities, the initiation of a comprehensive identification process begins with teacher referral (Zumeta, Zirkel, & Danielson, 2014).

Similarly, in gifted education, it is common for the identification process to take place in two stages (National Association for Gifted Children [NAGC], 2013). The first stage is commonly known as the “nomination” stage, in which a group of potentially gifted students, or even an entire student population, is screened for further consideration. This is typically done on the basis of teacher or parent nominations, but in some settings can involve automatic nominations on the basis of high achievement test scores, parent nominations, peer nominations, or student self-nominations (McBee, 2006). In other settings, the nomination stage consists of teachers completing a standardized checklist of gifted behaviors for each student in the class (e.g., Peters & Gentry, 2010; Pfeiffer & Jarosewich, 2003). Those with scores above some threshold on the Stage-I assessment will be eligible to undergo the additional testing required to be identified for a particular gifted and talented intervention.

Diagnostic tests, whether in education or in medicine, sometimes result in incorrect decisions called false positives and false negatives. In medicine, a false positive error implies that the test indicates that a healthy patient has the condition. In gifted education, a false positive error could result in a student being placed in a gifted education program even though she does not meet the criteria, indicating that the student might not benefit from the program. Likewise, false negative errors occur when a sick patient is classified as healthy by the test, or when a “truly” gifted student² is missed by the test, and therefore does not receive the needed curricular modifications and services. Two fundamental metrics describe the performance of diagnostic tests. Sensitivity describes the proportion of “true positives” that receive a positive test result. Specificity describes the proportion of “true negatives” that receive a negative test result. A test with

high sensitivity will produce a low false negative rate, where a test with high specificity will produce a low false positive rate. These metrics will be discussed more extensively later in this article.

Nissen-Meyer (1964) suggested that the purpose of a screening phase is to divide an entire population into two groups: Those who are in need of further evaluation because they might have a certain condition and those who, based on a certain level of confidence, can be assumed to not have the condition and therefore exempted from further testing. The screening phase is meant to solely separate a population into two groups—those for whom additional assessment is warranted and those for whom it is unnecessary. The goal of the nomination stage in gifted education is to avoid the needless expenditure of assessment resources on students who are unlikely to need any additional services offered under the umbrella of gifted education. Accordingly, any test or process to be used as a screener needs to possess certain characteristics. It must be quick and simple to implement, relatively inexpensive, and not overly onerous to the administrator or the person being screened (Nissen-Meyer, 1964; Peters & Gentry, 2013). Above all, screening tests require very high sensitivity, because any false negatives that occur at this stage will not have a chance to be corrected—if a student who needs gifted services is not passed through the screening or nomination phase to the Phase-II confirmatory assessment, she has no chance of being served. If these criteria are met, an optimal balance is achieved between efficacy and efficiency. Qaseem et al. (2012) argued that a major benefit of judicious use of screening phases is a significant decrease in the cost of identification. Rather than using diagnostic evaluations on the entire population—some of which are very expensive and/or uncomfortable for the person being tested—shorter, more logistically and physically palatable screeners can be used first to pare down the number of people who need to be tested using the confirmation assessments. With gifted education already funded at very low

levels in most states (NAGC, 2013), it is highly desirable and in most cases, necessary, to limit assessment costs as much as possible.

Screening in Gifted Education

In the 2012-2013 *State of the States of Gifted Education Report* (NAGC, 2013), 33 states responded to a question regarding when students were typically identified for gifted and talented services. Although multiple responses to this question were allowed, the two most common were following a teacher referral (20 states) and following a parent referral (19 states). These teacher referrals vary widely in formality and could consist of anything from a generic teacher recommendation for testing to a decision based on the responses to a formal teacher-rating instrument. There exists an entire genre of teacher rating scales (see Hoge & Cudmore, 1986; Peters & Gentry, 2010) that can act as population-level screeners before further diagnostic assessment is implemented. Both the *State of the States Report* and a 2013 report from the National Research Center on the Gifted and Talented (Callahan, Moon, & Oh, 2013) noted that “parent or teacher nomination or referral is still a common entry point in the identification process at the elementary school level” (p. 14) where all students at a particular grade are screened using a nomination phase before then being tested further. While the exact extent of screening or two-stage identification systems is not known, it is clear that it is extremely common and likely influences the pathway through which the majority of students in the United States are identified as needing additional advanced educational services.

The benefits of screening phases are only realized if the performance and efficacy of the assessment system as a whole is not severely degraded by the screener’s inclusion. In other words, saving money and time is a false economy if large numbers of students who need services are missed because a low-quality screener placed them in the group that did not need the full diagnostic testing (false negatives). This concern is shared by most fields that use two-phase systems including psychology and special education, particularly when the screener is a classroom teacher (Madelaine & Wheldall, 2005; VanDerHeyden, Witt, & Naquin, 2003; Zumeta et al., 2014). As we show later in this article, screeners can only reduce the sensitivity of the integrated assessment system. The inclusion of a screening or nomination phase can only result in more gifted students being missed than if the screener was skipped and all students in the population were given the confirmatory test(s), regardless of the quality of the screening tool or procedure used. Despite this fact, screeners may be justified on a cost-benefit basis. Well-designed screening systems sacrifice minimal sensitivity in return for large reductions in the cost and time devoted to assessment. This tradeoff should always be considered when assessing the value of a particular screening system. Screeners do result in a reduction of the false positive rate since some students who would otherwise

achieve a qualifying score on the confirmatory assessment via a positive measurement error will not receive a nomination, but this improvement is typically small relative to the potentially large detriment in sensitivity as we will show later in this article.

This article investigated how the nomination stage and its characteristics affect the quality and cost of gifted identification and had the following goals:

1. To describe how psychometric features of the nomination stage, including the reliability, validity, and cutoff, impact the performance of two-stage gifted identification systems.
2. To estimate the “typical” efficacy of the two-stage identification system commonly used to identify gifted students.
3. To describe how to select nomination cutoffs that result in the desired compromise between lost sensitivity and cost savings.

Method

Background Framework

Because all test scores contain measurement error, decisions made on the basis of a test are sometimes wrong. Classical test theory (CTT) provides a useful means of analysis, as it conceptualizes all observed test scores as consisting of a student’s *true score* plus measurement error. The following application of CTT is based on the outline of the model by Crocker and Algina (1986) and Lord (1980). The true score is what we want to measure. It is the construct of interest (e.g., giftedness, intelligence, school readiness). Error represents “noise” in the measurement and includes everything that affects the observed test score besides the construct of interest. With respect to gifted program identification, the true score defines who *is* gifted, the observed score defined who is *identified* as gifted, and the error causes these to be different. Our goal in assessment is to always try and decrease error as much as possible in order to obtain a purer measure of the construct of interest.

The CTT model can be written as

$$X = T + E \quad (1)$$

where

$$E \sim N(0, \sigma^2) \quad (2)$$

In the CTT model, X represents the observed score, T the underlying true score, and E the measurement error. The measurement errors are conceptualized as random draws from a normal distribution with a mean of zero. Because the mean of the measurement errors are zero, the observed test scores are unbiased estimates of the true scores when averaged across many hypothetical measurement occasions.

In other words, the observed scores will not tend to be systematically too high or too low. Because measurement errors are assumed to follow a normal distribution, small errors are much more common than large errors. The variance of the normal distribution for the measurement errors (σ^2) depends on the reliability of the assessment. High reliability implies a smaller ratio of measurement error variance to true score variance, so the observed scores will tend to be very close to the true scores. Low reliability means that the ratio of measurement error variance to true score variance is larger, so the observed scores are on average quite different from the true scores. More formally, reliability (ρ_{XX}) is the proportion of true score variance in the observed scores (whose variance is the sum of the true score variance and error variance) and can be written as

$$\rho_{XX} = \frac{\text{Var}(T)}{\text{Var}(T) + \text{Var}(E)} \quad (3)$$

This is equivalent to describing the reliability as the squared correlation between the true scores and the observed scores because the squared correlation coefficient is interpreted as the proportion of variance shared between a pair of variables. In a perfect world, 100% of the variance in an observed test score would be explained by the level of the construct present in the test taker with no influence of measurement error. In a gifted education identification setting, this would mean a student's identification as "gifted" (observed score) would depend solely on whether or not that student is actually gifted. However, in practice this is never the case since no test is completely error free.

Assessing the Performance of Identification Systems

As previously described, the performance of any identification system can be quantified using four metrics. First is *sensitivity*, the proportion of truly gifted students who are admitted to the program. Second is the *false positive rate*, the proportion of identified students who are not actually gifted. The false positive rate is the complement of *specificity*, which tends to be very high in even poor-performing gifted identification systems because only a small proportion of students are typically identified. For that reason, our analysis focuses on the related concept of the *incorrect identification rate*, which is the probability that a non-gifted student is identified (McBee, Peters, & Waterman, 2014). Finally, the *false negative rate* is the proportion of truly gifted students who are not identified and is the complement of sensitivity. Figure 2 provides a graphical representation of the possible results of an identification system.

The upper right quadrant includes those students who were identified based on their observed score and should have been because their true score is above the set criteria.



Figure 2. The possible results of an identification system.

These students represent true positives because we wanted to find them and we were successful in doing so. The bottom left quadrant is the largest in gifted education because it represents the students who we did not want to identify (because their true score was below the criteria), and they were successfully labeled as not being gifted (based on their observed score). The other two quadrants represent false positives and false negatives—cases in which the decision made based on observed scores is different from that which would be made if we knew students' true scores. These students are the primary focus of this article.

Identification System Performance With No Nomination Stage. As referenced earlier, screeners cannot improve the sensitivity of an assessment system; they can only reduce it because of their position as an additional hurdle for gifted students to clear. Because of this fact, the important question to consider when evaluating the impact of a screening phase is just how badly it degrades the overall efficacy of the identification system. The particulars of a given screening procedure determine whether the detrimental impact of the screener on system performance is negligible or extreme. Therefore, in order to evaluate empirically the effect of adding a nomination phase, we must first consider the performance of a single-assessment identification system, sans screener, to serve as a point of comparison. This imaginary scenario of universal screening with a high-quality confirmation assessment serves as a best-case scenario since the addition of a screening phase can never improve the quality of the overall system.

Although many states and school districts now use some form of multiple-criteria assessment to determine which students qualify for gifted education programs, here we consider a simple single-assessment system. We choose a single-assessment system because multiple-criteria systems can make use of a number of different methods for combining scores, such as the "and," "or," or "mean" rules, or even

Table 1. Performance of a Single-Assessment System by Cutoff and Reliability.

Cutoff	Reliability	Sensitivity	False negative	Incorrect ID rate
90th percentile (z = 1.280)	1.00	1.000	.000	.000
	.95	.843	.157	.157
	.90	.776	.224	.224
	.85	.725	.275	.275
	.80	.680	.320	.320
	.75	.641	.359	.359
	.70	.604	.396	.396
95th percentile (z = 1.645)	1.00	1.000	.000	.000
	.95	.815	.185	.185
	.90	.738	.262	.262
	.85	.679	.321	.321
	.80	.628	.372	.372
	.75	.582	.418	.418
	.70	.541	.459	.459
99th percentile (z = 2.330)	1.00	1.000	.000	.000
	.95	.763	.237	.237
	.90	.665	.335	.335
	.85	.592	.408	.408
	.80	.530	.470	.470
	.75	.476	.524	.524
	.70	.428	.572	.572

Note. Bolded row indicates maximum performance for a plausible identification instrument. Table adapted from McBee et al. (2014).

complex combinations of rules. Each of these rules brings considerable complexity that is for the moment, irrelevant (McBee et al., 2014).

Our hypothetical single-assessment system is one in which gifted program placement decisions are made on the basis of a lone measure which will be administered to each student in the population—such as an entire school or grade level. A cutoff score is set, often at the 90th, 95th, or 99th percentile, to translate the continuous scores generated by the test into a binary placement decision (identified or not identified as gifted). This is the simplest possible identification system. The psychometric performance of such a system is governed by the reliability of the test and the cutoff for program entry, with higher reliability and lower cutoffs leading to better sensitivity and lower incorrect identification rates and false negative rates. Table 1 (taken from McBee et al., 2014) illustrates the performance of single-assessment systems at varying cutoffs and reliabilities.

Assuming that the cutoff is held constant (i.e., the performance standard considered to be indicative of giftedness does not change), the system performance depends only on the reliability of the test. The optimal performance of a single-assessment system under optimistic, but potentially realistic, conditions (i.e., the reliability $\rho_{xx} = .95$) is that only about 84% of qualifying students are actually admitted to the program. This is as efficacious as a single-assessment system can get in the real world because it is extremely difficult to design instruments that can achieve reliabilities of greater

than .95. In fact, most identification instruments do not achieve reliabilities that high. We use this 84% sensitivity as a baseline when considering the degree of the negative impact of adding a screener to the system.

Parameters That Influence Nomination Stage Performance. Our analysis in this section is driven by the following example. A school identifies students for its gifted program using a two-stage system. First, teachers are asked to nominate the students in their classes whom they believe to be potentially gifted and in need of formal assessment to determine program eligibility. Next, the nominated students are formally tested via a single-assessment system or a single composite score derived from multiple assessments. Those students who score higher than the 90th percentile on the assessment are admitted to the program. Alternately, the nomination procedure can require that teachers complete a formal teacher rating scale on each student rather than the teacher simply advancing a “list of names.” In this case, students whose teacher ratings exceed some cutoff are able to advance to the confirmation phase. In either case, the process and resulting methods listed below are the same.

A few key parameters determine how well such a system functions. In addition to the reliability and the cutoff of the confirmatory testing stage, there are now three additional parameters: (1) the reliability of the nomination, (2) the validity of the nomination, and (3) the nomination cutoff. We next consider each in turn.

Nomination reliability. If the nomination consists of a published teacher rating scale, the meaning of this term is clear and the reliability can be estimated from the data or approximated using information published in the rating scale's test manual (which we do later on for some popular teacher rating scales). However, if the nomination stage is an informal process consisting of teachers supplying a list of names of students who they are recommending for further testing, we frame the term "reliability" based on a cognitive model of how teachers decide whom to nominate. In this model, the teacher's decisions are based on that individual teacher's experiences and preconceptions rather than research and formal guidelines. This understanding of "reliability" is in concordance with research-based theories about categorization decisions (Murphy & Medin, 1985; Nosofsky & Palmeri, 1997; Rosch, 1978). We assume that teachers have an implicit theory of giftedness (which may or may not be valid) that they use to tacitly "score" each student in the class. In other words, teachers implicitly assign to each child a number representing the favorability of a comparison between that child's behaviors and perceived abilities versus the teacher's implicit theory of what "gifted" means and looks like (Miller, 2009). The "reliability" of the teacher nomination represents the *precision* of this judgment, or the degree to which small gradations of differences in perceived standing are related to corresponding responses in the teacher's subjective assessment of the students. Because we view this subjective judgment as a kind of latent continuous variable, we model it using the classic test theory concept of reliability—the proportion of variance in the teacher's determination that is related to true differences between students rather than random error. For the purposes of our analysis, whether the nomination decision is based on a realized score on a formal nomination instrument or an implicit comparison against a subjective standard, nomination reliability means essentially the same thing and can be analyzed using the same psychometric model.

Nomination validity. In the case of rating scales, validity simply refers to the correlation between the rating scale score and the score on the confirmation assessment (or assessments, in the case of multiple criteria). For generic teacher nominations, validity represents the degree of correspondence between the teacher's implicit theory of giftedness and the operational definition of giftedness to be imposed in the confirmation assessment phase. If the confirmation assessment is a single intelligence test, then the nomination validity represents the degree to which the teacher's subjective judgment is responsive to the same skills and abilities that the confirmation assessment(s) measures. For example, if the teacher's implicit theory of giftedness involves students with excellent memory, ability to detect patterns, large vocabulary, and so on, and the confirmation test is the Stanford-Binet Intelligence Test (Roid, 2003), then that teacher's nomination validity is high. However, if the teacher instead

nominates the most creative students in the class (something not measured by typical intelligence or achievement tests), the nomination validity is low. We represent the nomination validity as a correlation between the teacher's latent "scores" (or ratings) for each student with respect to his or her own implicit theory of giftedness and the score that child will receive on the confirmation assessment(s).

Nomination cutoff. When published rating scales are used, the cutoff is often explicitly specified by school district or state policy (e.g., 90th percentile). For informal nominations, the nomination cutoff represents how "gifted" the student must be perceived by the teacher to be in order to earn a nomination. If a teacher nominates students that he perceives as being slightly above average with respect to his implicit theory of giftedness, that teacher has a low nomination cutoff. A teacher who only nominates the "once in a generation" excellent student has a very high nomination cutoff. For modeling purposes, the cutoff represents the *z*-score on the latent variable, which is required before students may advance to the formal assessment phase. This operationalization is consistent with current psychological theories regarding implicit conceptions and categorization (Kruschke, 1992; Medin & Schaffer, 1978; Nilsson, Juslin, & Olsson, 2008; Nosofsky, 1986).

Performance Analysis

The strategy for analyzing the performance of an identification system is based on the calculation of conditional probabilities. For example, the definition of sensitivity is the probability of being identified conditional on being gifted. The conditional probability of event *a* given *b* is calculated by dividing the joint probability of *a* and *b* by the marginal probability of *b*. In other words, $p(\text{identified}|\text{gifted}) = p(\text{identified} \& \text{gifted})/p(\text{gifted})$. If we assume that the nomination and confirmation test scores follow a multivariate normal distribution, the required joint and marginal probabilities can be calculated by numerically integrating the multivariate normal density. We performed these calculations using the R command `sadmvn` included in the library `mnormt`. We are considering here an identification system with a single nomination and a single assessment, but according to classic test theory each has both an observed score and an underlying true score. Thus, there are four variables in the system to consider. Just as the normal distribution is governed by its mean and variance, the multivariate normal distribution is governed by a mean vector, which provides the mean of each constituent variable, and the covariance matrix, which provides information on the variances of each variable and its correlations with the other variables in the system. The construction of the covariance matrix begins by specifying *a priori* the following parameters: the nomination reliability (ρ_{oo}), the test reliability (ρ_{tt}), and the validity coefficient ($r_{n,t}$), where the validity coefficient represents

the correlation between *observed* scores. It is necessary to apply the correction for attenuation (Crocker & Algina, 1986) formula to calculate the corresponding correlations between true scores and observed scores as well as true scores and true scores. If the variables are to be on a z-score metric, the diagonal elements of the matrix should all be one, implying that the variance (and standard deviation) of each variable is one, and that the off-diagonal covariance elements are equal to correlations. Letting the order of variables be nomination true score (n_{true}), nomination observed score (n_{obs}), confirmatory test true score (c_{true}), and confirmatory test observed score (c_{obs}), the variance-covariance matrix is as follows.

$$Cov(n_{true}, n_{obs}, c_{true}, c_{obs}) = \begin{bmatrix} 1 & \sqrt{\rho_{nn}} & \frac{r_{n_{obs}, c_{obs}}}{\sqrt{\rho_{nn}\rho_{cc}}} & \frac{r_{n_{obs}, c_{obs}}}{\sqrt{\rho_{nn}}} \\ \sqrt{\rho_{nn}} & 1 & \frac{r_{n_{obs}, c_{obs}}}{\sqrt{\rho_{cc}}} & r_{n_{obs}, c_{obs}} \\ \frac{r_{n_{obs}, c_{obs}}}{\sqrt{\rho_{nn}\rho_{cc}}} & \frac{r_{n_{obs}, c_{obs}}}{\sqrt{\rho_{nn}\rho_{cc}}} & 1 & \sqrt{\rho_{cc}} \\ \frac{r_{n_{obs}, c_{obs}}}{\sqrt{\rho_{nn}}} & r_{n_{obs}, c_{obs}} & \sqrt{\rho_{cc}} & 1 \end{bmatrix} \quad (4)$$

The values in this covariance matrix bear some explanation. The correlation between the observed nomination score and the observed confirmatory test score is $r_{n_{obs}, c_{obs}}$. The reliability coefficient for the nomination is ρ_{nn} ; the covariance (correlation) between the nomination observed score and the nomination true score is therefore the square root of the reliability. The reliability coefficient for the confirmatory test is ρ_{cc} . The remaining covariances involve one or more true scores. The correlation between observed scores is suppressed by measurement error. Therefore, the covariances involving true scores must be corrected for attenuation (Crocker & Algina, 1986). The disattenuated correlation coefficient is calculated by dividing the observed correlation by the square root of the product of the reliabilities of the involved variables. When one of the involved variables is a true score, its reliability appears in the denominator.

After specifying this covariance matrix and mean vector of zeros, the conditional probabilities can be computed by numerical integration of the multivariate normal distribution. By fixing the mean vector to zero and the variances to one, the variables exist on a familiar z-score metric. It then becomes quite convenient to identify cutoff values corresponding to common percentiles required for gifted program entry (i.e., 90th percentile) by referring to a table of normal curve areas such as can be found in the appendix of most any introductory statistics textbook (e.g., Aron, Coups, & Aron,

2013). For example, the z-score cutoff corresponding with the 90th percentile is $z = 1.28$.

Letting the order of variables be as before, sensitivity can be calculated as follows:

$$\text{sensitivity} = \frac{p(\text{gifted, identified})}{p(\text{gifted})} = \frac{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} N_4(\mu, \Sigma) d_{c_{obs}} d_{c_{true}} d_{n_{obs}} d_{n_{true}}}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} N_4(\mu, \Sigma) d_{c_{obs}} d_{c_{true}} d_{n_{obs}} d_{n_{true}}} \quad (5)$$

where τ is the test cutoff z-score, ν is the nomination cutoff z-score, μ is a 4×1 mean vector of zeros, and Σ is a 4×4 variance-covariance matrix as specified in Equation 4.

The logic of this is as follows: a gifted individual is operationally defined as one for whom the confirmation test true score (c_{true}) on the test is above the cutoff (τ). The nomination true score (n_{true}), confirmation test true score (c_{true}), and nomination observed scores (n_{obs}) are irrelevant. The integral in the denominator represents this idea. These quantities are integrated from negative to positive infinity, whereas the test true score is integrated from τ to positive infinity. This integral, found in the denominator of the sensitivity equation, computes the proportion of students who would be expected to have true scores above the cutoff (based on whatever cutoff is selected as the criteria for “gifted”), without requiring that they satisfy any other criterion. The numerator of the sensitivity equation computes the proportion of students that will be true positives, meaning that they are gifted and identified. Being gifted requires that the true test score (c_{true}) be above τ . Being identified requires that the observed nomination score (n_{obs}) be above ν and that the observed test score (c_{obs}) also be above τ ; the former constraint requires that gifted students receive a nomination while the latter that the student receives a qualifying observed score on the confirmatory test. These constraints appear as the limits of integration. Similarly, the incorrect identification rate can be calculated as

$$\text{incorrect identification rate} = \frac{p(\text{identified, not gifted})}{p(\text{identified})} = \frac{\int_{-\infty}^{\infty} \int_{\nu}^{\infty} \int_{-\infty}^{\tau} \int_{-\infty}^{\infty} N_4(\mu, \Sigma) d_{c_{obs}} d_{c_{true}} d_{n_{obs}} d_{n_{true}}}{\int_{-\infty}^{\infty} \int_{\nu}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} N_4(\mu, \Sigma) d_{c_{obs}} d_{c_{true}} d_{n_{obs}} d_{n_{true}}} \quad (6)$$

because an incorrect identification means that the student receives a nomination and a qualifying observed confirmation

test score but has a true test score below the cutoff (τ). The positive predictive value (PPV) can be calculated as

$$PPV = 1 - \text{incorrect identification rate} \quad (7)$$

and the proportion of students identified as

$$p(\text{identified}) = \int_{-\infty}^{\infty} \int_{\nu}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\tau} N_4(\boldsymbol{\mu}, \boldsymbol{\Sigma}) d_{c_{obs}} d_{c_{me}} d_{n_{obs}} d_{n_{me}} \quad (8)$$

Our analysis was performed by specifying the values of the nomination reliability (ρ_{nn}), nomination validity ($r_{n_{obs}, c_{obs}}$), and confirmatory test reliability (ρ_{cc}). We could then compute the values necessary to populate the variance-covariance matrix ($\boldsymbol{\Sigma}$, Equation 4). Next we specified the nomination cutoff (ν) and test cutoff (τ), and then used R's `sadmvn` function to approximate the values of the integrals in Equations 5, 6, and 8. We fixed the confirmatory test reliability to .95 (again, ambitious but also achievable) and the test cutoff to the 90th percentile ($z = 1.28$) for all our computations. Our analysis program looped through ranges of plausible values for the nomination reliability (ρ_{nn}), the nomination validity, ($r_{n_{obs}, c_{obs}}$), and the nomination cutoff (ν), allowing the creation of graphs displaying the influence of these three parameters on the integrated identification system's performance.

Results

Effects of Nomination Reliability, Validity, and Cutoff on System Performance

Figure 3 summarizes how the system sensitivity responds to change in the three parameters discussed above. In this figure, the parameters of the testing phase are set to their most optimistic plausible values with test reliability of .95 and a cutoff at the 90th percentile. According to Table 1 and as referenced earlier, with the test reliability fixed at .95 and the test cutoff set at the 90th percentile, the sensitivity of an identification system with no screener is .843. This value is the maximum possible performance and is represented in the figure by dashed horizontal lines. Figure 3 presents five panels, one for each level of nomination reliability (representing different levels of measurement error present in the screening phase). Nomination validity is plotted on the x-axis, integrated system sensitivity on the y-axis, and the values for the nomination cutoff are represented as separate lines.

The reliability of an assessment limits its maximum correlation with other variables (overall system validity). Since the system validity coefficient represents the correlation between the (observed) nomination "scores" and the confirmation observed test scores, low reliability for nominations puts an upper limit on the overall system validity coefficient (Crocker & Algina, 1986). This is reflected in the figure, as

the plots do not extend across the full horizontal range of validity when reliability is low. The figure reveals that sensitivity is mostly a function of the nomination validity and the nomination cutoff. Nomination reliability itself had little impact other than by constraining the validity coefficient. Nomination validity and the nomination cutoff clearly interact. If the nomination cutoff is set low enough, perhaps at the 50th percentile, sensitivity remains at reasonable levels even when validity is moderately poor. In such a situation enough students pass through the nomination stage that they are then "found" by the confirmation phase, but this comes at the cost of having to test a much larger number of students. Few students are missed due to failure to pass through the nomination stage because the nomination cutoff is set at a relatively low level (50th percentile). As is shown next, the same cannot be said when nomination cutoffs are set at more typical levels.

High nomination cutoffs, in which the standard for nomination approaches the confirmation test threshold (e.g., at the 90th percentile), lead to extremely low sensitivity. In this case, the psychometric quality of the confirmatory assessment hardly matters because so few students who are capable of meeting the cutoff for that phase would ever be tested. We find this alarming as we believe that teachers often only nominate those students that they perceive to be "truly" or extremely gifted (Carman, 2011; Neumeister, Adams, Pierce, Cassady, & Dixon, 2007; Peterson, 1999; Plata, Masten, & Trusty, 1999; Siegle & Powell, 2004). In this case, the system sensitivity can easily fall into the 30% range, meaning that 70% of qualifying students—those who the system was explicitly designed to find—do not gain access to the program. Increased validity allows the system to tolerate increasingly stringent nomination cutoffs—the more accurate the assessments, the higher the cutoffs can be while still "catching" the right students. Even at the highest validity, however, a nomination cutoff equal to the identification cutoff (at the 90th percentile for both) leads to unacceptably poor sensitivity.

Next, we considered the impact of a screener on the system incorrect identification rate: the proportion of identified students who do not actually qualify (with respect to their true score). In a single-assessment system with no nomination phase, the incorrect identification rate and the false negative rate are balanced (as can be seen from Table 1). This is no longer the case when a nomination phase is added. As shown in Figure 4, the nomination validity and cutoff again prove to be the key parameters in affecting the incorrect classification rate, with higher cutoffs and low validity both resulting in fewer incorrect identifications, but at the expense of more false negatives.

While reducing incorrect identification rate is always in itself desirable, we think it is more important to consider the tradeoff between incorrect identifications and false negatives. Decreasing the incorrect identification rate necessarily increases the false negative rate. The ideal balance between

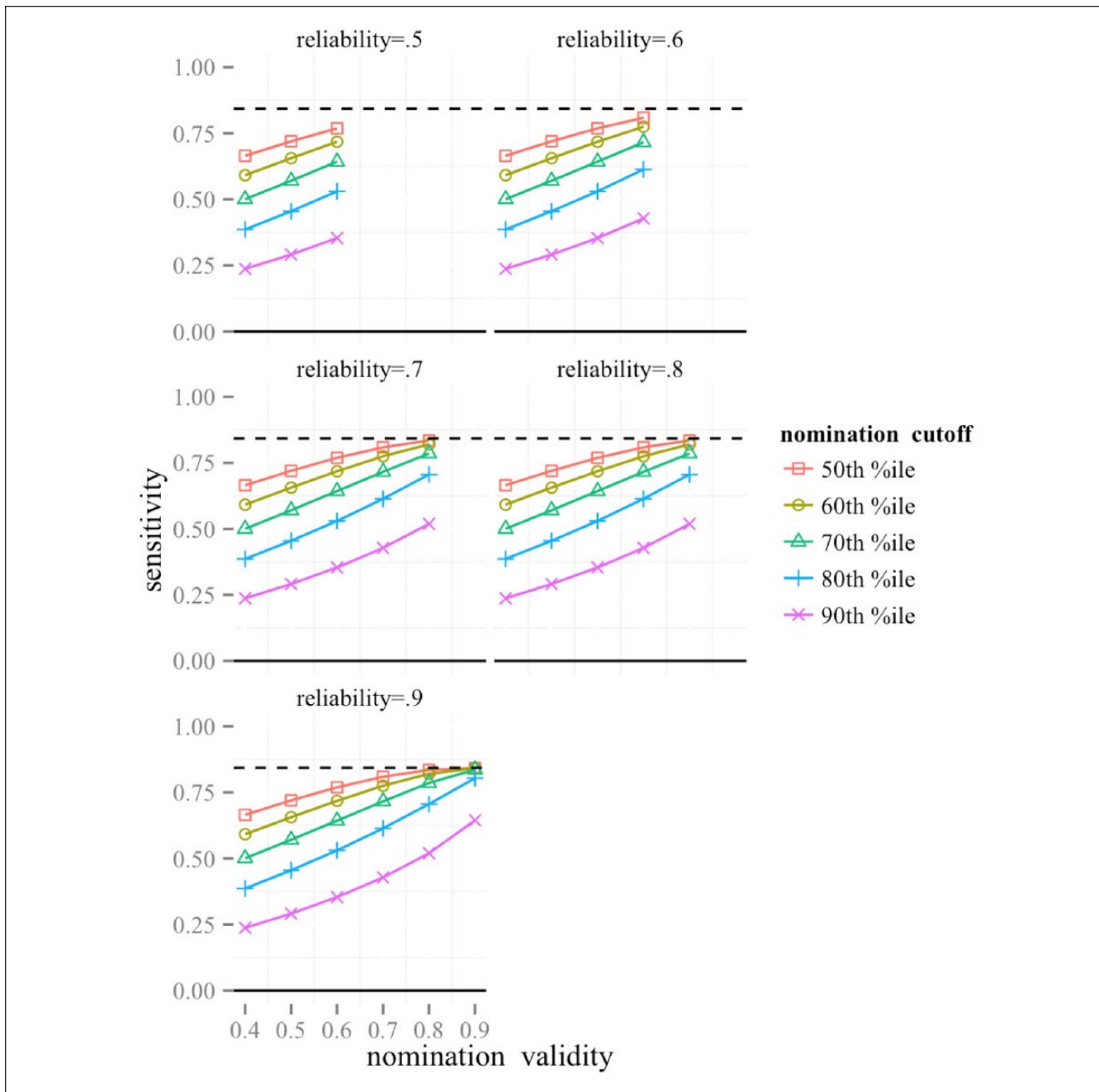


Figure 3. Sensitivity of a two-stage, single-assessment system by nomination reliability, validity, and cutoff.
 Note. Test cutoff set to 90th percentile and test reliability set to .95. Horizontal dotted lines indicate system performance with no screener.

incorrect identifications and false negatives depends on the specific nature of the educational programming to follow a successful identification and particularly on the consequences of failing to thrive in the program (Peters, Matthews, McBee, & McCoach, 2014). For example, a child’s unsuccessful attempt at an after-school enrichment program may have minimal negative consequences, whereas a child who cannot perform after skipping a grade may require a logistically and socially fraught process of reintegration into her

old classroom placement. In the former situation, it would be wise to design an identification system that tends to favor incorrect identifications rather than false negatives, whereas in the latter, incorrect identifications are very problematic. Therefore there is no globally correct decision with regard to incorrect identifications and false negatives; one must take into account the program to be provided and its potential for negative outcomes as a result of an incorrect placement in order to decide which is preferable.

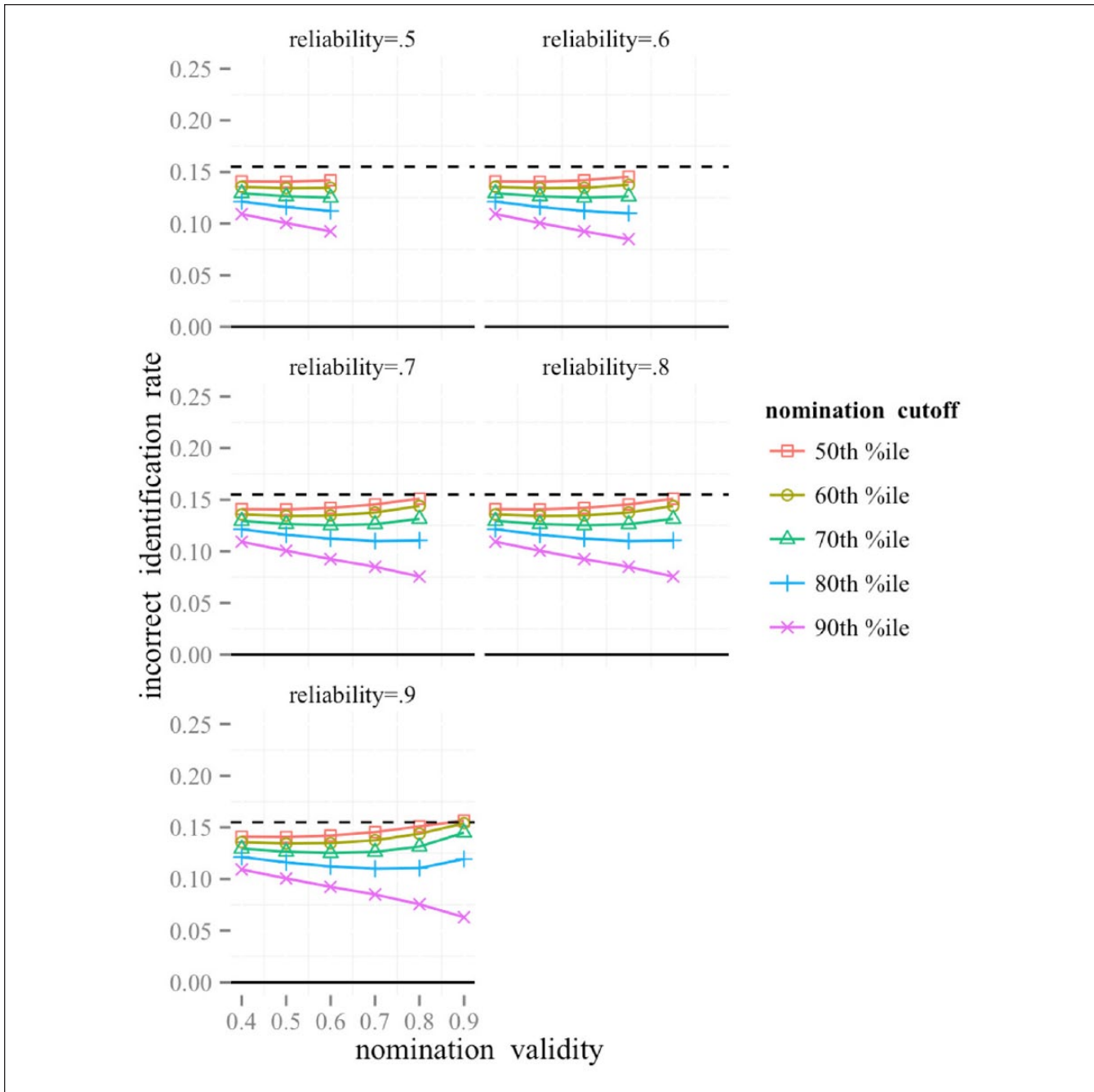


Figure 4. Incorrect identification rates for a two-stage, single-assessment system by nomination reliability, validity, and cutoff. Note. Test cutoff set to 90th percentile and test reliability set to .95. Horizontal dotted lines indicate system performance with no screener.

Finally, we considered the impact of the nomination stage on the proportion of children identified for the program. In a single-assessment system, the expected program size is a function of the cutoff only. If the cutoff is set at the 90th percentile, then the expected proportion of students qualifying for the program should be roughly 10%, assuming that the school resembles the norm group for the test or that local norms are used as the point of comparison. Figure 5 illustrates

the consequences of adding a nomination stage with varying reliability, validity, and cutoff.

As can be seen from the figure, the story remains the same. High cutoffs at the nomination stage can strongly limit the program size (below what it is seemingly intended, based on the criteria for “gifted” being set at the top 10% in this example). Setting a high nomination cutoff means that some of those students who would have met the 90th percentile

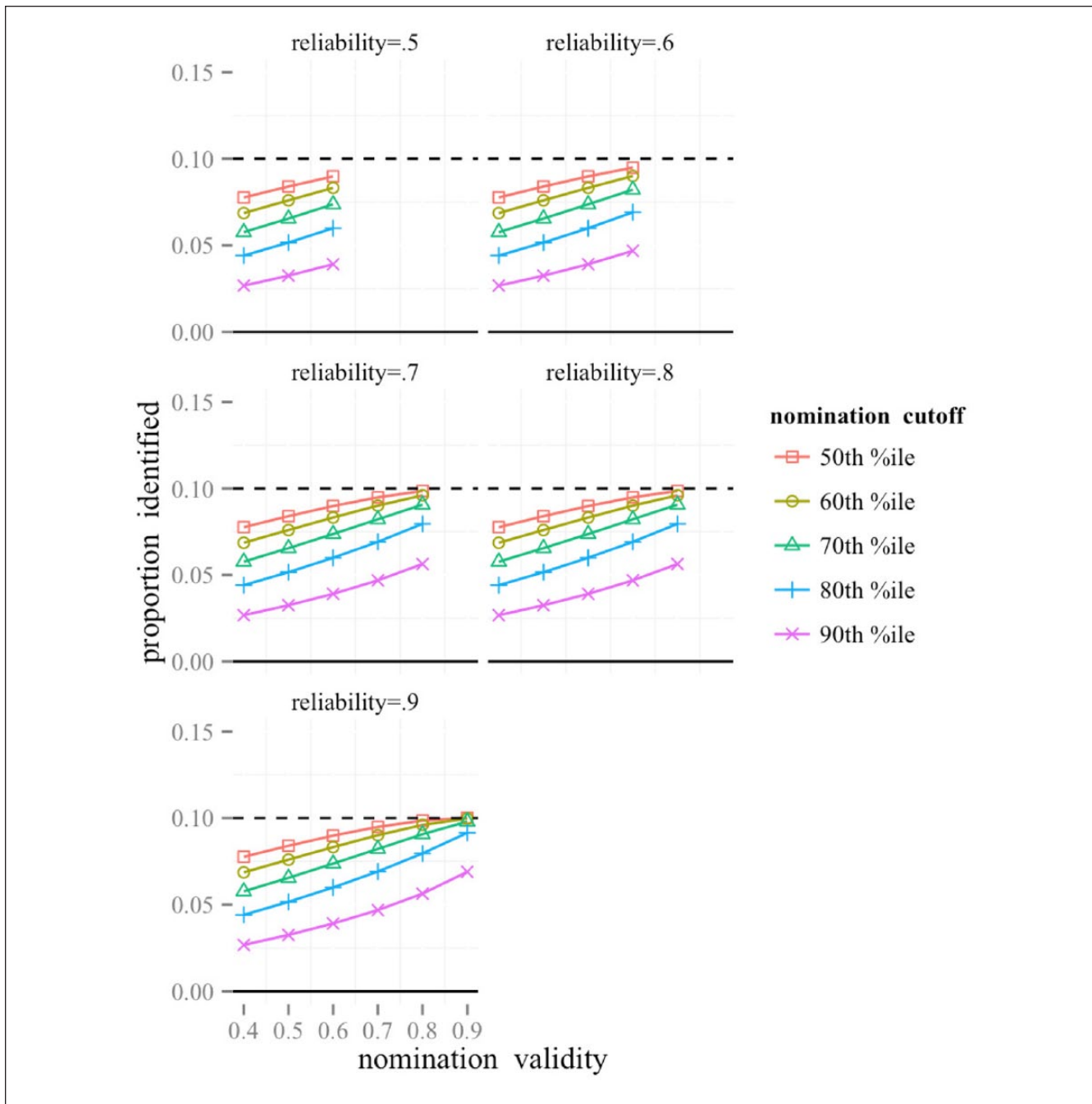


Figure 5. Proportion of students identified in a two-stage, single-assessment system by nomination reliability, validity, and cutoff. Note. Test cutoff set to 90th percentile and test reliability set to .95. Horizontal dotted lines indicate system performance with no screener.

criteria on the confirmation assessment are never given a chance to take the test. This means they are never identified and thus the size of the resulting “gifted” population is decreased from what it should be. Furthermore, Figure 5 shows that the degree of program size reduction at a given nomination cutoff is also inversely proportional to the nomination validity. This demonstrates that the nomination phase can significantly decrease the size of the identified population

in addition to decreasing the overall efficacy of the identification system.

Realistic Estimates of System Performance

Given the above theoretical analysis, how well does the typical identification system, as implemented in thousands of schools across the United States, actually function in finding

the students it sets out to find? Here we consider such a question using evidence of reliability and validity from published teacher rating scales as they might be used in a nomination phase. Informal teacher nominations are probably much worse than this but there is no way to know for sure. The validity of implicit theories is mediated by experience (Medin & Schaffer, 1978). As teachers only rarely receive training in gifted education (NAGC, 2013), their implicit theories of giftedness are unlikely to have close correspondence with the school or district's operational definition, leading to low system validity. Research on the performance of classroom teachers as referral agents indicates there is a large degree of variability in the validity of teacher recommendations (Carman, 2011; Hunsaker, Finley, & Frank, 1997; McBee, 2006; Miller, 2009; Siegle, Moore, Mann, & Wilson, 2010). In most cases, we suspect that the nomination cutoffs in general are much higher than their optimal levels. This is a direct consequence of requesting that teachers nominate those students whom they suspect of being "gifted" rather than those students who are above average. Table 2 contains validity information for the *Scales for Identifying Gifted Students* (SIGS; Ryser & McConnell, 2004) rating scale. We chose this scale because we believe it has a high-quality research base and because its test manual provides an extensive set of reliability and validity information.

For the following analysis, we set the SIGS reliability to .95, as this is a rough average for the alpha reliability at the elementary level (Ryser & McConnell, 2004). We set the reliability of the confirmatory test to .95 and its cutoff to 90%. We use the published correlations between the SIGS General Intellectual Ability subscale and the Wechsler Intelligence Scale for Children, 3rd Edition (WISC-III; $r = .67$), Test of Cognitive Skills, 2nd Edition (TCS-2; $r = .73$), and Cognitive Ability Test (CogAT; $r = .48$) as the validity coefficients (Ryser & McConnell, 2004). We can then calculate the performance of many potential identification systems in which the SIGS General Intellectual Ability subscale is used at the nomination instrument with various cutoffs. The results are displayed in Table 3.

The performance analysis indicates that when nomination cutoffs are high, the overall system performance can only be described as unacceptable, particularly when the WISC-III or CogAT are the confirmation stage instruments, as around 60% to 70% of the qualified students are missed when high cutoffs are used with those instruments. For example, if the confirmatory assessment was the CogAT, and the nomination cutoff on the SIGS general intellectual ability subscale was set to the 90th percentile, the integrated system sensitivity would be .28. Only 28% of the "truly" gifted students² would be identified. This sensitivity should be contrasted against the 84% sensitivity that would have been achieved if the screener had not been used and instead the CogAT had been given to all students; the relative integrated system sensitivity that would result would be only 33% of its optimal value. The TCS-2 as a confirmatory assessment following the SIGS

Table 2. Validity Coefficients (Correlations) of SIGS Subscales With Measures of Academic Ability.

SIGS subscale	WISC-III	TCS-2	CogAT
General Intellectual Ability	.67	.73	.48
Language Arts	.62	.72	—
Mathematics	.47	.88	—
Science	.56	.89	—
Social Studies	.53	.83	—
Creativity	.38	.86	—
Leadership	.53	.84	—

Note. SIGS = Scales for Identifying Gifted Students; WISC-III = Wechsler Intelligence Scale for Children, 3rd Edition; TCS-2 = Test of Cognitive Skills, 2nd Edition; CogAT = Cognitive Ability Test. Table adapted from Ryser and McConnell (2004).

fared somewhat better because its validity coefficient is higher, but still does not approach a reasonable level of performance with high nomination cutoffs. When a 90th percentile cutoff is used for the TCS-2, the integrated system sensitivity is 45%, implying that system sensitivity is roughly halved as a result of the inclusion of the screener. In either case, the addition of a screening phase has devastated the quality of the resulting gifted identification decisions. In both cases the identification system failed to find the majority of students it set out to find.

This analysis has been limited to systems with a single confirmatory test. However, many school districts have now adopted multiple criteria assessment systems. When multiple scores are combined using "and" rules (McBee et al., 2014), the lowest correlation between the nomination and the instruments in the assessment package sets an upper bound on system performance. Many multiple criteria systems, such as Georgia's, involve measures of creativity, which are often poorly correlated with scores on gifted rating scales because they are inherently looking for different things. For example, Table 4 provides correlations between the *Gifted Rating Scales–School Form* (GRS; Pfeiffer & Jarosewich, 2003) subscales and the Torrance Test of Creative Thinking–Figural. Again, we chose the GRS in this case because it too has a strong research base and is likely of higher quality than many other published rating scales. Though we do not provide an analysis of the consequences of these low validities, based on Figure 3, it should be clear that the psychometric performance of any identification system involving the use of the GRS or any of its subscales as a nomination instrument for an assessment system that includes the Torrance Test of Creative Thinking in an "and" rule could only be described as abysmal.

Setting Optimal Nomination Cutoffs

In the previous section, we described how severely the performance of a gifted identification system can be damaged when a nomination stage is added with a cutoff higher than

Table 3. Performance of Systems Using SIGS General Intellectual Ability Subscale for Nomination by Nomination Cutoff and by Phase-2 Assessment.

Instrument	Nomination cutoff	Sensitivity	False negative rate	Incorrect ID rate	Proportion identified
WISC-III (<i>r</i> = .67)	50th percentile	.798	.202	.144	.094
	60th percentile	.759	.241	.136	.088
	70th percentile	.694	.306	.126	.080
	80th percentile	.588	.412	.110	.066
	90th percentile	.404	.596	.087	.044
TCS-2 (<i>r</i> = .73)	50th percentile	.818	.182	.147	.096
	60th percentile	.790	.210	.139	.092
	70th percentile	.738	.262	.127	.085
	80th percentile	.640	.360	.110	.072
	90th percentile	.453	.547	.082	.049
CogAT (<i>r</i> = .48)	50th percentile	.709	.291	.141	.083
	60th percentile	.643	.357	.134	.075
	70th percentile	.556	.444	.127	.064
	80th percentile	.441	.559	.117	.050
	90th percentile	.280	.720	.102	.031

Note. SIGS = Scales for Identifying Gifted Students; WISC-III = Wechsler Intelligence Scale for Children, 3rd Edition; TCS-2 = Test of Cognitive Skills, 2nd Edition; CogAT = Cognitive Ability Test. Assumed reliability of .95 and cutoff at 90th percentile for each instrument. Validity coefficients for each instrument were estimated from data in Table 2.

Table 4. Validity Coefficients (Correlations) for GRS-S Subscales With TTCT, Figural.

GRS-S subscale	TTCT Figural
Intellectual	0.19
Academic	0.21
Creativity	0.14
Artistic	0.15
Leadership	0.14
Motivation	0.27

Note. GRS-S = Gifted Rating Scales–School Form; TTCT = Torrance Test of Creative Thinking. Table adapted from Pfeiffer and Jarosewich (2003).

can be supported by its validity coefficient. We also provided some evidence that this is probably the usual state of affairs in gifted education. In this section we describe how the use of a nomination stage could be improved in order to make it successful at decreasing the cost and time of assessment while having only a minimal negative impact on the system sensitivity.

Background. In order to perform a rigorous analysis of classification tests, we assume that an individual can be in one of two states, such as being gifted or not being gifted. That individual takes an assessment that yields a continuous score. A cutoff value is specified, such that scores on the assessment above the cutoff are taken to indicate that the condition (e.g., giftedness) is present while a score at or below the cutoff indicates that the condition is absent. The choice of the cutoff directly affects the test’s sensitivity and specificity. For example, if the cutoff is set at the lowest possible score that

the assessment can generate, all test takers will be classified as having the condition. Therefore there will be no false negatives, and sensitivity will be a perfect 100%. However, the test will generate a huge number of false positives. In fact, because everyone will be classified as having the condition, the false positive rate will be 100%. Using the assessment’s highest possible score as the cutoff will have the opposite effect. No one will be classified as having the condition, so the false positive rate will be 0%, but sensitivity will also be 0%. The test’s sensitivity and false positive rate can be computed for every choice of cutoff between these extremes. This information can be summarized by a plot called the receiver operating characteristics (ROC) curve, which plots sensitivity on the y-axis against the false positive rate on the x-axis. An example ROC curve is provided in Figure 6.

The diagonal line on the ROC curve is the line of no discrimination; it represents what the ROC curve would look like for a completely noninformative test. The better the test, the further the ROC curve will deviate from that line. Each point on the ROC curve represents the sensitivity and false positive rate that would result from a different choice for the cutoff score. The optimal cutoff score is the one that results in the best compromise between sensitivity and the false positive rate, and can be identified as the point that is most distant from the line of no discrimination. The psychometric quality of the test (i.e., reliability and validity) determines the shape of the ROC curve, with higher-quality tests having ROC curves that curve sharply away from the line of zero information.

Application to Two-Stage Identification Systems. A modified version of ROC curve analysis can be used to better understand

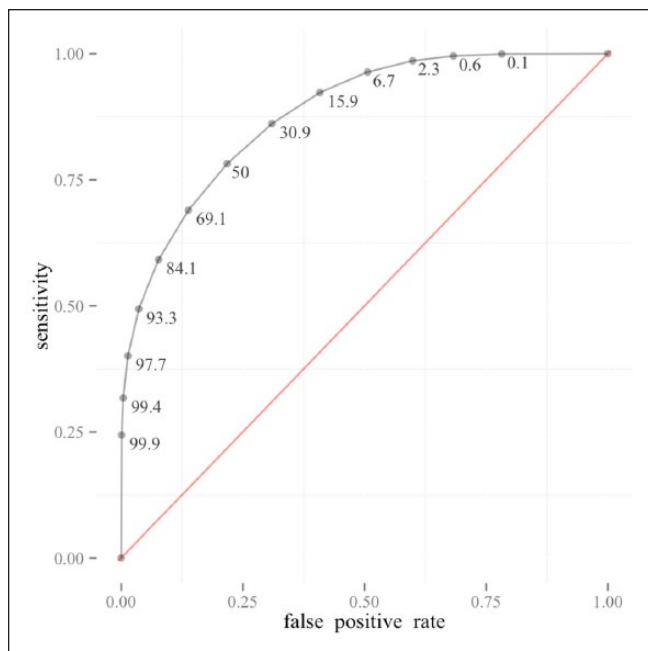


Figure 6. Example ROC curve.

Note. Numbers above each plotted point provide the percentile cutoff for that point.

nomination cutoff changes the integrated sensitivity and false positive rate in two-stage identification systems. The principle is simple. Begin by fixing the values of the confirmatory test reliability, cutoff score, the nomination validity, and nomination reliability coefficients. Then compute the sensitivity and incorrect identification rates of the integrated identification system using the methods previously outlined in this article. Performing this computation across a range of possible nomination cutoff scores yields a series of points which, when plotted, yield the modified ROC curve. Results of this analysis for an identification system in which nomination reliability is set to .90, nomination validity is set to .70, the test reliability is set to .95, and the test cutoff is set to the 90th percentile, are presented in Figure 7.

The reference lines at 84% sensitivity and 16% incorrect identification rate display the performance that would occur if the nomination stage was abolished and instead the confirmation assessment was given to all students. The numbers next to each point indicate the nomination cutoff score (as a percentile) corresponding with those values for the computed sensitivity and false positive rate.

The pattern is interesting. In traditional ROC curve analyses of single instruments, there is a clear tradeoff between sensitivity and the false positive rate (or the related incorrect identification rate) as cutoffs vary. In the two-stage ROC curve, the confirmatory test provides strong protection against incorrect identifications because false positives with respect to the nomination are unlikely to successfully qualify on the test. In contrast, the confirmatory assessment can provide no protection whatsoever against false negatives

because those students that get a false negative with respect to the nomination are denied access to the confirmatory test. As a result, the points of the two-stage ROC curve are confined to the “box” imposed by the maximum sensitivity and maximum incorrect identification rates that would occur in the absence of the screening test, and the usual sensitivity-specificity reciprocity is altered such that sensitivity varies with the cutoff much more strongly than specificity.

The optimal nomination cutoff, where “optimality” is defined as the best compromise between sensitivity and the incorrect identification rate, is therefore found at the point of maximum sensitivity, and this value will be found at the lowest possible cutoff on the nomination instrument. A ROC curve analysis for the two-stage identification system would therefore suggest that the screening process be removed, although as can be seen in Figure 7, sufficiently low nomination cutoffs result in nearly imperceptible reductions in sensitivity.

Optimal Nomination Cutoffs are a Values Proposition. Earlier in this article, we described useful screeners as those that substantially reduce the time and cost of assessment while imposing only a *mild* performance penalty. The ROC curve analysis locates an optimal tradeoff between sensitivity and specificity. The adoption of a nomination phase as prelude to formal evaluation is motivated by concerns of cost and efficiency rather than maximizing psychometric performance—as we have stated several times, adding a screening phase can only harm the efficacy of the resulting data. Optimizing the nomination process with respect to the two primary considerations of sensitivity vs. cost becomes a values proposition with no simple answer. The real question should be, “how much sensitivity am I willing to sacrifice to achieve a given level of cost savings?” On this point, we can provide guidance. Table 5 provides the absolute and relative sensitivity and false positive rates by nomination cutoff for a two-stage identification system with nomination validity equal to .7. Different levels of nomination validity would result in different values in this table. We chose a validity coefficient of .7 because, on consideration of the literature regarding teachers’ beliefs about giftedness (i.e., Berman, Schultz, & Weber, 2012; Miller, 2009; Siegle et al., 2010), this is about as high of a validity coefficient as one could expect to encounter. One could expect the absolute and relative sensitivities to degrade even more rapidly as the nomination cutoff is raised.

Table 5 indicates that if the nomination validity is equal to .7, setting the nomination cutoff to the 55th percentile (barely above average) would reduce the relative integrated system sensitivity to 95% of its optimal value. Raising the nomination cutoff to the 65th percentile would reduce the relative integrated system sensitivity to 90%.

Table 6 and Figure 8 provide the nomination cutoffs necessary to achieve a tolerable decrease in relative sensitivity for nomination instruments with varying levels of validity. As before, *relative sensitivity* is defined with respect to the

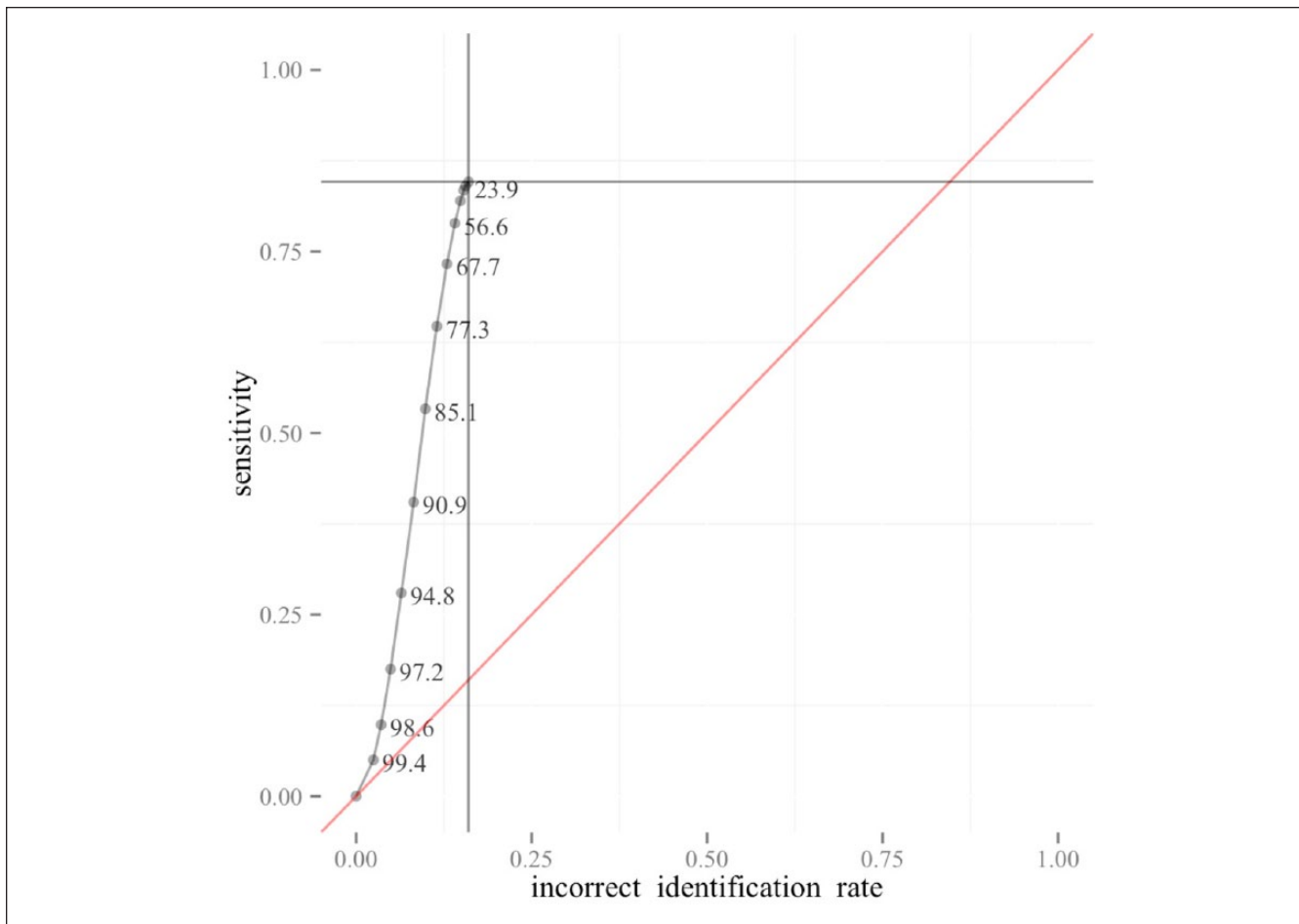


Figure 7. Modified ROC curve for nomination cutoffs in a two-stage identification system.

Note. Nomination reliability = .90; nomination validity = .70; test reliability = .95; test cutoff = 90th percentile. Numbers beside each plotted point give the nomination cutoff at that point. The horizontal and vertical reference lines display the sensitivity and incorrect identification rate, respectively, of a corresponding single assessment system without a nomination stage.

sensitivity that would have been achieved with no nomination stage at all, which is 84.3% when the test cutoff is at the 90th percentile and the test reliability is .95. This analysis was performed by computing the integrated system sensitivity at 10,000 equally spaced nomination cutoffs from $z = -0.5$ (30.8th percentile) to $z = 3.0$ (99.9th percentile) for 19 choices of nomination validity—from .00 to .95 in .05 increments. We used values of .90 for the nomination reliability and .95 for the test reliability. We identified the nomination cutoffs necessary to achieve sensitivity reductions of 2%, 5%, 10%, 15%, 20%, 30%, 40%, or 50% by specifying a loss function defined as $(0.843 - (0.843 * \text{target sensitivity reduction}) - \text{sensitivity})^2$ and then used R to find the nomination cutoff corresponding with the minimum value of the loss function, resulting in the selection of the nomination cutoff that most closely results in the desired sensitivity reduction.

The bottommost plotted curve in Figure 8 indicates the nomination cutoffs that must be used to achieve no more

than a 2% drop in relative sensitivity, which corresponds with an absolute sensitivity of 82.6% in this case. For example, if the nomination instrument has a validity of .60, the highest tolerable nomination cutoff is about 30th percentile. This implies that 70% of the students can be expected to pass the nomination stage and should receive the Stage-II confirmatory assessment. If policy makers are willing to accept a 20% loss of relative sensitivity, a cutoff at the 67th percentile could be used. The school could now expect to bear the expense of testing 33% of the students, meaning they spend less than half on testing compared with the previous example, but at the cost of many more false negative identification decisions. The 20% loss of relative sensitivity implies an absolute sensitivity of 67.4%, so roughly a third of the school’s gifted students would not be identified even though they should have been. Using Table 6 and Figure 8 most schools and districts should be able to find a rough approximation for their practice in order to inform their own identification policy.

Table 5. Absolute and Relative Sensitivity and Incorrect Identification Rates for a Two-Stage Identification System With Nomination Validity of .70 by Nomination Cutoff.

Nomination cutoff	Sensitivity		Incorrect identification rate	
	Absolute	Relative	Absolute	Relative
.05	.843	1.000	.157	1.000
.10	.843	1.000	.157	1.000
.15	.842	.999	.157	.998
.20	.841	.998	.156	.995
.25	.840	.996	.155	.990
.30	.837	.993	.154	.983
.35	.833	.988	.153	.973
.40	.828	.982	.151	.961
.45	.820	.972	.148	.945
.50	.809	.960	.145	.926
.55	.795	.942	.142	.904
.60	.775	.920	.138	.876
.65	.750	.889	.132	.843
.70	.716	.850	.126	.804
.75	.672	.797	.119	.757
.80	.614	.728	.110	.700
.85	.535	.635	.099	.630
.90	.428	.508	.085	.541
.95	.272	.323	.063	.404

Note. Nomination reliability = .90, test reliability = .95, test cutoff = 90th percentile. Other values of nomination validity would result in differing sensitivity and false positive rates.

Table 6. Optimal Nomination Cutoffs by Nomination Validity and Maximum Tolerable Decrease in Relative Sensitivity.

NV	Maximum tolerable decrease in relative sensitivity							
	.02	.05	.10	.15	.20	.30	.40	.50
.30	—	—	25.1	33.1	40.2	52.2	62.4	71.2
.35	—	18.3	28.8	37.2	44.3	56.2	66.0	74.3
.40	—	21.8	32.9	41.4	48.6	60.2	69.5	77.2
.45	—	25.7	37.3	45.9	53.0	64.1	72.8	79.9
.50	18.9	30.1	42.1	50.6	57.4	68.0	76.0	82.4
.55	23.3	35.1	47.1	55.4	61.9	71.7	79.0	84.7
.60	28.4	40.6	52.4	60.3	66.3	75.2	81.7	86.7
.65	34.3	46.6	57.8	65.2	70.6	78.6	84.3	88.6
.70	41.1	53.0	63.4	70.0	74.8	81.7	86.6	90.3
.75	48.7	59.7	68.9	74.6	78.7	84.5	88.6	91.7
.80	57.0	66.6	74.4	79.1	82.4	87.2	90.5	93.0
.85	65.8	73.5	79.6	83.3	85.8	89.5	92.1	94.1
.90	75.0	80.4	84.6	87.1	89.0	91.6	93.5	94.9
.95	84.6	87.2	89.4	90.7	91.7	93.3	94.5	95.6

Note. NV = nomination validity. Values in the table are the maximum allowable nomination cutoff percentile required to achieve the desired decrease in relative system sensitivity (relative to an identification system with no screener) and are also plotted in Figure 8; Assumes nomination reliability = .90, test reliability = .95, test cutoff = 90th percentile; Missing values indicate that no solution existed within the search space, which began at a minimum nomination cutoff of $z = -1.0$ (15.9th percentile).

Discussion

The analysis presented in this article shows that, unless nomination stages are carefully constructed with high validity (which requires high reliability) and low cutoffs (low especially when compared to traditional gifted education cutoffs of 90th percentile or higher), they are almost always extremely detrimental to identification system performance. However, screening/nomination stages can be quite helpful when they adhere to psychometric principles. Given that the use of nomination stages is extremely common in gifted education, and that these issues have not been widely understood by practitioners in the field, we find it likely that most identification systems that have incorporated a nomination stage have not performed well and, in fact, have missed larger percentages of gifted students than they have actually successfully identified. Adapting a slight variation on Freedman's (1991) comments on regression models, we offer the following four possibilities for identification systems involving nomination stages:

1. Nomination stages usually work, although they are (like anything else) imperfect and may sometimes go wrong.
2. Nomination stages sometimes work in the hands of skillful practitioners, but aren't suitable for routine use.
3. Nomination stages might work, but they haven't yet.
4. Nomination stages can't work (Freedman, 1991, p. 292).

Like Freedman, we believe that the truth is bracketed by options two and three, and we lean toward option three. In common identification practice, cutoff criteria are set at the "gifted" range for both the nomination and the confirmation assessment phases. If nominations are to be implemented at all, it is of utmost importance that only high-validity instruments and procedures should be used. In most cases, the nomination cutoffs will require substantial reduction in order to achieve reasonable performance given the low correlations that often exist between screening and confirmation assessments. We believe this is a key point that is not widely appreciated.

Currently implemented nomination practices likely result in very few false positive identification decisions (meaning few students are identified "on accident") at the cost of a tremendous numbers of false negatives (many students are missed who should have been identified). In some cases this may be an appropriate balance—such as when programs could be dangerous or harmful for those students who are not truly ready. However, we believe that in most cases, this practice is needlessly exclusionary and should be modified to more correctly balance false negatives with incorrect identifications. Addressing this has a relatively simple set of solutions.

1. Increase the validity of nominations.
 - a. When informal nominations are solicited from teachers, the teachers should be trained to recognize the qualities that will be assessed during

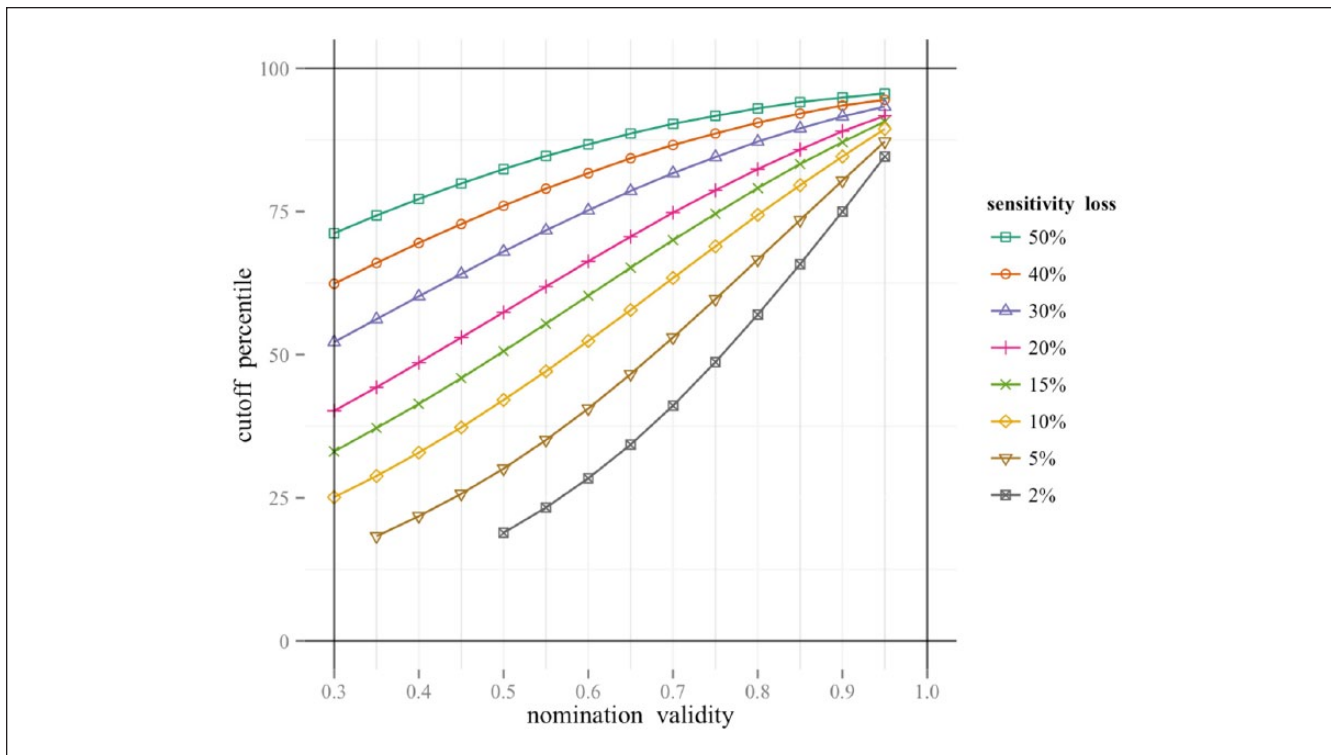


Figure 8. Optimal nomination cutoffs by nomination validity and desired system sensitivity. Note. Assumes nomination reliability = .90, test reliability = .95, and test cutoff = 90th percentile. Each plotted line displays the maximum allowable nomination cutoff for each level of nomination validity that preserves the integrated system sensitivity (relative to a system with no screener) at the desired level. Plotted values appear in Table 6.

- the confirmatory assessment(s) and should be asked to nominate students accordingly. These two phases must align as closely as possible.
- b. Formal nomination instruments should be selected based on the degree to which the scores they produce are correlated with scores in the confirmatory assessment. Strong phase-I to confirmatory alignment is key.
 2. Select appropriate nomination cutoffs based on the desired compromise between sensitivity and assessment cost (see Table 6 and Figure 8). For informal teacher nominations, teachers may frequently need to be instructed to nominate all children in the class that are above average rather than perceived to be potentially in the “gifted range” in order to achieve the required low cutoff. In general, we believe nomination/screening cutoffs need to be lowered.
 3. Consider abolishing the two-stage system completely. Testing all students with simple and (relatively) lower-quality assessments will almost always result in better system sensitivity than two-stage systems in which a poor nomination process is followed by elaborate, expensive, and high-quality testing.

Solution 1a has been suggested many times by researchers examining teachers’ recommendations and beliefs about

giftedness (i.e., Harradine, Coleman, & Winn, 2014; Michael-Chadwell, 2011; Neumeister et al., 2007). It is a generally accepted fact that classroom teachers would benefit from more training in gifted education. Solution 1b may encounter some debate by those individuals who believe that teachers are able to provide information about children that is distinct from that measured by standardized tests. Foreman and Gubbins (2015) found evidence that teachers can make a distinct contribution to identification when teachers are asked to cast a wide net, *such as the top 25% of the class*. This research supports Solution 2 in which teachers nominate a larger percentage of their students for further assessment. Solution 3 is likely to be rejected by many school districts as it seems to go against the desire to identify a wide range of gifted students, but the point is that a two-stage system *with a poor nomination phase* is restricting, not diversifying the population of students identified as gifted.

In closing, the use of nomination stages in gifted program identification can result in a stunningly high false negative rate unless the process is designed very carefully and with a focus on the psychometrics described throughout this article. The solutions (as described above) are simple: increase the validity of the nomination stage and/or lower the nomination cutoff. That said, increasing the overall system validity may be challenging given the constraints of modern psychoeducational assessment. Lowering the nomination cutoff logically

implies that more students will require testing using the confirmatory measures and therefore increased assessment costs for schools. However, perhaps this is not as much of a problem as might be initially assumed since in many cases the per-form cost of popular teacher nomination instruments is roughly equivalent to the per-form cost of some popular group ability tests. These parallels are complicated by the fact that many such tests involve software, scoring manuals, and so on, and as such the price comparison is not a simple one. However, in such cases where the confirmation assessment is only marginally more expensive than a quality teacher nomination or screening phase, abolishing the nomination stage completely is likely to be the best policy. What can certainly be said from the perspective of this work is that if the goal is to identify a particular group (say the top 2%, 5%, or 10% in a particular domain) because educators have reason to believe those students need specialized services, then steps need to be taken to determine what will happen when this goes wrong (due to measurement error). In other words, what will be done to respond to the inevitable false negatives that will result? What additional measures, procedures, programs, or identification system modifications need to be implemented or undertaken in order to make sure what all students who have a need for a particular service receive it? As we have outlined

above, commonly-implemented practices in gifted education create alarmingly large numbers of false negatives, which not only degrades the internal consistency and integrity of gifted education services, but results in large numbers of students failing to receive needed educational intervention.

The authors of this article are on record in opposition to a model of gifted education which begins with an attempt to “identify the gifted,” because we believe that the usual conception of giftedness as a trait of individuals, with stable manifestation across academic domains, lifespan, and educational arrangements (cf., Peters et al., 2014), is not educationally useful though it is scientifically interesting. Indeed, the prominence that the identification process receives in gifted education strikes us as misguided and counterproductive especially in light of the results presented in this article. Regardless of our position, however, we recognize the dominance of the current paradigm. Given that the practice of gifted education, as widely implemented, begins with an identification process in which fine distinctions of true ability, achievement, or creativity are held to be meaningful, in this article we use quantitative techniques to ask, “how well do these identification systems work?” Our answer, briefly summarized, is “probably not very well.” We leave critique of the system itself and its goals to other venues.

Appendix

R Script for Calculations

```
### You must install the mnormt library for this script to work ###
library(mnormt)

### These values should be changed as desired ###
tau <- 1.28 # the confirmatory test cutoff is z=1.28 (90th percentile)
nu <- 1.28 # nomination cutoff is also at z=1.28 (90th percentile)
relyc <- .95 # confirmatory test reliability is 0.95
relyn <- .90 # nomination reliability is 0.90
r <- 0.70 # nomination validity coefficient

### Do not change code beyond this point ###
mean <- c(0,0,0,0) # mean vector is all zeros

### Creates the covariance matrix (eqn 4) ###
# (The order of variables is nomination true score, nomination observed score,
# confirmatory test true score, and confirmatory test observed score)
cov <- matrix(c(1, sqrt(relyn), r/sqrt(relyn*relyc), r/sqrt(relyn),
               sqrt(relyn), 1, r/sqrt(relyc), r,
               r/sqrt(relyn*relyc), r/sqrt(relyc), 1, sqrt(relyc),
               r/sqrt(relyn), r, sqrt(relyc), 1), nrow=4, ncol=4)

# Calculate sensitivity (eqn 5). Note that infinities are replaced by
# +/- 5 to stabilize the numerical integration
```

```

sensitivity <- sadmvn(lower=c(-5, nu, tau, tau), upper=c(5,5,5,5), mean=mean,
varcov=cov) / sadmvn(lower=c(-5, -5, tau, -5), upper=c(5,5,5,5), mean=mean, varcov=cov)
# Calculate incorrect identification rate (eqn 6)
incorrectid <- sadmvn(lower=c(-5, nu, -5, tau), upper=c(5, 5, tau, 5), mean=mean,
varcov=cov) / sadmvn(lower=c(-5, nu, -5, tau), upper=c(5, 5, 5, 5), mean=mean,
varcov=cov)
# Calculate positive predictive value (eqn 7)
PPV <- 1 - incorrectid[[1]]
# Calculate proportion identified (eqn 8)
proportionid <- sadmvn(lower=c(-5, nu, -5, tau), upper=c(5, 5, 5, 5), mean=mean,
varcov=cov)
# Print results to screen
sensitivity[[1]]
incorrectid[[1]]
PPV
proportionid[[1]]

```

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.

Notes

1. Although we are using many references to medical practices and research in this article, we hope it is obvious that we do not see “giftedness” or the need for gifted or advanced academic services as a parallel for any kind of medical disease. Instead, we simply seek to learn from practices used in other fields in order to help assure more children receive appropriate educational services.
2. By “truly” gifted student we mean a student who actually does meet the standards specified by the school or school district, regardless of whether or not that student has been identified as such. We contrast this with the colloquial and unfortunate usage of the term “truly gifted” to indicate students of exceptionally high ability.

References

- Aron, A., Coups, E. J., & Aron, E. N. (2013). *Statistics for psychology* (6th ed.). New York, NY: Pearson.
- Berman, K. M., Schultz, R. A., & Weber, C. L. (2012). A lack of awareness and emphasis in preservice teacher training: Preconceived beliefs about the gifted and talented. *Gifted Child Today*, 35, 18-26. doi:10.1177/1076217511428307
- Callahan, C. M., Moon, T. R., & Oh, S. (2013). *Status of elementary gifted programs: 2013*. Charlottesville: University of Virginia. Retrieved from <http://www.nagc.org/sites/default/files/key%20reports/ELEM%20school%20GT%20Survey%20Report.pdf>
- Carman, C. A. (2011). Stereotypes of giftedness in current and future educators. *Journal for the Education of the Gifted*, 34, 790-812. doi:10.1177/0162353211417340
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart, & Winston.
- Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). “Mental state”: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12, 189-198. doi:10.1016/0022-3956(75)90026-6
- Foreman, J. L., & Gubbins, E. J. (2015). Teachers see what ability scores cannot: Predicting student performance with challenging mathematics. *Journal of Advanced Academics*, 26, 5-23. doi:10.1177/1932202X14552279
- Freedman, D. A. (1991). Statistical models and shoe leather. *Sociological Methodology*, 21, 291-313. doi:10.2307/270939
- Harradine, C. C., Coleman, M. R. B., & Winn, D. C. (2014). Recognizing academic potential in students of color: Findings of U-STARS~PLUS. *Gifted Child Quarterly*, 58, 24-34. doi:10.1177/0016986213506040
- Hoge, R. D., & Cudmore, L. (1986). The use of teacher-judgment measures in the identification of gifted pupils. *Teaching and Teacher Education*, 2, 181-196. doi:10.1016/0742-051X(86)90016-8
- Hunsaker, S. L., Finley, V. S., & Frank, E. L. (1997). An analysis of teacher nominations and student performance in gifted programs. *Gifted Child Quarterly*, 41(2), 19-24. doi:10.1177/001698629704100203
- Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, 99, 22-44. doi:10.1037/0033-295X.99.1.22
- Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Mahwah, NJ: Lawrence Erlbaum.
- Madelaine, A., & Wheldall, K. (2005). Identifying low-progress readers: Comparing teacher judgment with a curriculum-based measurement procedure. *International Journal of Disability, Development, and Education*, 52, 33-42. doi:10.1080/10349120500071886

- McBee, M. (2006). A descriptive analysis of referral sources for gifted identification screening by race and socioeconomic status. *Journal of Advanced Academics, 17*, 103-111. doi:10.4219/jsgc-2006-686
- 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*, 69-89. doi:10.1177/0016986213513794
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238. doi:10.1037/0033-295X.85.3.207
- Michael-Chadwell, S. (2011). Examining the underrepresentation of underserved students in gifted programs from a transformational leadership vantage point. *Journal for the Education of the Gifted, 34*, 99-130. doi:10.1177/016235321003400105
- Miller, E. M. (2009). The effect of training in gifted education on elementary classroom teachers' theory-based reasoning about the concept of giftedness. *Journal for the Education of the Gifted, 33*, 65-105. Retrieved from <http://eric.ed.gov/?id=EJ856177>
- Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*, 289-316. doi:10.1037/0033-295X.92.3.289
- National Association for Gifted Children. (2013). *State of the states in gifted education: 2012-2013*. Washington, DC: Author.
- National Cancer Institute. (2014). *Mammograms*. Bethesda, MD: Author. Retrieved from <http://www.cancer.gov/cancertopics/factsheet/detection/mammograms>
- Neumeister, K. L. S., Adams, C. M., Pierce, R. L., Cassady, J. C., & Dixon, F. A. (2007). Fourth-grade teachers' perceptions of giftedness: Implications for identifying and serving diverse gifted students. *Journal for the Education of the Gifted, 30*, 479-499. Retrieved from <http://eric.ed.gov/?id=EJ769920>
- Nilsson, H., Juslin, P., & Olsson, H. (2008). Exemplars in the mist: The cognitive substrate of the representativeness heuristic. *Scandinavian Journal of Psychology, 49*, 201-212. doi:10.1111/j.1467-9450.2008.00646.x
- Nissen-Meyer, S. (1964). Evaluation of screening tests in medical diagnosis. *Biometrics, 20*, 730-755. doi:10.2307/2528126
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39-57. doi:10.1037/0096-3445.115.1.39
- Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104*, 266-300. doi:10.1037/0033-295X.104.2.266
- Peters, S. J., & Gentry, M. (2010). Multigroup construct validity evidence of the HOPE Scale: Instrumentation to identify low-income elementary students for gifted programs. *Gifted Child Quarterly, 54*, 298-313. doi:10.1177/0016986210378332
- Peters, S. J., & Gentry, M. (2013). Additional validity evidence and across-group equivalency of the HOPE Teacher Rating Scale. *Gifted Child Quarterly, 57*, 85-100. doi:10.1177/0016986212469253
- Peters, S. J., Matthews, M. S., McBee, M. T., & McCoach, D. B. (2014). *Beyond gifted education: Designing and implementing advanced academic programs*. Waco, TX: Prufrock Press.
- Peterson, J. S. (1999). Gifted—Through whose cultural lens? An application of the postpositivistic mode of inquiry. *Journal for the Education of the Gifted, 22*, 354-383. doi:10.1177/016235329902200403
- Pfeiffer, S. I., & Jarosewich, T. (2003). *Gifted Rating Scales* (3rd ed.). San Antonio, TX: Pearson.
- Plata, M., Masten, W. G., & Trusty, J. (1999). Teachers' perception and nomination of fifth-grade Hispanic and Anglo students. *Journal of Research & Development in Education, 32*, 113-123.
- Qaseem, A., Alguire, P., Dallas, P., Feinberg, L. E., Fitzgerald, F. T., Horwitch, C., . . . Weinberger, S. (2012). Appropriate use of screening and diagnostic tests to foster high-value cost-conscious care. *Annals of Internal Medicine, 156*, 147-149. doi:10.7326/0003-4819-156-2-201201170-00011
- Roid, G. H. (2003). *Stanford-Binet Intelligence Scales: Examiner's manual* (5th ed.). Itasca, IL: Riverside.
- Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 27-48). Hillsdale, NJ: Lawrence Erlbaum.
- Ryser, G. R., & McConnell, K. (2004). *Scales for Identifying Gifted Students*. Waco, TX: Prufrock Press.
- Siegle, D., Moore, M., Mann, R. L., & Wilson, H. E. (2010). Factors that influence in-service and preservice teachers' nominations of students for gifted and talented programs. *Journal for the Education of the Gifted, 33*, 337-360.
- Siegle, D., & Powell, T. (2004). Exploring teacher biases when nominating students for gifted programs. *Gifted Child Quarterly, 48*, 21-29. doi:10.1177/001698620404800103
- VanDerHeyden, A. M., Witt, J. C., & Naquin, G. (2003). Development and validation of a process for screening referrals to special education. *School Psychology Review, 32*, 204-227.
- Zumeta, R. O., Zirkel, P. A., & Danielson, L. (2014). Identifying specific learning disabilities: Legislation, regulation, and court decisions. *Topics in Language Disorders, 34*, 8-24. doi:10.1097/TLD.000000000000006

Author Biographies

Matthew T. McBee is an assistant professor of quantitative psychology at East Tennessee State University, where he teaches undergraduate and graduate courses in multilevel modeling, regression, longitudinal data analysis, and psychometrics. He is assistant director of Experimental Training in the Experimental Psychology PhD program. Prior to his arrival at East Tennessee State University, he served as a statistician at the UNC Chapel Hill Frank Porter Graham Child Development. He is coeditor of the *Journal of Advanced Academics*. His publications have appeared in a wide range of education journals, including *Gifted Child Quarterly*, *Reading and Writing*, *Journal of Autism and Developmental Disorders*, *Exceptionality*, *Research in Autism Spectrum Disorders*, and *Annals of Dyslexia*. He is coauthor of *Beyond Gifted Education* (Prufrock) along with Scott Peters, Michael Matthews, and Betsy McCoach.

Scott J. Peters is an associate professor of educational foundations at the University of Wisconsin–Whitewater where he teaches courses on measurement and assessment, research methodology, and gifted education. His research work focuses on educational assessment, identification, student underrepresentation, and educational policy. He has published in *Teaching for High Potential*, *Gifted Child Quarterly*, *Journal of Advanced*

Academics, Gifted and Talented International, Gifted Children, Journal of Career and Technical Education Research, Ed Leadership, Ed Week, and Pedagogies. He is the past recipient of the Fedlhusen Doctoral Fellowship, the NAGC Research and Evaluation Network Dissertation Award, the NAGC Doctoral Student of the Year Award, the NAGC Early Scholar Award, and the UW-Whitewater College of Education Innovation and Outstanding Research Awards. Along with Matthew McBee, Michael Matthews, and D. Betsy McCoach, he is the first author

of *Beyond Gifted Education: Designing and Implementing Advanced Academic Programs* (Prufrock).

Erin M. Miller received her PhD in educational psychology from the University of Virginia in 2006. She is an assistant professor of psychology at Bridgewater College. Her research interests involve methodologies of measuring implicit conceptions of intelligence and talent and the implications of these conceptions for motivation and achievement.