5-2016

Can Curriculum-based Measures and Teacher Ranking Predict End-of-year Achievement for Students Who are Gifted or High-achieving?

Bruce Alan Ewing
University of Tennessee - Knoxville, bewing@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Bruce Alan Ewing entitled "Can Curriculum-based Measures and Teacher Ranking Predict End-of-year Achievement for Students Who are Gifted or High-achieving?." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Education.

Sherry M. Bell, Major Professor

We have read this dissertation and recommend its acceptance:

David F. Cihak, Tara C. Moore, Jennifer A. Morrow

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

Can Curriculum-based Measures and Teacher Ranking Predict End-of-year Achievement for Students Who are Gifted or High-achieving?

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Bruce Alan Ewing

May 2016

Acknowledgements

Heartfelt thanks to the committee for their advice and support; Dr. David Cihak, Dr. Tara Moore, and Dr. Jennifer Morrow; most especially to Dr. Sherry Mee Bell for her expert guidance and persistence tempered with patience.

Abstract

The lack of cohesion and oversight in federal and state laws that outline identifying and serving gifted/high ability students have been cited by researchers and practitioners as a hindrance in the development of programming designed to serve these populations (National Association for Gifted Children, 2014). Controversy over definitions of giftedness and the role of schools in identifying and serving gifted students indicate that policy and practice in gifted education are highly inconsistent. In partial response, researchers in gifted education have begun to call for the extension of the response-to-intervention (RTI) model to identify and serve gifted students, leading to questions centered on the validity of curriculum-based measures (CBM) used for gifted screening (Burns, Jacob, & Wagner, 2008). This study expands the literature of the field by examining the validity of CBM for gifted screening, the accuracy of teacher perception, and the adequacy of measures taken early in the school year for gifted screening when necessitated by the absence of formal measures traditionally used. Two early measures in reading and math, a qualitative, domain-specific teacher rank and a CBM universal screener, were administered to 372 third graders in a rural school district and results were compared to a quantitative, norm-referenced measure taken at the year's end. The relation between early and late measures is examined to assess the utility of early measures for making educational placement decisions for gifted students. The CBM examined here demonstrated appropriate psychometric properties (sufficient item gradient at the upper end and scores greater than 2 standard deviations above the mean) for effective use in gifted screening. Teacher ranking proved to be a strong predictor of future performance on standardized testing, and when used in combination with the CBM as early measures, yielded an 80% accuracy rate in group assignment when using the later measure as a standard of determination, though reading measures performed more strongly than math measures. Results generate increased confidence in the efficacy of early indicators of student performance for making quotidian planning and placement classroom decisions regarding gifted and high-ability students.

Table of Contents

List of Tables

List of Figures

Chapter 1
Rationale, Methodology, Assumptions, Limitations

Rationale

One theme that emerges quite early in a review of literature in the field of gifted education is that though much is known about what might be considered as best practice, the implementation of such is immediately conditioned by caution against its universal application. As now required by No Child Left Behind Act (NCLB, 2001) the legislation that currently dominates education policy in the United States, state education agencies are compelled to utilize service delivery models, teaching practices, and program designs that attain the standard of best practices, defined as those based upon "… research that involves the application of rigorous, systematic, and objective procedures to obtain reliable and valid knowledge relevant to education activities and programs " (NCLB, Title IX General Provisions, Part A Sec. 9101). The mandate continues that suitable practices derive from research that "employs systematic, empirical methods that draw on observation or experiment" (NCLB, 2001).

Borrowing methods long used by other social sciences, critical reforms in education now require increased attention to scientifically-based research that produces reliable and valid outcomes, uses experimental or quasi-experimental designs, involves rigorous data analyses, allows for replication, and that has been evaluated in a peer-review process or approved by a panel of independent experts (NCLB, 2001). This legislation applies to all programming and services offered in schools, including special education which frequently subsumes gifted education, though noting that gifted

education, per se, is not addressed by NCLB. However, the methods typically utilized to assess the success of educational interventions for discovery of best practice may, in gifted education, be difficult to execute, appear to lack rigor, and be complicated by several important considerations. Thus, when compared to investigations of other educational practices, identification of best practice for gifted education may be significantly more recondite.

Other issues exacerbate these concerns. Comprehensively, the literature on gifted education suggests that decisions about policy and practice are best made at a local level in response to individual needs and in consideration of the resources available at the local education agency (LEA). Yet, coincidentally, a discernable trend in the literature decries the need for stronger federal and state levels of policy decision-making. Virtually every research article on gifted education practices includes commentary on the lack of a federal definition and the ancillary vagaries of state directives. Information currently found on the website for the National Association of Gifted Children (NAGC) states that all program and service decisions for gifted learners are made at state and local levels noting the wide variability between state policies, and in many cases, even wider unevenness between districts in the same state (see 2012-2013 State of the States in Gifted Education). This leads to policies that are unevenly applied; confusing; lack funding, monitoring, and oversight; or that are in some cases contradictory. Though federally commissioned reports empirically document the need for gifted services (U.S. Department of Education, 1993), federal and state governments continue to fail to

provide sufficient support for gifted education through legislative policy and program funding (Brown, Avery, Van Tassel-Baska, Worley, & Stambaugh, 2006).

Additional to federal, state, and district level concerns appertaining to gifted education are those arising from classroom practices. The Individuals with Disabilities Education Improvement Act (2004) increases teacher accountability through increased teacher accountability and annual monitoring of student progress, and supports the implementation of the response to intervention model (RTI) within the general education classroom. RTI focuses teacher attention more specifically toward performance levels present in the classroom to design grouping structures and lesson differentiation that remediate struggling learners in both short- and long-term tier placements through increased intensity, duration, and frequency of instruction. The use of the RTI process and the documentation of student attainment collected thereby have become a new path for the screening, identification, and remediation of many special education disability categories, and can even serve as a platform for disabilities service delivery. The use of RTI, however, as part of a referral process to screen, identify, and/or serve gifted and high-ability (G/HA) students has only recently become a topic of research.

Kavale and Spaulding (2008) note that the closer alignment of RTI with NCLB, as opposed to IDEA, brings a stronger emphasis on scientifically valid practices and increased rigor when identifying best practice (though they also note several unintended negative consequences). Though ostensibly designed to develop specialized classroom strategies for low-performing students, the need to develop an analogous RTI

plan for neglected gifted children is obvious (Kavale & Spaulding, 2008). However, Volker, Lopata, and Cook-Cottone (2006) assert that the current conceptualization of RTI makes it more suited for identifying children who have learning difficulties, adding that "on its own, RTI is not particularly well suited for identifying gifted children at Tiers 1 and 2" (p 863). The Association for the Gifted, a division of Council for Exceptional Children (CEC-TAG), has recognized the potential of adapting the RTI framework for gifted learners, and the CEC recommends that the "RTI model be expanded in its implementation to include the needs of gifted children" (Council for Exceptional Children, the Association for Gifted, 2009, p. 1). The RTI framework can support the advanced learning needs of gifted students by facilitating such accommodations as a faster pace, and more complex content presentation in greater depth and/or breadth with respect to the curriculum (Council for Exceptional Children, the Association for Gifted, 2009, p. 1).

More important, however, are the consequences that accrue to gifted children while waiting for issues in gifted education to be resolved in any practicable manner, a wait that can possibly delay appropriate interventions or differentiation in the meantime. When screening and identifying students for many of the special education categories, NCLB and state policy may require documentation completed by general education teachers from all three tiers of RTI interventions (each between six and nine weeks long) before the referral process to special education even begins (noting explicitly that for identification of some disabilities this step is not required or may be abbreviated).

Additionally, information from end-of-year standardized, criterion- or norm- referenced testing typically used to make many educational decisions may be absent or unavailable sufficiently early in the academic year to be of use. In certain grades, due to the schedule of grades tested and the nature of the tests administered, documentation of student performance may lack any formalized or standardized measures. For specific cases, such as for transfer students as an example, some states allow or mandate additional testing or screening upon take-in providing more information for decision making. However, there are many potential exigencies in which a reliance on measures taken early in the school year becomes necessary when designing and implementing lessons and interventions, or for screening and identification as part of the referral process. Though reliance upon early-in-year measures may be unavoidable for many teacher decisions, moderate to high rates of predictability of such measures for late-in-year measures may provide increased credibility to gifted screening and identification practices and improve the likelihood of increased levels of intervention for gifted and high ability learners to begin earlier in the school year.

Purpose of the Study

It is logical to assume that teacher perception of student abilities influences many quotidian decisions such as grouping, classroom seating, and assignment differentiation in many informal ways that have not yet been researched. In such cases, teachers may have their own anecdotal evidence, or that from previous years, to make initial decisions about general student ability that affect many of these classroom

practices. This entails more sensitive awareness of the range of student ability present within the classroom and escalates the need for instruments and processes to document student attainment, as part of either the referral process for formal identification for special education, or as justification for the implementation of other, less formal strategies (such as homogeneous grouping or enrichment) within an RTI model consistent with meeting the needs of gifted learners.

In this study the efficacy of early-in-year measures is examined and their relation to end-of-year measures to enhance confidence in their use for making educational decisions for gifted and high-ability (G/HA) students. For optimum usage, these instruments should be easy to administer in a variety of group settings, able to screen for multiple levels of ability, able to be administered frequently without negative testing effects, and able to track student progress in an RTI setting. The instructional decisions based on progress monitoring using curriculum-based measures (CBM) taken quickly and frequently to assess student acquisition of single-subject content have been shown to be effective for remediating struggling students in both short- and long-term applications. It should not be summarily assumed, however, that all instrumentation and protocols for screening, identifying, and serving students for other special services are equally valid when extended to above-grade-level applications within the context of giftedness. To meet the guidelines presented by NCLB these strategies require explicit exploration.

Studies in gifted education present unique challenges to researchers and may necessitate embracing methodologies and methods not found in traditional empirical

studies. These include: the degradation of random group assignment required by experimental and quasi-experimental research designs; problems associated with small sample sizes by virtue of examining only the top 2-5% of any given student population making it difficult to show statistical significance or strong effect sizes; atypical variance in samples, as giftedness manifests in such individual ways that group membership may not reflect any similarity of student profile other than gifted/high ability (G/HA) identification; restriction of range produced both by the inability of instrumentation to adequately assess the upper limits of student ability (known as *ceiling effects*) and the tendency of metrics (such as achievement scores) used to measure gifted students to cluster at the highest levels of performance resulting in a lack of heterogeneity which reduces variability and "…leads to attenuated reliability coefficients" (Kieffer, Reese, & Vacha-Haase; 2010), and a lack of consistent construct definitions and operationalization of giftedness between studies (see Subotnik & Thompson, 2010).

In Tennessee, the state department of education provides for gifted services within the purview of special education. However, implementation decisions are left to individual districts, including the fundamental decision to identify and serve gifted students. The state provides a separate manual with additional guidelines for gifted identification that defines specific protocols, as these differ from other special education categories. A matrix entitled *Tennessee K-12 Intellectually Gifted Assessment Scoring Grid* (https://www.tn.gov/assets/entities/education/attachments/se_eligibility_gifted_res_pkt.p

df.; 2010) provides the documentation of required elements for identification. The gold-standard for identification on the identification matrix is the state designed, end-of-year, high-stakes test the *Tennessee Comprehensive Assessment Program* (TCAP; latest iteration, 2013) given annually in grades 3 through 12. Many districts provide gifted services for only elementary grades 3 through 5. Identification of gifted third graders then becomes problematic as the most frequently used identification metric, the TCAP, is not administered to second graders. Thus, third grade teachers must rely on other measures or procedures when identifying third graders for gifted services. The question, then, centers on the efficacy of early-in-year CBM and measures of teacher perception in predicting gifted status in reading and math as defined by the parameters set forth by the state.

Research Questions

The literature review generated the following question: What is the efficacy of early-in-year CBM and measures of teacher perception in screening for gifted status in reading and math? Specifically:

1. Do CBM of reading and math (as measured by the Monitoring Instructional Responsiveness: Reading (MIR:R) and Monitoring Instructional Responsiveness: Math (MIR:M)) provide sufficient ceiling to serve as screeners for gifted and high ability students (G/HA) in a general education classroom sample?
2. To what degree or extent are the domain-specific (reading and math) scores for the measuring instruments related to each other for the entire sample?

a) To what degree or extent are the early domain-specific (reading and math) MIR scores related to each other for the entire sample?

b) To what degree or extent are early Teacher Rankings (TR) as measures of teacher perception of student performance in domain-specific (teacher rank reading TR:R, and teacher rank math TR:M) scores related to each other for the entire sample?

c) To what degree or extent are the late-in-year domain-specific (reading and math) Tennessee Comprehensive Assessment Program (TCAP, TCAP:R, TCAP:M) scale scores related to each other for the entire sample?

d) To what degree or extent is the magnitude of the MIR inter-correlation comparable to that of the TCAP inter-correlations for the entire sample?

3. To what degree or extent are the MIR:R and MIR:M related to TR in reading and math as a measure of teacher perception (TR:R, TR:M) for the entire sample?

4. To what degree or extent can early-in-year measures predict end-of-year measures?

a) To what degree or extent can early-in-year CBM (as measured by MIR: R and MIR:M) predict the TCAP scores as an example of end-of-year measure?

b) To what degree or extent can TRs of reading and math as examples of early-in-year measures predict the TCAP scores as an example of end-of-year measure?

c) To what degree or extent can the MIR and TR collectively predict TCAP scores?

5a) Is there a significant difference in the rate MIR, TR, and TCAP identify G/HA students based on dichotomous gifted group assignment? Group assignment is defined as

attainment at or above the 85Th percentile for MIR and TCAP and as the top two ranks for the TR.

5b) Do the MIR, TR, and TCAP identify the same cases of G/HA students based on dichotomous gifted group assignment (assignment is defined as at or above the 85th percentile for MIR and TCAP, and as the top two ranks for the TR)?

Significance

The search for universally applicable *best methods* in gifted education is significantly hampered by important characteristics inherent within the target population and not yet validated applications and extensions of other practices and metrics. Lack of legislation and legal mandates requiring states to identify and serve gifted children are frequently cited as fundamental issues that degrade the development of the field (Brown & Van Tassel-Baska, 2006). Brown and Van Tassel-Baska lament the "paucity of research" of state policies regarding their relative strengths, limitations, and effects on practice in gifted education. Moreover, Brown and Van Tassel-Baska assert that effective state and federal policies can legitimize the perception of the need for gifted services and dispel misconceptions associated with giftedness such as elitism.

Researchers in gifted education seem to be embattled with policy makers who seek restrictions in the operationalization of the term *gifted*, ready instrumentation, and clearly explicated practice. The language of NCLB (2001) suggests that interventions need to be evaluated using quantitative metrics and analyses that produce reliability and validity coefficients, as is much the case for the evaluation of interventions designed for

general education and other special education classrooms. In compliance with federal legislation which seems to favor empirically attained findings, and state and district policies that encourage and perpetuate an abiding interest in quantitatively measurable outcomes, policymakers increasingly insist on quantitative metrics to evaluate gifted children and the interventions used to meet their learning needs within public school systems. However, many researchers perceive that qualitative research is still best suited to address the generally misunderstood (and perhaps more important) affective concerns of gifted and high ability children as a way to better meet their needs through an increased understanding of their psychosocial-emotional dispositions. Adherents of this position argue that gifted children are by definition qualitatively different from their peers, and that these significant differences may only be discerned and catalogued through qualitative methods. The confluence of two such diametrically opposed positions (i.e., those of researchers and policymakers) cannot help but militate against much needed progress in serving gifted children. There remains a clear need for the validation of metrics or processes designed to screen and identify gifted students. This is predicated by a consensus definition of giftedness, increased investment of stakeholders, and clarity and unanimity in federal, state, and district policy.

Assumptions

It is assumed that identifying, serving and promoting gifted students is a worthwhile endeavor. Brown and Garland (2015) contend that society loses human capital when gifted children are not nurtured, an easily defensible position, with James

Gallagher going so far as to claim that failure to act on behalf of gifted students is a threat to national security (Gallagher, 2005). In Book III of *The Republic*, Plato (~380 BCE) comes forward as the first advocate of gifted children (though in a restricted manner that would be untenable in society today), recognizing that some children possess advanced abilities, whom he labeled as "children of gold," gifted children whose talents should be developed in service to the city-state. Two millennia later, Maslow (1964) proposed via elaborations of his *Hierarchy of Needs* that the goal of human development is to answer the question "… of what the human being should grow toward" (p. 7), a position that marks the beginnings of his theories of *self-actualization* eventually articulated as the "farther reaches of human nature" (Maslow, 1971). Though it may be possible to question the purpose of identifying and serving gifted and high ability children, whether it may be for societal or individual benefit, the questions of whether or not giftedness exists and if gifted children should be identified and served has never been at issue. Implied by these assertions is the fact that giftedness exists and benefits accrue to the individual, community, and society at large when the talents and abilities of gifted individuals are developed and exercised to the improvement of society as a whole. It is assumed, then, that constructs such as giftedness can be assessed reliably and validly, and that measurement science is adequate for this purpose.

Methodology

In his seminal essay, *The Structure of Scientific Revolutions*, Thomas Kuhn (1966) explicates the qualities of any given domain of scientific endeavor, as well as the

characteristics of change leading to the development of new paradigms within a field. Established scientific fields are characterized by a consensus of the scientific or professional community concerning the fundamental tenets, knowledge base, governing rules, qualifications for experts, etc. Practitioners or experts in the field contribute to the development of theoretical underpinnings and epistemology, a common and functional vocabulary, and worldview. Kuhn theorizes that a field attains the level of *normal science* when practitioners, researchers, and society at large no longer feel compelled to explain the principles or conditions of the field, but assume that these are fully understood by all who operate within the field.

Philosophical assumptions become important in preparing and completing a study, because they guide the use of research methodologies and methods. It is vital to understand the difference between a researcher's *methodology* and the *method* used for discovery, as the terms are not interchangeable, despite frequent usage that lacks clarity on this point. The term *methodology* refers to ''the general logic and theoretical perspective'' (Bogdan & Biklen, 2007, p. 35) and reflects a set of epistemological and ontological assumptions. Three research methodologies receive general approbation: *quantitative*, *qualitative*, and *mixed-methods* (Cohen, Manion, & Morrison, 2011; Creswell, 2014).

Quantitative researchers answer their research questions through the use of measurement, experiment, and statistical analysis; though observations, interviews, and content analysis are preferred by qualitative researchers. Mixed-methods attempts a

middle ground between the two, combining elements from each (Long, 2014). *Method* refers to the specific strategies and procedures for analyzing and interpreting data utilized during an investigation (Bogdan & Biklen, 2007; Lincoln & Guba, 1985; Merriam, 2002). Mixed-method approaches often demonstrate a systematic progression beginning with case studies and correlational research, then concluding with full interventions and laboratory-based experimental trials (Mullan, Todd, Chatzisarantis, & Hagger, 2014).

Current research within the literature base of gifted education indicates that the field is potentially experiencing a *crisis* within its paradigm, Kuhn's term for divagations from normal science that signal at a paradigmatic shift. Though the word *crisis* might indicate the eristic transition that generally typifies scientific revolution, such does not appear to be the case in the field of gifted education. Many of the established leaders of the field are working cooperatively to refine existing concepts, adapt current research methodologies from other social sciences to the new demands of the field, clarify and expand the working vocabulary associated with the field, and to develop new strategies to answer better current questions in gifted education.

In Kuhn's terms, gifted education does not operate as normal science; or, more accurately perhaps, the field has yet to attain the level of normal science, which cannot help but create a lack of unqualified affiliation with either a wholly quantitative or qualitative paradigm. Yet to reconcile this tension by putting forth that the field is tolerant of *mixed methods*, may be to deny the fact that there has been for some time a state of flux in the field while it validates methods that are currently imported from

special education and other social sciences. Contemporary researchers in gifted education must evince higher tolerance for new perspectives in the theoretical underpinnings of their research, as well as the forbearance necessary to resolve the accompanying epistemological tensions. Still, while the nature of "knowing" may yet be as elusive as ever, it is both convenient and necessary for researchers currently operating amid this theoretical quagmire to assume that some phenomena associated with the field are able to be both defined and measured.

Borland (1990) asserts that *post-positivism* is an appropriate research paradigm for studies of gifted children and the programming serving them in an article thoroughly comparing and contrasting post-positivism to *positivism*, the paradigm most dominant in empirical research. The post-positivistic position does not entirely jettison positivism, rather critiques, partially refutes, and/or elaborates positivism and its associated empiricism, the idea that observation and measurement are at the core of scientific endeavors (Trochim, 2006). Post-positivism incorporates many of the basic assumptions of positivism including an ontology that beliefs are approximations of reality and new observations deepen the understanding that reality; however, knowledge is attained often through the use of what positivists might describe as "experimental" methodologies (Christ, 2014). Following is a brief, general summary from Borland (1990) of post-positivism noting, again, that this is given explicitly within the context of an application to studies in gifted education research. Other contributors' relevancies are noted with citations.

Post-positivists believe that the phenomena of interest should be examined only in their natural settings (alternative to "laboratory" settings), which provide a more holistic context for consideration; this has the additional advantage of eliminating artificial control imposed by the isolation of single variables and facilitates an understanding of the variables as they may interact with each other. The researcher himself or herself becomes the data-gathering instrument, as opposed to paper-and-pencil measures and instruments. Because all measurement might be viewed as fallible, the post-positivist emphasizes the importance of multiple, converging measures and observations (known as *triangulation*), realizing that each measure may possess different types of error, but which collectively aggregate to inform a construct (Trochim, 2006).

The post-positivist also believes that all researchers and subjects are inherently biased by their culture, experiences, world views, mores, etc. (Trochim, 2006). Moreover, the paradigm legitimizes the researcher's use of intuition (also known as *tacit knowledge*) and the interactions occurring between the inquirer and the setting.

Qualitative methods (defined by Miles and Huberman in 1984 as data in the form of words, not numbers) are also identified as an acceptable way to generate data. Data from target populations may be gathered through *purposive sampling* which endeavors to attain maximum variation when sampling, intentionally seeking sites, subjects, or observations that differ to the greatest degree possible one from another. Contrasting sharply with positivism is an approach known as *grounded theory* by which the post-positivist might not begin the inquiry with a theory in place, but may allow one to

develop from the data. This prevents the imposition of a priori decisions that may obscure important, relevant, or unforeseen factors that may be of interest (Glaser & Strauss, 1967). Similarly, even the research design itself may not be preplanned, but may be allowed to emerge from interactions of interest between inquirer and setting.

Postpostivism is enhanced by the use of a case study design adjudging elements such as the research context, the environmental interaction between the site, researchers, and subjects, the values and biases of the researcher, subjects, and those of the community to be associated with a specific investigation. This restricts the interpretation of the findings to the particulars of the case study setting, known as an *idiographic interpretation*, which necessarily limits any *generalization* of findings, though case study findings will add to the general body of knowledge by considering outcomes and determinations aggregated with other relevant research conclusions, known as *transferability*.

Traditional positivist research uses internal validity, external validity, reliability, and objectivity as the criteria for trustworthiness. The post-positivist paradigm is dismissive of these as they spring from a belief in a single reality and linear causality. Lincoln and Guba (1985, as cited by Borland, 1990) proposed analogous criteria for trustworthiness; internal validity is replaced by *credibility*, defined as an adequate description of the multiple constructed realities in a manner that is credible to the constructors of those realities. External validity, as suggested by the generalization of finding to other settings, is analogous to transferability as described above. Reliability,

usually reported as *error variance* by the positivist, is replaced by *dependability*. *Confirmability* replaces objectivity and is defined as the degree to which the data can be validated as “true” within the given context.

This study is a correlational study conducted within the post-positivistic paradigm of the relation between two domain-specific, early-in-year measures, and the relation between those two measures and one late-in-year measure in the context of determining gifted eligibility. In correlational research, the researcher does not try to influence the variables in any way, but attempts only to measure them and look for relations (correlations) between them. (Experimental research may also calculate "correlations" between variables, but usually these are between the manipulated variables and those affected by the manipulation.) Another feature present in experimental methods, the division of the target population into experimental groups, is also absent from correlational research. The expected relation between the examined variables may be theory driven; tests are performed to determine whether the variables expected to be related are, in fact, related, and reporting shows or describes these associations.

Correlational studies must be interpreted cautiously as spurious correlations often occur reflecting a relation that may be attributable to some unmeasured, yet shared factor. Correlational studies usually show much weaker effect sizes than experimental studies, also attributable to unmeasured or uncontrolled factors, or confounding interactions between factors. Additionally, it must be explicitly stated that correlations, even very high correlations, cannot be interpreted as causation; i.e. the most a researcher can claim

about two variables is that they relate to one another to some degree or extent. A causal direction to the relationship between variables cannot be expressed; a correlational study does not manipulate one variable to precede the other. However, correlational research can help form the starting point for research, leaving causal mechanisms to be explored at a future date under different experimental conditions (Mullan, 2014). Additionally, it is a useful method for predicting the levels of one variable based on knowledge of another variable.

Limitations

This study, its interpretation, and use of the results are limited by several important considerations. At the time of measure, no demographic information was collected on the teachers; thus, the unknown characteristics of the teachers (such as years in service, previous experience with gifted populations, levels of comfort with researchers in their classrooms, etc.) that may have had an effect on outcomes were not captured. The population for the study represents a sample of convenience and may feature variables unmeasured or undocumented that affect outcomes. Data cleaning may have had some unintended effects on the distribution of scores from the sample, specifically that lower-performing students may be overly represented in excluded cases. As per recommendation from the test authors those students with a MIR:R reading comprehension percentage of <20, (see below) were dropped from the data, as were incomplete cases with missing data, which may reflect a high degree of absenteeism, suspensions, etc. that affects distribution. It is also important to note that the data were

collected while screening for at-risk students, thus their use within a gifted context was not a consideration at the time of collection. No information about the reality of gifted services used by the district or the identification status of the students involved was collected.

There are other limitations related to the instrumentation. Results and interpretations must be limited to the utilized instrumentation; it should not be assumed that the same patterns and findings would hold when utilizing other forms of teacher ranking scales, other forms of CBM, or other state tests. Additionally, the study extends the intended scope and purpose of the instrumentation to validate the application of the metrics and processes for G/HA students, an extension planned by test authors but investigated here for the first time.

Delimitations

The definition of giftedness is intentionally restricted to academic and achievement measures in reading and math and does not address other aspects of the construct such as leadership, motivation, talent, or creativity; a practice much criticized by researchers in the field, but one justified by current practice as detailed below. Unlike the student data used for analyses, which reflects through data cleaning for this study some level of selection, data collected from teachers via the teacher ranking instrument used the entire third grade teacher population (no teachers were excluded for any reason). Similarly, the data were collected from every public elementary school in the district serving grade three (no schools were excluded for any reason). The study was conducted

on one school district in a south eastern state and grade specific for grade 3, though the instruments have measurement capabilities that extend to other grades. The results may not be interpreted as generalizable to other states, to other districts within the state, or even other grades within the same school.

Summary

Despite general consensus about the importance of implementing appropriate educational programming for students who are intellectually gifted and federal expectations (e.g., NCLB, 2001) that teachers use evidence-based practices, there is a lack of consensus about the quiddity of giftedness and how to identify it. As RTI increasingly becomes standardized practice around the country, researchers in gifted education are calling for protocols delimiting the use of RTI to screen for giftedness in a manner similar to the current use for screening at-risk students at the low end of the academic performance continuum. Many questions are as yet unaddressed or not addressed satisfactorily. For example, can traditional in-grade level CBMs typically used in RTI adequately screen for giftedness? Do they possess adequate ceiling and do they predict high stakes group achievement performance? What value do teacher perceptions, commonly used in gifted identification, add to the early identification of giftedness?

Chapter 2

Review of the Literature

Gifted Definition

The NAGC estimates that 3,000,000 academically gifted students may be found in U.S. classrooms (NAGC, 2012). The variance in their cognitive and academic profiles is exceeded only by that of the educational programming designed to serve them. It is self-evident that many societal and personal benefits might accrue to effective development of the advanced abilities exhibited by gifted children. Researchers have provided evidence that supports the contention that gifted students are at an increased risk for dropping out of high school or underachieving by a school's failing to meet their needs (Russo , Harris, & Ford, 1996; Stambaugh, 2001). "There is no absolute or universal definition of giftedness or system of identification" (Assouline & Whiteman, 2011). Giftedness is a "highly value laden term" (Volker, Lopata, & Cook-Cottone, 2006). Within the United States, most definitions derive from a federal law that defines gifted and talented children as those who are highly capable in general intellectual ability, specific academic domains, creative and productive thinking, artistic pursuits, or leadership (originally promulgated by Marland in1972). The identification as gifted indicates a high level of performance in an ability or domain of competence that is valued in a particular cultural or subcultural context.

Gifted, high ability (G/HA), and talented children are those who possess, or are capable of developing, a set of traits (above-average general or specific abilities, high

levels of motivation, or high levels of creativity) and of applying them to any potentially valuable area of human performance. The term refers to children and youths who, regardless of gender, or cultural or ethnic diversity, give evidence of higher levels of performance (or the capacity for higher levels of performance) in such areas as intellectual, creative, artistic, or leadership capacity, or in specific academic fields. They demonstrate atypical development in which advanced intellectual abilities and heightened intensity combine to create inner experiences and awareness that are qualitatively different from same aged peers, an exceptional ability to reason and learn, or competence (in top 10% or rarer) in one or more domains. Domains include any structured area of activity with its own symbol system (e.g., mathematics, music, or language) or set of sensorimotor skills (e.g., painting, dance, or sports). Outstanding talents are present in children and youth, regardless of gender, from all cultural groups and socio-economic levels, and in all areas of human endeavor (United States Department of Education (DOE), 1993). Gifted learners are a heterogeneous group who manifest their abilities in particular areas or pursuits; that is, gifted students are typically gifted *in* something (Tomlinson, 2005).

Within education at the federal, state, and district levels, giftedness is often defined in terms of intelligence and/or academic achievement, intelligence as demonstrated performance on any standardized test or other psychometric instrument providing an intelligence quotient (IQ) of at least two standard deviations above the mean (e.g. Stanford-Binet Intelligence Scales (SB5), Fifth Edition, published by Houghton

Mifflin Harcourt, 2003), advanced student achievement measured by grade point average (GPA) or eligibility for and participation in advanced coursework (e.g., honors and advanced placement classes), or by student performance on state-mandated testing. This definition, however, is not without detractors.

Adelson, McCoach, and Gavin (2012) found that *on average,* gifted programming provided no effects on *achievement* or *attitudes* in either mathematics or reading, regardless of the level (school or student) or the population of interest (gifted students or non-gifted students). Adelson also examined the opposite proposition and found that *on average,* gifted programming did not have negative effects on the achievement or academic attitudes of non-gifted students; that is, gifted programming does not appear to have detrimental effects on non-gifted students. "Thus, gifted programming, as the United States currently provides it, does not appear to affect gifted students' achievement…" (Adelson, McCoach, & Gavin, 2012, p. 33). However, Adelson notes the operative phrase is *on average*; suggesting that though some gifted education programs are increasing student achievement in reading and mathematics, the effects are neutralized by programs having either negative effects or no effects. Adelson states that this is all the more unfortunate, as research indicates the positive effects of many gifted education practices in specific contexts.

Federal Definitions of Giftedness

Individuals with Disabilities Education Improvement Act (IDEA); No Child Left Behind Act (NCLB).

Neither of the legislative acts that most dominate education policy and practice in the United States today, the *Individuals with Disabilities Education Improvement Act* of 2004 (IDEIA, IDEA 2004; revised and updated IDEA 2013; IDEA should be read as referring to this act in its latest iteration unless otherwise noted) and the *Elementary and Secondary Education Act* (ESEA; Public Law 107-110), commonly known as *No Child Left Behind Act* (NCLB, 2001) have specific mandates to identify and serve gifted students. Though little of actionable practice in gifted education emanates from these acts, they, in combination with other federal policies in funding and educational testing, do contribute some important ideas, if only to recognize that giftedness exists, can ostensibly be quantified and measured, and may be manifested to such a degree as to make the unaccommodated, general education classroom a restrictive environment. NCLB provides a definition of giftedness and requires progress monitoring of all students; IDEA fails to include giftedness, but provides frameworks of screening, identification, and programming used for other disabilities that may be adapted for gifted populations: the J. K. Javits Act (1988) funded gifted programming for states for many years and, after being unfunded for several years, now supports research on gifted education through grants; the National Assessment of Educational Progress (NAEP) tests establish reading and math as essential domains when evaluating educational progress,

allow for inter-state comparisons between domains and grades tested, and delimit parameters of advanced attainment.

IDEA is a federal law detailing 13 disabilities designated as special education categories. IDEA mandates that children with these disabilities must be identified and served through specific programming designed to meet their needs in the least restrictive environment and provides that federal and state funding be allocated to do so. The programs, the academic progress and affective disposition of the recipients, and the dispersion and use of the allocated funds are all closely monitored by federal, state, and local agencies. IDEA does not mention giftedness in any way. NCLB legislation created a new, achievement-based definition of giftedness, however it does not mandate that states use its definition:

> The term "gifted and talented", when used with respect to students, children, or youth, means students, children, or youth who give evidence of high achievement capability in areas such as intellectual, creative, artistic, or leadership capacity, or in specific academic fields, and who need services or activities not ordinarily provided by the school in order to fully develop those capabilities. (Title IX, Part A, Section 9101(22), p. 544).

The act neither includes nor specifically excludes mandates for gifted learners (NAGC), 2003). Consequently, many states compromise services for the gifted, focusing resources on the attainment of the more specifically detailed metrics for mandates in the

legislation such as those for lower-achieving students (Brown et al., 2006). However, research evidence supports the contention that gifted students are at an increased risk for dropping out of high school or underachieving by a school's failing to meet their needs (Russo, Harris, & Ford, 1996; Stambaugh, 2001).

Gifted education once received discretionary and other limited federal funding through the Jacob K. Javits Act (1988). However, recent funding for the Act has been sporadic. The Act received no funding for the years 2011 through 2013, $5 million in 2014 for research and discretionary grants, and $10 million in 2015 for research and new awards (U.S. Department of Education, 2015) causing the NAGC to state, "With the lack of a federal policy, mandate, or funding as a backdrop, the current condition of gifted education in the states is mixed" (2007). It is worthwhile to note that the Act has not provided any funding for state programs to provide services for gifted students, but instead offered funding for, on average, six research grants per year.

National Assessment of Educational Progress (NAEP).

Beginning in 1969, the Department of Education has produced a report entitled the *Nation's Report Card* using data from a series of specially designed NAEP tests. The report is intended to inform the public about the academic achievement of elementary and secondary students in the United States. Since 2003, NAEP national and state assessments have been conducted in reading and mathematics at least once every two years at grades 4 and 8; some NAEP assessments are conducted at the national level for grade 12, as well. Assessment domains include reading, mathematics, science, writing,

U.S. history, civics, geography, and other subjects. NAEP collects and reports academic achievement results at the national level, and for certain assessments, at state and district levels.

Federal law initially specified that participation in NAEP testing was *voluntary* for every student, school, school district, and state. However, federal law now requires participation of any state or school district receiving Title I funds. Currently, as per NCLB (2010) testing is *required* in two domains; all states must administer NAEP reading and mathematics assessments for grades 4 and 8 every other year. NAEP assessments are also used to monitor results from state testing programs through comparison between state tests and NAEP results in the corresponding grades and content areas. In addition, NAEP tests must be administered in reading and math on a nationally representative basis at grade 12 at least every four years. Mandatory participation in NAEP testing has been in place since the 2003 testing cycle.

NAEP results are important and germane to gifted research for several reasons, least of which is that by requiring testing in reading and math these two domains are established as essential when measuring academic progress. This is to the exclusion of other domains such as science and social studies. NAEP tests also establish performance criteria that operationalize levels of ability, including advanced-level performance metrics. Thus the NAEP testing comes closer than any other metric to a nation-wide assessment of identical content while establishing criteria for advanced levels of performance. Moreover, the domain specific operationalizing of advanced performance

provides metrics that may be used for comparison when seeking to validate new instruments as the NAEP measures are an example of norm-referenced, domain-specific testing. This study uses performance metrics in the two domains, reading and math, that seem indicated as most important as evidenced by practices dictated at the federal level.

State Definitions of Giftedness

State definitions display a wide range of attributes when defining giftedness. A survey by the National Association for Gifted Children for its report *State of the States in Gifted Education* (2014) received self-reported responses from states regarding the characteristics included in the state definition of giftedness and identification practices. Of the 43 states responding to the question, only three states reported that the state has no definition of giftedness. Thirty-nine states responded to questions about individual identifiers of giftedness included within the state definition from a selection including *intellectual achievement, academic achievement, creativity, leadership*, advanced abilities in the *performing and fine arts, other indicators* such as advanced abilities in technology, and an explicit acknowledgement that giftedness might be found in *diverse populations* such as low socio-economic status, linguistically and culturally diverse populations, or concomitant with other disabilities such as specific learning disabilities. Of these seven characteristics, no single attribute was present in all the definitions of reporting states, though most reporting states included intellectual giftedness as a primary marker. Happily, 30 of 39 reporting states have a definition that includes some attention to diversity in the profile of gifted characteristics. Only two states, California and

Colorado, include all seven indicators in their definition. The mean number of identification characteristics present in most state definitions was three; meaning most reporting states (9) include three or more attributes of giftedness in their state definition of giftedness. However, there is no consensus upon which three attributes are included. Of the states reporting, six states have a definition that includes only one (four states) or two (two states) of the characteristics listed. *Leadership abilities* and *advanced levels of performance in the performing and fine arts* are the least frequently included. Sixteen states with mandates to serve gifted students report that the mandate extends to grades K-12. An additional four states include Pre-K. Two states serve grades 3-12; one state serves grades 1-8. Other states either have no policy or no explicitly stated policy.

Table 1. Gifted Attributes Included in State Definitions of Giftedness

Attribute	Included	Not included	Percent Included
Intellectually Gifted	38	1	74.5%
Academically Gifted	27	12	69.2%
Creativity	24	15	47.1%
Leadership	15	24	29.4%
Performing/Fine Arts	18	21	41.2%
Diversity	30	9	76.9%
Other Indicators	24	15	47.1%

Note: 39 states responding

To answer the research questions presented in this research, the definition of giftedness will be restricted to one most generally used by states in their legal mandates, that is one of an intellectual or academic nature. While acknowledging that to many this definition is unacceptably limited, it is consistent with what is mandated in most states and falls within the purview of the legislated responsibilities of state and local education agencies.

Identification Practices

As with everything related to gifted education, it is difficult to determine who the gifted are and how to best identify them. Though an aggregation of federal policy may support gifted education in theory, the dearth of applicable legislation certainly stalls gifted education in practice. Identification protocols vary widely across the United States. As it is typical to develop programming and services around the needs of the service recipients, it becomes prudent to ask questions concerning the nature and characteristics of gifted students who will be the ultimate benefactors of such programming. Operationalizing giftedness has been a process fraught with contention; researchers complain that definitions are too restrictive, failing to include the many permutations of giftedness; local and state education agencies (LEA, SEA) seek more limited definitions that include parameters that fall within the scope of the legal mandate of the services schools are charged to provide. Zirkel (2005) states that all decisions concerning gifted programming are local decisions. The veracity of this statement is evidenced by an examination of state legislation for gifted identification and services.

Two of the primary goals for providing gifted education are 1) to increase learning and achievement to a level matching students' potential, and 2) to enhance the self-concept of gifted students by allowing them to interact and learn with like-ability peers with similar interests (see Delcourt, Cornell, & Goldberg, 2007; Rogers, 2007). A goal of this study is to provide an increased understanding of gifted screening and identification practices at the state level through a test case examining the correlation between early- and end-of-year measures, a comparison of formal and informal measures, qualitative and quantitative measures, as well as experimental probes and norm-referenced tests.

Principles of Identification

As the definition of giftedness continues to expand and enfold other aspects of the construct, such as *creativity* or *leadership*, for example, it is now commonly accepted that the identification process will involve multiple measures. The use of a *multiple criteria method* includes the consideration of a wider variety of cognitive abilities, as well as other facets of the construct such as *creativity*, *achievement*, *motivation*, *leadership*, etc. (Volker et al., 2006). Best practice in gifted identification now also includes the use of *multiple sources* of data, such as academic progress (grades); nominations from teachers, parents, or peers; test data; or school products and portfolios (Gallagher, 1994; Sternberg , 1998). The interpretation of test results should involve sensitivity to important factors in the examinee's profile such as cultural background; possible motor, sensory, or

learning disabilities; known errors in instrumentation (Kaufman & Harrison, 1986), and inconsistencies in the development of intelligence.

Researchers from the NAGC (2010) provide a summary of the major principles of identification for underserved gifted students, which is based on available research:

- Select a broad definition of giftedness with which to assess, going beyond cognitive abilities.
- Use a multiple criteria approach (performance assessment, portfolios, dynamic assessment, nominations).
- Use unique and appropriate identification strategies to identify different aspects of giftedness, making use of reliable instruments and strategies, while considering the reliability and validity data for the populations assessed, norms, and cultural bias in instrumentation.
- View each child as an individual and recognize limitations of a single score on any measure.
- Recognize the serious limitations of matrices in the identification process.
- Identify and place students based on student need rather than by a pre-determined program limit.
- Create a larger talent pool to allow more students to further develop their talents and abilities.

- Design an identification process that is varied, wide-reaching, and ongoing to ensure that students whose abilities are masked by environmental circumstances are part of the process.

District-wide considerations from the NACG (2010) include:

- A district-wide operational definition of giftedness used for screening and identification purposes.
- Selection of suitable screening and identification protocols that include multiple metrics that align with the district's working definition of giftedness and identify reliable and valid screening and identification instruments.
- Hiring of personnel who have been specifically trained in the affective and educational needs of gifted and high ability children and in teaching or counseling methods proven to be the most efficacious.
- "Buy-in" from invested stakeholders at all levels; that is, educators who are invested in meeting the needs of G/HA students.
- Flexibility to accommodate needs-based acceleration; be it in a single domain, or through grade skipping, radical acceleration, early entrance to Kindergarten or Pre-kindergarten levels, or dual enrollment with a college or university.

State Methods of Identification

As reported through the survey by the NACG for its report *State of the States in Gifted Education* (2014), 29 states have a legal mandate for the identification of gifted students. However, as might be expected, there is much variation in the methods

specified by states for use in the identification process. For example, only 14 states require LEAs to use the same identification method, and state policy leaves most identification protocols to be decided at the local level. As has been noted, best practice in identification indicates that the identification process should utilize multiple methods, which is to say that reliance upon any single score or metric is discouraged. Generally, this is known as the *multiple criteria method* (MCM) and is a prevalent legal stipulation in the identification practices of many other special education categories as well. Thirteen of 42 responding states provide specific language at the state level regarding the methods used for gifted identification. Despite this, most states leave these decisions to the LEA through three mechanisms; 1) by a failure to include any specific language in state policy, 2) through the use of language that explicitly grants this power to the district level, or 3) the inclusion of language that provides districts the ability to override state policy, even when state policy explicitly disallows a given policy.

In an exploration into the nature of talent identification programs such as *Carnegie Mellon University Institute for Talented Elementary and Secondary Students* (C-MITES), the *Center for Talent Development* (CTD) at Northwestern University, the *Center for Talented Youth* (CTY) at the Johns Hopkins University, the *Connie Belin & Jacqueline N. Blank International Center for Gifted Education and Talent Development* at the University of Iowa, the *Rocky Mountain Talent Search* at the University of Denver, and the *Talent Identification Program* (TIP) at Duke University, Lee, Matthews, and Olszewski-Kubilius (2008) found identification protocols align with the type of

programming offered. Early administration of norm-referenced tests such as SAT® (The College Board, www.collegeboard.org), ACT®, ACT Explore®, ACT Plan® (ACT, http://www.act.org/research/) form the principle component of identification for these programs. The authors also observe that the identification criteria are "generally conservative, centering on indicators of academic achievement" (Lee, Matthews, & Olszewski-Kubilius, 2008, p. 65). They report frequent use of

- scores on in-grade achievement tests for enrichment-oriented weekend programs
- student portfolios for summer and leadership programs
- scores on off-level tests for accelerated classes
- teacher and school recommendations for summer, Saturday and weekend, and leadership programs
- in-grade standardized achievement tests in distance education and summer programs
- parent nomination for entrance to Saturday and weekend programs

The NAGC *State of the States Report* (2014) also includes survey results from questions concerning the criteria used by states in the identification of gifted and high ability students. Thirty-nine states responded to questions about acceptable methods of identification mandated by state legislation from a selection including the use of *IQ* scores, *academic achievement*, *multiple criteria method*, allowing the LEA to *select* from a state provided list of appropriate tests, and *nomination*. It should be noted that the processes described by these categories may overlap. Nearly half of the responding states

specifically allow the use of the multiple criteria method (MCM) as indicated by best practice. Though no method of identification is explicitly included in the language of the *National Research Center on the Gifted and Talented* (NRC/GT) summary of the major principles of identification, the use of nomination as an identification method logically proceeds from the recommendations included such as the use of multiple measures. Three states (Indiana, Oklahoma, and Oregon) allow all five methods to be used. Only seven states specifically allow teacher nomination, though most state laws contain language that allows LEAs to make this determination as well.

Table 2. Methods of Identification of Giftedness Allowed in State Definitions

Method	**Allowed**	**% Allowed**
IQ	17	33.3%
Academic Achievement	15	29.4%
Multiple Criteria Method	23	45.1%
Selection from List	14	27.5%
Nomination	7	13.7%

Note: 39 states responding

IQ measures are not typically available for teachers to use when making initial educational programming decisions. The focus of this study is two measures that can be used as part of a multiple criteria method. As important as which measures are used, is the question of when these measures are available. This study uses an informal teacher ranking form that could, theoretically be available at any point during the academic year, but one which was taken in September, and subsequently defined as an early-in-year

measure. A second measure of academic progress is included, universal screeners (US) in reading and math, the *Monitoring Instructional Responsiveness: Reading* (MIR:R; Bell, Hilton-Prillhart, McCallum, Hopkins, 2012) and the *Monitoring Instructional Responsiveness: Math* (MIR:M; Bell, Hilton-Prillhart, McCallum, Hopkins, 2012), measures collected concurrently, early in the school year with the teacher ranking. The relation of these measures to the Tennessee Comprehensive Assessment Program (TCAP, 2013), a state-mandated, criterion-referenced high-stakes test given in the spring each year in grades three through 12 was examined. TCAP data become a compelling factor in gifted identification in the state as testing results are a primary consideration when completing a gifted identification matrix as required by the state for identification. It is not, however, available to third grade teachers due to testing grades and schedules. TCAP information is available to teachers of the fourth grade and above as a guide for decision making in placements and program eligibility.

Identification by Teacher Nomination

This study uses a teacher ranking instrument as described below. Teacher nomination is here predicated by the inference that teachers are more likely to nominate students whom they rank towards the top of the class. The effectiveness of teacher nomination as a method of gifted student identification is another topic in gifted education that lacks a definitive consensus. The literature base fails to make any distinction between *teacher nomination* and the ancillary topic of *teacher perception* of giftedness. The effectiveness of nomination seems highly contingent upon teacher

perception, such that experienced teachers (Siegle, Moore, Mann, & Wilson, 2010) with some training in the characteristics of gifted students (Speirs -Neumeister, Adams, Pierce, Cassady, & Dixon, 2007) generally nominate more students to gifted programs. Historically, nomination has been viewed as inaccurate and was initially viewed with much disfavor (Pegnato & Birch, 1959; Hoge & Cudmore, 1986). However, Hoge and Cudmore (1986), in an extensive literature review of research concerning teacher perception measures, concluded that "the use of teacher judgments in the identification of gifted children should be continued, and, in fact, expanded" (p. 192). Significantly, when re-evaluating the original Pegnato and Birch data, Gagné (1994) revealed several erroneous conclusions and found that teacher nomination was as effective as other methods of identification of gifted students, a finding since corroborated by other researchers (Rohrer, 1995; Hodge & Kemp, 2006). Peters and Gentry (2012) state that teacher ratings of explicit behaviors, as opposed to general teacher opinions and perceptions, more consistently identify gifted characteristics successfully. Lacking concrete parameters, the resulting teacher nominations do not appear to be especially accurate (Peters & Gentry, 2012).

Many different gifted observation and nomination scales have been developed for use by teachers, parents, and others, which can provide valuable insights about a student's specific strengths (Elliott, Busse, & Gresham, 1993; Feldhusen & Heller, 1986). These scales require teachers to observe and rate general gifted characteristics such as learning, motivation, and creativity. Renzulli et al. (2009) note that research-based scales

in specific content areas have limited previous research, and that most checklists for these purposes, if available, provide only anecdotal information. The findings of several researchers who have investigated the validity of teacher nomination confirm that when given specific rating criteria, teachers were better able to identify talented students (Borland, 1978; Gagné, 1994; Hoge & Cudmore, 1986; Hunsaker, Finley, & Frank, 1997; Johnsen, 2004; Pegnato & Birch, 1959; Renzulli & Delcourt, 1986; Rohrer, 1995; Siegle & Powell, 2004). However, Speirs -Neumeister et al. (2007) caution that teachers may "rely exclusively on characteristics of gifted students that appear on published checklists without realizing that all gifted kids do not demonstrate all of the characteristics" (p. 480). Schroth and Helfer (2008) examined the gifted identification beliefs of school personnel and found that *teacher nomination* was believed to be the second most effective identification method, ranking ahead of *standardized tests*. *Performance assessment by experts* was reported as the most effective. This contradiction elaborates an important theme prevalent in all gifted education research, to wit, any decision about gifted children is best made on an individual basis by those who understand giftedness generally and who have a specific familiarity with the child in question.

Teacher training in giftedness either during pre-service coursework, continuing education, or professional development is clearly an additional factor in the accuracy of teacher nomination. Teachers with training in giftedness are more likely to recognize different expressions of giftedness (Siegle, Mann, & Wilson, 2010). Still, Speirs-Neumeister et al. (2007) report that even experienced teachers often hold a "narrow

conception of giftedness" and are not aware "how culture and environmental factors may influence the expression of giftedness in minority and economically disadvantaged students" (p. 479). It is worthwhile to note that much of the research of nomination and, hence, teacher perception in identification protocols investigates under-representation of minorities and those with cultural or linguistic differences as a function of teacher bias in nomination.

Bianco (2010) states that persistent concerns about under-representation and the lack of diversity in gifted identification and gifted programming, including students with disabilities, perpetuate the conceptualization of gifted programming as *elitist* (see Bernal, 2002; Bianco, 2005; Ford, Grantham, & Whiting, 2008; Sapon-Shevin, 1996). Ford and Grantham (2003) and Valdés (2003) also report that a lack of teacher referrals for students of color and those who are culturally and linguistically diverse as factors contributing to under-representation in gifted programs. Additionally, teachers whose focus is on what students cannot do (labeled as a *deficit model*) may have a "blurred" perception of student ability, and, as a result, the gifted abilities of some students may go unrecognized (Bianco, 2010). This is one reason to implement universal screenings throughout the year. In fact, Volker et al., (2006) state that gifted identification should be possible throughout the entire school experience of a child and should not be a based on a single qualifying opportunity.

Only five states require *all* teachers to receive pre-service training in gifted and talented education (Delcourt, Cornell, & Goldberg, 2007). In a national survey of

teachers conducted by the NAGC and available on their website, 73% of teachers agree with the statement, “Too often, the brightest students are bored and under-challenged in school; we’re not giving them a sufficient chance to thrive.” Still, 77% of teachers also agree that “getting underachieving students to reach proficiency has become so important that the needs of advanced students take a back seat” (Farkas & Duffett, 2008). The existence of such a counter-intuitive dichotomy is indicative of the dilemma faced by gifted children and their parents in the programming policies of and services provided by public schools today.

Many school districts allow teacher ratings of students as part of the selection criteria using teacher input as an additional tool to screen a pool of students to be further tested for gifted programs. Teacher nomination may be one measure of identification protocols that may also include standardized achievement tests, portfolio review, performance assessment, intelligence tests, etc. Teacher nomination may be included as a metric in a formal matrix of required components used to identify gifted students (Ash & Huebner, 1998, Bain & Bell, 2004; Wu & Elliott, 2008) or as one measure of the multiple criteria method (Frasier & Passow, 1994; Maker, 1996; Plucker, Callahan, & Tomchin, 1996).

An informal teacher rank form (TR) that requests teachers to rank students by their performance, independently in reading (TR:R) and math (TR:M), with “1” as the highest, was used in this study. To help in triangulating data sources, and in a manner consistent with the use of multiple criteria, this qualitative measure was correlated with

more quantitative data using results from the MIR:R and MIR:M, and was included as a metric taken early in the academic year. These sources are related to end-of-year measures.

Rates of Gifted Identification

Rates of Identification on the Federal Level: NAEP Data

For comparison to state gifted identification rates, rates to measure actual performance were collected from NAEP for the year 2011, as available from the National Center for Education Statistics website (NCES, https://nces.ed.gov/nationsreportcard/; 2015), were examined. It must be explicitly stated that this comparison serves merely as a reference. The measures below from Davidson are neither subject nor grade specific, while the following NAEP data are disaggregated by subject and are reported for the fourth grade only. NAEP tests establish performance criteria that operationalize levels of ability, including advanced level performance metrics. Thus the NAEP testing comes closer than any other metric to a national assessment of the same content while establishing criteria for advanced levels of performance.

Outcome measures utilized by this research include the use of achievement levels, which are performance standards set by the National Assessment Governing Board of NAEP. Mean state scores are used to compare average state performance. Cutoff scores are used to delimit four levels of student performance into easily understood yet general categories: *Below Basic*, *Basic*, *Proficient*, and *Advanced.* Achievement-level percentages, the percentage of students within the given population

who meet or exceed the performance indicators of each level, are included as a selection parameter for data searches. These percentages reflect the weighted percentage of students with NAEP composite scores that are equal to, or exceed, the achievement-level cut scores specified by the National Assessment Governing Board. Of interest is the percentage of the total state population that attains the *advanced level*, denoting superior performance. The cut scores for determining advanced levels of achievement are specific to the domain of the test and, unlike other outcome measures, do not typically vary from year to year.

The NAEP reading assessment measures a student's reading and comprehension skill, defined on the NCES website as "… a dynamic cognitive process that allows students to understand written text, develop and interpret meaning, and use meaning as appropriate to the type of text, purpose, and situation." The achievement-level descriptions were updated in 2009 to reflect a new reading framework. The specific descriptions of reading achievement in grades 4, 8, and 12 are presented on the website. Average reading scale score results are based on the NAEP reading scale, which ranges from 0 to 500. The cut score indicating the lower end of the score range for each level *Basic*, *Proficient*, and *Advanced* reading achievement levels are presented below. The fourth category, *Below Basic*, is comprised by those scores lower than the *Basic* cutoff.

- Grade 4: Basic (208), Proficient (238), Advanced (268)
- Grade 8: Basic (243), Proficient (281), Advanced (323)
- Grade 12: Basic (265), Proficient (302), Advanced (346)

The NAEP mathematics assessment measures a student's knowledge and skills in mathematics and the ability to apply knowledge in problem-solving situations. Questions are designed to measure one of the five mathematics content areas: number properties and operations, measurement, geometry, data analysis, statistics and probability, and algebra. Some aspects of mathematics, such as computation, occur in all content areas. Specific definitions of the *Basic*, *Proficient*, and *Advanced* achievement levels for grades 4, 8, and 12 are presented on the website; cut points are included below. The achievement-level descriptions and cut points for grade 12 were updated in 2005; consequently, the 2011 test cycle results for grades 4 and 8 are reported on a 0–500 scale, while results for grade 12 are reported on a 0–300 scale.

- Grade 4: Basic (214), Proficient (249), Advanced (282)
- Grade 8: Basic (262), Proficient (299), Advanced (333)
- Grade 12: Basic (141), Proficient (176), Advanced (216)

Rates of Identification on a State Level

Data on the total enrollment and gifted enrollment by state were collected from the individual state reports found on the website of the Davidson Institute for Talent Development (2014). The institute was formed in 1999, with the stated mission to "… recognize, nurture and support profoundly intelligent young people and to provide opportunities for them to develop their talents to make a positive difference." The Davidson Institute is a 501(c) 3 private operating foundation funded by Bob and Jan Davidson. The institute engages in several activities. The Davidson Academy of Nevada

is a free public school for profoundly gifted middle and high school students, those who score in the 99.9th percentile on IQ or college entrance tests, such as the SAT or ACT. The academy was created by the Nevada State Legislature in 2005 to provide direct support to profoundly gifted young people 18 and under. Thirty-three states reported data on the number of gifted children identified; this was divided by the reported total enrollment to calculate the percentage of gifted students. These percentages should be interpreted with caution and likely represent a very gross measure. The range of gifted identification rates in the United States is wide. West Virginia and Tennessee identify giftedness at the lowest rates, 1.92% and 2.02% respectively. In contrast, Kentucky and Virginia identify students at a rate above 16%. The mean rate of gifted identification for 33 reporting states is 8.43%.

Table 3. Comparison of State and the National Assessment of Educational Progress (NAEP, 2010) Identification Rates

	N	Range	Min	Max	*M*	*SD*
Average % gifted by State	33	14.53	1.92	16.46	8.43	4.68
Average NAEP:R Grade 4, % advanced	51	12.05	3.46	15.52	7.38	2.49
Average NAEP:M Grade 4, % advanced	51	11.35	2.12	13.47	6.32	2.49

Note: NAEP:R National Assessment of Educational Progress: Reading, NAEP:M National Assessment of Educational Progress: Math, *N* = number of states with available data (includes Washington D. C.)

NAEP testing indicates that rates of advanced performance on norm-referenced testing show a comparable range, minimum, maximum, and mean when compared to the less formal state measure from the Davidson website reported above. The mean percentage of students who attain at the advanced level performance on NAEP is 7.38% for reading and 6.32% for math for the 51 states (including the District of Columbia) in 2011. Identification rates range from 3.46% to 15.52% in reading and from 2.12% and 13.47% in math. The rates of advanced performance on NAEP reading and math tests are significantly correlated (Pearson $r = .837$, $p < .001$). The state, Maine, with the highest percentage rate of actual performance in reading and math (15.52% and 13.47% advanced, respectively) reports a gifted identification rate of 3.60%. Eighteen of 33 states report identifying giftedness at higher rates than seems indicated as appropriate by actual performance. These comparisons of identification rates of giftedness indicate that, in this case, policy, performance, and practice are too inconsistent to determine anything practicably meaningful.

This lack of consistency necessitates a return to theoretical practices as established by leaders in the field of gifted education. Renzulli (2010) recommends a screening cutoff of the top 15% and the use of local norms. He is also a proponent for the inclusion of high ability students who, when provided appropriate interventions, may attain at gifted levels. Renzulli (2010), Subotnik (2010), and Adams (2010) all stress the importance of using local norms when establishing cut points, a stance consistent with the idea that in order to be effective, programming decisions are best made at a local level to meet the specific

needs of the students served. Screening rates are generally less rigorous than identification rates to include high ability. Identification rates are frequently reported with an expected rate of 3%-5%, delimited in part by the statistical proportion of the population who perform two standard deviations above the mean.

The Tennessee gifted identification matrix uses several paths for gifted identification. Much weight is given to TCAP performance and cutoff scores either at the 90^{th} percentile in two academic domains, or the 95^{th} percentile in a single domain for identification; thus, screening at the recommended cutoff of 85% seems appropriate when searching for students whose performance may warrant closer scrutiny. To answer the research questions the screening rate closest to the 85^{th} percentile was used for the MIR:R and MIR:M. For the TCAP, the *Advanced* category in reading and math scale scores was utilized for the analyses. The top two rankings on the Teacher Rank form were used, with class sizes ranging from 9 to 16, thus the top two ranks represent a variable percentage of the class ranging from 22% to 12.5% respectively.

Response to Intervention, Curriculum-Based Measures, and Universal Screening

The extension of processes and protocols designed to screen, identify, and serve at-risk students to above-grade levels of performance are examined in this study. The processes established for special education referral for at-risk students require specific validation as appropriately applicable to gifted and high-ability students.

Response to Intervention (RTI)

IDEA embeds a process that has become known as Response to Intervention (RTI, RtI, or in Tennessee, Response to Instruction and Intervention- RTI^2) introduced in the 2004 amendments that went into effect on July 1, 2005. However, the language of IDEA does not explicitly refer to "response to intervention." The specific learning disability (SLD) category had come to represent the highest percentage of students receiving special education services, including over half of the total students served. Congress was moved to act concerning the increasing numbers of students in the SLD category under IDEA (identification increases of approximately 200%; Kavale & Spaulding, 2008), and an ancillary concern that many students might have avoided the need for special education services if appropriate instructional supports and interventions had been provided to them earlier. To militate against such increases, more specific identification protocols now known as response to intervention (RTI) were conceived as a special education reform to the screening and identification process of potential SLD identification with mandates for increased documentation by the general education teacher before referral to special education.

It is a misconception that IDEA mandates the use of RTI. IDEA merely permits the use of RTI, rather than mandates its use; that is, according to IDEA, a state may not prohibit the use of RTI (34 C.F.R. § 300.307[a]). It is another common misconception that IDEA provides for the use of an RTI process that extends beyond the identification of SLD (Daves & Walker, 2012). Zirkle (2011) observes that a "… careful review of the

IDEA legislation and regulations clearly reveals that the only reference to and recognition of the use of… RTI… is limited to the identification of students with SLD" (see 20 U.S.C. § 1414[b][6][B]; 34 C.F.R. §§ 300.307, 300.309, and 300.311).

Multiple reciprocal references exist between the NCLB and IDEA creating a commonality of ideas and language, such as an emphasis on scientifically rigorous interventions, which blur the origins of RTI and may contribute to the surrounding confusion. Though, like IDEA, NCLB does not specifically articulate the process, NCLB includes language that has come to describe the RTI process: (a) the use of scientific, research-based interventions in general education, (b) measurement of student response to the intervention, and (c) the collection, then use of, response data to modify the type, frequency, and intensity of interventions.

RTI is also linked by IDEA and NCLB through common funding sources, especially Title I funds, used to support the lowest-achieving, at-risk students, those most likely to benefit from the RTI process. NCLB authorizes Title I funds for staff, training, and resources for students struggling in reading and math in low-income schools. LEAs may use up to 15% of funds from Title I and Title III of NCLB and from Coordinated Early Intervening Services (CEIS) to support RTI in public schools, assisting students who are not currently identified as needing special education or related services, but who may need additional academic or behavioral supports for improved success in the general education environment.

NCLB expands the application of the RTI model from special education to the general education classroom; thus, what began as a special education reform of the methods used to identify students with specific learning disabilities, evolved into a primary feature of the manner in which general education programs identify and remediate struggling students, increase teacher accountability, document student progress, and provide funding sources for the resources needed to do so. RTI enfolds many concepts relevant to this research. Interventions and grouping practices; progress monitoring through CBM; frequent, informal assessment of student progress including universal screening; and data-driven decision making are all integrated within the RTI model. This legislation applies to all programming and services offered in schools, including special education which frequently subsumes gifted education, though noting again that gifted education, per se, is not included.

The Tennessee State Board of Education approved changes to Special Education Guidelines and Standards such that as of July, 2014, all districts and schools are required to replace the use of a discrepancy model to that of RTI to determine eligibility for special education services in the SLD category. Additionally, the 2015 revision of Tennessee State Common Core Standards (TNCore), Response to Instruction and Intervention (RTI^2) manual clearly states that RTI is a path for providing instructional opportunity to "… any student struggling for success…" and should not be construed as a path to special education eligibility (p. 7).

RTI has typically maintained a focus on those performing significantly below their peers on CBM, but more recent applications of RTI have expanded the model for above-grade students who may be gifted. The Association for the Gifted, a division of Council for Exceptional Children (CEC-TAG), has recognized the potential of adapting the RTI framework for gifted learners, and the CEC now recommends that the "RTI model be expanded in its implementation to include the needs of gifted children (CEC, the Association for Gifted, 2009, p. 1)." However, a clearer distinction should be made between recommendations for the use of RTI as a process used to *identify* gifted learners (such as requirements of universal screening tools for identifying gifted status) and the use of RTI to *serve* gifted learners (such as suitable progress slopes toward accelerated learning goals derived from CBM data), which are, in fact, two different considerations. Many scholars recognize the need for developing RTI protocols designed to screen and identify G/HA students in a manner similar to other current special education identification procedures. A lack of consensus exists concerning validated procedures for the role RTI might potentially have in screening and identifying gifted and high ability students. Research is needed to demonstrate that features of RTI such as universal screening, CBM and progress monitoring requiring multiple data sources can be used effectively to adequately screen for and serve gifted students.

In the RTI^2 Implementation Guide, composed for implementation of RTI^2 for academic year 2013 and revised in July, 2014 (still available on the Tennessee Department of Education website at http://www.tn.gov/education/article/special-

education-evaluation-eligibility), the provisions for serving gifted students are muddled but perhaps tenuously hopeful in terms of meeting the needs of the gifted population. A *Forward* by the State Commissioner states that, "It is my fundamental belief that all students are able to reach higher levels of academic achievement…" (p.6, emphasis retained), but afterward refers only to the handbook's elucidation of "…best practices in closing gaps for students who struggle." However, both the 2013 and revised 2014 manuals specifically mention the application of the model to *gifted* students, a classification attained, it is assumed, by using the state's legislated definition. In *Section 2.7 Resources for High Achieving Students within an RTI² Framework* guidelines are found to ensure that gifted students have access to "differentiated curriculum, flexible pacing, cluster grouping, acceleration and other universal interventions available to all students in the regular classroom" (p. 114). The section continues to outline many other instructional strategies and interventions that are widely recognized in the literature as best practice in gifted education including the use of formative assessments that continually provide new data for monitoring progress such as those available from RTI settings.

Significantly in yet another revision of the RTI^2 manual in January, 2015, as part of the adoption of the Common Core State Standards (and currently available at http://tncore.org/rti.aspx; though as per this link, the site is migrating early in 2016 to http://www.edutoolbox.org/), gifted students and their place in the RTI^2 process are not mentioned. Instead, references are made to *advanced* students, who are not specifically

labeled as *gifted* but rather obliquely described as those who *exceed* expectations. Clarity ensues, as the manual explicitly states that these students (as well as those who *meet* expectations, it should be noted) are served utilizing strategies and enrichment provided within the general education classroom, with tier placement limited to Tier 1; that is, advanced students recursively retain their Tier 1 placement after enrichment. The protocol provides for appropriate, even generous use of universal screening for identification of struggling students; however, the protocol fails to mention any application of universal screeners to gifted or advanced students.

Clearly, the current 2015 Tennessee RTI^2 model is not intended as a process for *identifying* gifted students or as a requisite process for gifted identification before a special education referral and its attendant diagnostic evaluation. The RTI^2 process for advanced students is intended to limit services to in-class differentiation and enrichment provided in Tier 1.

Gifted identification and acquisition of services as such, then, is only possible through the paths outlined in the 2010 Tennessee State Plan for the Education of Intellectually Gifted Students (available at http://www.tn.gov/education/article/special-education-evaluation-eligibility and considered as current and viable at this time). These paths are detailed in matrix discussed below, which allows for the use of data collected during the universal screening process, through teacher/parent/student recommendation, consideration of academic performance, and other indicators, but with state-mandated end-of-year testing (TCAP) as the most heavily weighted component. Negative

implications about the state's commitment to identifying and serving gifted learners necessarily accrue to the devolution of state protocols and policies evidenced as the changes between 2013 and 2015.

Ideally, research proceeds from the known to the unknown, from the concrete to the abstract, from hypothesis to theory. Through an examination of law, public policy, peer-reviewed research, actual practice in the real world, and an understanding of human development, the preceding analyses force the conclusion that in gifted education, such is not the case. Experts, researchers and leading authorities seek a broad understanding of giftedness encompassing a range of manifestations and concomitant affective considerations. Legal statutes and court definitions, as well as state definitions of giftedness constrict the operationalizing of giftedness to a circumspect set of characteristics that not only fails to attain a uniform consensus throughout, but also fails to receive widespread support among academics. Though some conclusions may be made from an examination of actual practice in gifted education, there is little consistency and few agreed upon practices. There is no consistent rate of identification as shown by percentages of gifted students identified within state populations. When percentages of actual gifted identification are compared to actual incidence of advanced performance on testing metrics, the dissonance is resounding. Methods for identification of gifted students necessarily spring from the definition, which having been established as inconclusive, logically indicates that any given identification practice may be shown to be applicable and appropriate in some, but not all, situations. This study, then, is

conducted under parameters that reflect only the broadest understanding of, and within areas of general agreement in, giftedness and gifted education, applying theory in instances where practice fails. Concisely, as justified by this research and as explained below, giftedness is defined as academic success at an advanced level, significantly different from that of same-age peers, and limited to two subject domains, reading and math, that are expected to develop in a commensurate manner. Identification methods are accepted as requiring concurrence from multiple sources, specifically teacher opinion and a universal screener shown by this study to have adequate psychometric properties. Performance on state mandated, legally enfranchised testing serves as the standard of determination. Finally, as outlined by theory rather than practice, the highest attaining 15% of the participant population is deemed as the target population of interest.

Curriculum-Based Measures (CBM)

As the name implies, *curriculum-based measurements* are developed within the context of a school's curriculum; that is, the measures derive from and provide formative assessments of the students' attainment of the learning goals of a specific curriculum in use by a given school. In advancing the idea of CBM, Deno wanted "to create a simple, reliable, and valid set of measurement procedures that teachers could use to frequently and repeatedly measure the growth of their students in the basic skills of reading spelling and written expression" (Deno, 1985). He sought to design a system of measurement that would be both time and cost efficient, as the probes must be administered frequently and repeatedly. Repeated administration also necessitates that multiple, equivalent forms of

the probes must be developed to eliminate the potentially adverse effects of practice associated with multiple administrations of an instrument. Additionally, the protocols for utilizing and administrating the CBM must be easily learned by teachers and students, and, as CBMs are administered within the context of ongoing instruction, the tasks must be of short duration (Deno, 1985). Probes are written to reflect end-of-year achievement goals, but are administered throughout the academic year.

Administration of each series begins with a universal screener (US) designed to introduce testing protocols, establish baseline performance, and identify levels of student competency. This outcome measure may be used to screen for at-risk students, who are then monitored more frequently in an RTI setting. Subsequent probes in the series serve as CBM that provide data intended for use in graphing student progress as detailed by NCLB, a process known as *progress monitoring*. Individual student performance is plotted as *performance X time* to visualize student progress. Steady student progress will create a calculable slope and establish a rate of progress. Progress monitoring using CBM is a quantifiable way to assess the efficacy of RTI interventions used to improve student attainment.

Once data have been collected, a decision must be made about levels of performance; that is, the determination of measurement points whereby acceptable performance is separated from unacceptable performance or into multiple levels. *Cutoff* points are at the termini of divisi in a range of scores which are comprised by a scale. Scientifically valid instrumentation will possess a range of scores that are described as

"normal" or "average," and scores above or below that range may begin to be considered atypical and warrant investigation. Screening and progress monitoring tools need established cut points to guide tier placement decisions such as whether the student demonstrates an adequate response, if changes are needed in instruction, or if a change of tier placement is appropriate (NCRTI, 2007).

CBM can be used as a means of *screening* (Ardoin et al., 2004), *identifying*, or *monitoring* (Fuchs & Fuchs, 1999; Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993; Hosp & Hosp, 2003; Stecker & Fuchs, 2000), and may be used to confirm or disconfirm students' status within RTI tiers. As an indication of the efficacy of instructional methods, data may also be used to direct decisions concerning teaching methods (NCRTI, 2007). Formative assessments using CBM are now essential components of RTI models (Burns, Dean, & Klar, 2004; Burns & Ysseldyke, 2005; Gresham, 2002). The research literature supports the utility of this type of measure (Burns, Jacob, & Wagner, 2008; Fuchs & Fuchs, 1986; Shinn, 2002), and the data have been demonstrated to be reliable for many populations *if* properly implemented (Burns, Jacob, & Wagner, 2008). Margolis (2012) stresses the need for assurance in the reliability and validity of CBM instrumentation. Improper use of CBM, or the use of invalid CBM can have negative ramifications in identifying and serving at-risk students. Researchers claim that CBM represents a scientifically validated form of progress monitoring (Fuchs, Seethaler, Fuchs, & Hamlett; 2008). However, progress monitoring should be distinguished from the special education referral process. Burns, Jacob, and Wagner (2008) question the use

RTI data when making formal eligibility decisions, and claim that this practice represents a yet-to-be validated use of these data. To these authors, it is necessary to evaluate the ethical and legal standards of acceptable assessment practices involving the use of RTI in special education decision making.

This study uses instruments designed as a series of grade-level probes based on curricular content that may be used in either whole- or small-group settings such as that used in RTI tiers. It is part of the process of establishing various forms of validity for these instruments as CBM, and its efficacy when used as a gifted screening tool. This study includes an examination of the relation between early measures of reading and math with data derived from instruments designed as CBM.

Universal Screening (US)

The National Center for Response to Intervention (NCRTI; 2007) defines *universal screening* as a process using "… brief assessments that are valid, reliable, and demonstrate diagnostic accuracy for predicting which students will develop learning or behavioral problems." An LEA must administer nationally normed, skills-based universal screeners as a brief screening assessment administered to *all* students to determine whether students demonstrate the skills necessary to achieve grade-level standards. Universal screening reveals which students are performing below, at, or above the level considered necessary for achieving long-term success (general outcome measures). The LEA should ensure that the screener administered is actually the most appropriate universal screener for the function it serves (NCRTI, 2007).

Screening of academic skills includes domains such as basic reading skills, reading fluency, reading comprehension, math calculation, math problem solving, and written expression. Screening is conducted with all students at the beginning of the school year in grades K-8 to identify those who are at risk of academic failure; however, some schools and districts administer a screener two or even three times throughout the school year. Jenkins et al. (2013) note in a survey of 62 elementary schools from 17 states that schools conformed closely to recommendations of Gersten et al., (2009) and the NCRTI (2010) that screening and benchmarking occur at least twice annually, finding that 98% of respondents indicated benchmark measures were given triennially with 90% using a form of CBM. NCTRI (2007) recommends implementing a two-stage screening process, by first using a universal screener to identify students, followed by additional, more in-depth testing (or short-term progress monitoring, see below) for students scoring below or above a pre-determined cut point.

An examination of universal screeners as a distinct type of CBM used for screening above-grade level performance becomes important. Evaluation of instrumentation deemed valid for the screening of below-grade level students for applications to above-grade level performance should be conducted with thoughtful intention to determine the applicability and validity of the instrument. The present study extends other research on the MIR:R and MIR:M universal screeners from the identification of struggling students to the identification of gifted and high-ability students.

Concurrent Development of Reading and Math

This study uses domain-specific data collected from the target population in reading and math. It is anticipated that student performance in reading will be strongly correlated to math performance based on the assumption that these skills develop synchronously. Difficulties in math have often been shown to be comorbid with reading difficulties (e.g., Ackerman & Dykman, 1995; Landerl & Moll, 2010; Räsänen & Ahonen, 1995), serving to promote the question as to whether these skills, which are often viewed as discrete, may have similar etiologies. If so, the same cognitive factors could mediate both academic skills.

Development of Reading Skills

Rapid serial naming or *rapid automatized naming* (RAN) is the ability to name as rapidly as possible highly familiar symbols such as digits, letters, colors, and objects. RAN has been shown to be a robust predictor of reading acquisition and future fluency (Georgiou et al., 2012; see also Compton, 2003; de Jong & van der Leij, 1999; Landerl & Wimmer, 2008; Savage & Frederickson, 2005). Reading acquisition is assumed to occur in three phases (e.g., Duncan & Seymour, 2000; Ehri & McCormick, 1998; Seymour , Aro, & Erskine, 2003). Earliest is the *alphabetic phase*, a period in which learners connect letters to sounds (sequential decoding). Following is the *orthographic phase*, when emergent readers are able to consolidate graphemes and phonemes into larger units (blending). Finally, reading becomes *fluent* when a child demonstrates the rapid ability to retrieve larger units and attach meaning (morphological skills).

As early as 1915, researchers suggested that set standards at different grade levels needed to be met for children to become successful readers (Starch, 1915). Outlined then was a series of sequential stages for successful reading development comprised by three components of reading: accurate word recognition and decoding, speed, and comprehension. LaBerge and Samuels (1974) argued that comprehension was difficult if children did not learn to rapidly recognize words. The mental preoccupation with the decoding process results in a failure to construct meaning. Practice in word identification allows readers to automatize the decoding process increasing the ability to focus on the construction of meaning. A fluent reader can simultaneously and efficiently process the two tasks of decoding and comprehension (Wang, Algozzine, & Porfeli, 2011). Oral reading fluency is widely accepted as the path to comprehension and overall success in reading (National Research Council, 1998; HHS, 2000a, 2000b); however, the relation between reading rate and reading comprehension needs additional research to be more fully understood.

Development of Math Skills

Similar to reading, development in math has been shown to occur in three stages. In Level I the features of basic numerical skills develop in which number words and sequences are isolated from quantities. Some researchers believe that infants are born with the capacity to discriminate quantities, the implication being that through early experience they can differentiate between discrete quantities (see Antell & Keating, 1983; Bijeljac-Babic, Bertoncini, & Mehler, 1993; Huntley-Fenner & Cannon, 2000; Xu,

Spelke, & Goddard, 2005). Others believe that this process is a differentiation between the spatial extent of the quantities, rather than between discrete amounts (see Feigenson, Carey, & Spelke, 2002; Rousselle, Palmers, & Noël, 2004; Xu et al., 2005). With language acquisition children develop the ability to verbally discriminate between quantities; using words such as *more*, *less*, and the *same amount* when comparing quantities (known as *protoquantitative comparison schema*; Resnick, 1989,). Learning to count, a sequential recitation of number words at around two years attaches precise number words to an exact number word sequence (Krajewski & Schneider, 2009). However, these number words may not be actually used to describe quantities; that is, the number words remain isolated from quantities.

In Level II the ability to link number words with a specific quantity emerges, and children are able to attach meaning to number words. This facilitates, for example, the arranging of numbers according to their size (see Gersten et al., 2005; Okamoto & Case, 1996). Level II skills typically are acquired in two phases. First is the development of an imprecise, vague conception of the correspondence between number words and quantities, and the assignation of number words to approximate quantity categories. Known as Level IIa, this develops at around three years of age. This process is refined during Level IIb as the ability to distinguish between adjacent numbers gradually develops, and number words become linked to exact quantities (Gersten et al., 2005; Okamoto & Case, 1996). At this time children are able to judge between quantities without reference to number words. Around four to five years of age, experience

promotes an understanding that a quantity can be divided into pieces which, when reassembled, will remain equal to the original quantity (known as *protoquantitative part/whole schema*; Resnick, 1989), and that quantities can only change when something is added or taken away (known as *protoquantitative increase/decrease schema*; Resnick, 1989).

Level III of successful math development is distinguished by a linking of quantity with the concept of number relations. Children begin to understand that precise number words can be used to represent part–whole relations (decomposition of numbers), that the difference between two numerical quantities will yield a third numerical quantity (differences between numbers), and that the difference between two numbers is another number.

Counting ability (the ability to count number words forward, backward, and in steps) has been identified as a strong predictor of later calculation fluency (Koponen, Aunola, et al., 2007; Krajewski & Schneider, 2009). Learning to *calculate* is a gradual developmental process with children first utilizing a counting-based calculation strategy such as verbal or finger counting (Ostad, 1999; Siegler, 1987; Siegler & Shrager, 1984). Frequent and repeatedly successful use of counting strategies has been hypothesized to increase representations for calculation facts in long-term memory (LTM). These, in turn, lead to the development of strategies to retrieve facts from LTM, becoming the basis for future fluency (Barrouillet & Fayol, 1998; Siegler & Shrager, 1984). Adequate representation in LTM facilitates an important developmental shift from a slower,

counting-based strategy to the faster calculation strategy of *fact retrieval*. Typically, children start to use fact retrieval as the primary calculation skill by the age of nine years (e.g., De Brauwer, Verguts, & Fias, 2006; Lemair & Siegler, 1995).

Mathematical reasoning is a critical skill that requires the use of all other mathematical skills, such as how to evaluate situations, select problem-solving strategies, draw logical conclusions, develop and describe solutions, and recognize how those solutions can be applied. Reasoning is usually divided into component parts; *inductive reasoning* which involves looking for patterns and making generalizations, and *deductive reasoning* involving processes related to forming logical arguments, drawing conclusions, and applying generalizations to specific contexts (Steen, 1999).

Metacognitive elements are also important to developing reasoning skills facilitating the recognition that mathematics makes sense and can be understood. Reasoning is built on domain-specific numerical skills and knowledge; however, researchers have shown the primacy of executive functioning cognitive factors in developing reasoning skills. Executive functioning skills are those required to monitor and control thought and action. Particularly important are *working memory*- the domain-general ability of holding and manipulating information in mind (Raghubar, Barnes & Hecht, 2010); *inhibition*- the ability to suppress distracting information and unwanted responses (Bull & Scerif, 2001; Gilmore et al., 2013), and *shifting*- the ability to flexibly switch attention between different tasks (Yeniad et al., 2013). Successful learners are mathematically active through the use of tasks such as discussion, projects, and teamwork

(Anderson, Reder, & Lebiere, 1996); passive strategies such as rote memorization and drill are less likely to produce either lasting skills or deep understanding. Reflective or "metacognitive" activities are also associated with increased success (Resnick, 1987). Bjork and Druckman (1994) showed that real competence comes only with extensive practice, and that students who utilize both recalled and deduced mathematical facts make more progress than those who limit themselves to one or the other (Askew & Williams, 1995).

Correlation of Math and Reading Skills

Krajewski and Schneider (2009) report findings to support the assumption that phonological awareness is a *domain-general* precursor variable of school achievement rather than a *domain-specific* precursor variable related to only literacy development, showing a close relation between the development of literacy and math from the early preschool years onward. Further, they provide evidence that there are also synchronous associations in the development of math and both intelligence and working memory skills. This developmental correlation is strengthened by the clearly established link between the presence of problems or delays in language, especially specific language impairments (SLI), and problems in the acquisition of early numeracy skills (Arvedson, 2002; Fazio, 1994). The common neurological base for phonological awareness and numeric representations in the brain (Dehaene et al., 2003) supports the assumption that limited access to the phonological formats of *counting words* (e.g. “one,” two,” “three,”

etc.) contributes to an inability to manipulate verbal codes required for counting (Geary, 1993; Simmons & Singleton, 2008).

Though less is known about the association between counting and reading, previous researchers' findings suggest that Kindergarten measures of counting ability are *more* strongly associated with reading performance than traditional, linguistic predictors of reading skills (Leppänen, 2006). Koponen, Aunola, Ahonen, and Nurmi (2007) found that counting ability predicted the co-variation of *both* single-digit calculation fluency and reading fluency. Though this particular research was conducted in Danish, other researchers have found these tendencies to be true in other languages as well. Research has also correlated RAN and calculation fluency at levels of significance (Bull & Johnston, 1997; Hecht, Torgesen, Wagner, & Rashotte, 2001; Koponen, et al., 2007; Swanson & Kim, 2007). This indicates that the skill acquisition process is similar in reading and math despite the differences in language representations in the brain. At present, reading and math abilities seem etiologically similar.

Researchers have repeatedly found substantial inter-correlations between literacy and math competencies reporting coefficients ranging between $r = .40$ and $r = .60$ (e.g., Berg, 2008; Koponen et al., 2007; Schneider, 2009), indicating that similar cognitive competencies influence performance and development in these two disparate areas of school achievement. Deficits in the relevant precursor variables (phonemic awareness and RAN) have been have been shown by researchers to be related to problems in both literacy and math development (e.g., Geary, 1993). Both the impact of working memory

skills and the role of phonological awareness have been emphasized for subsequent mathematical achievement (Krajewski & Schneider, 2009).

Implied is a process in which, during the early learning phases of both calculation and reading skill acquisition is based on an early one-to-one coding in memory representations that then gradually shifts toward the processing and retrieving of ever larger units; syllables or words in reading, and facts in calculation (Koponen, Salmi, Eklund, & Aro, 2013). It becomes logical then to expect reading and math skills to develop apace, with measures of student performance in the separate domains of reading and math to be correlated. Highly discrepant performance between reading and math could indicate a specific learning disability in either of the domains or single-subject giftedness. Discrepant performance could also be an indicator of dual exceptionalities (or twice-exceptional students, 2e; also known as gifted with learning disabilities [G/LD], etc.); those students who are gifted but who also have concomitant disabilities, and who often require, or would benefit from, adaptations, accommodations, and/or curricular modifications (Barton & Starnes, 1989; Baum, 1991, 2004; Cline & Schwartz, 1999; National Association for Gifted Children; 1998). The associated disabilities may include any of the disabilities commonly recognized in special education.

This research includes an examination of the correlation between measures of math and reading on the three instruments involved (Teacher Rank, MIR:R and MIR:M, and TCAP). Developmental theories indicate that valid measures should have high

correlations when administered to the same population. Significant correlations yield support for the construct validity of the instruments included.

Chapter 3

Participants, Instrumentation, Methods

Statement of Purpose

Comparisons of state rates of identification of giftedness, controversy over the definition of giftedness and the role of schools in identifying and serving gifted students, a lack of consensus concerning the extension of the response to intervention model to serve gifted students, and the questions centered on the validity of CBM as gifted screening instruments indicate that policy and practice in gifted education may lack clarity when attempting to determine best practice. This study expands the literature of the gifted education field by examining the validity of CBM when used as screening instruments, the accuracy of teacher perception, and the adequacy of measures taken early in the school year as gifted screening instruments when necessitated by the absence of formal measures. The purpose of this study is to examine the relation between two screening measures taken early in the school year and an end-of-year high-stakes test to assess the utility of early measures for making educational placement decisions for gifted students. The ability of CBM to identify gifted students in reading and math (i.e., the adequacy of ceiling and item gradient) is examined. The relation between student performance on two early measures of reading and math, a qualitative domain-specific teacher rank and a quantitative progress monitoring used as a universal screener, is examined. These are compared to a quantitative norm-referenced measure taken at the year's end. The inter-correlation between the domain-specific measures as compared to

each other is examined (i.e., the relation of reading or math when comparing the instruments to each other), as well as the intra-correlation between the domain-specific metrics (i.e., the relation of reading and math when comparing measures taken from the same instrument).

Participants

Consent and Approval

The initial research was conducted with Institutional Review Board approval (IRB), and typical IRB protocols were followed during the research and subsequent analyses. Multiple layers of consent were obtained from all participants (district, administrative, parental, etc.). Student and teacher participant confidentiality subsequently has been strictly maintained throughout the research process by the assignation of student, teacher, and school identification numbers, and vigilance in document security of both electronic and paper copies of the testing instruments.

Participant Demographics

Participants were third-grade students (n = 556) and their teachers (n = 28) enrolled in eight elementary schools during the 2010-2011 academic school year from a small, rural school district located in the southeast of the United States. Incomplete cases, that is those without measures from all three instruments, were deleted from the set. Incomplete cases represent absenteeism at the time of testing administration of MIR:R and MIR:M. Incomplete cases may also be attributable to ingress or egress from the district during the time between the collection of early and late measures. Those

responsible for data entry, erring on the side of caution, did not code responses of unresolvable confusion, such as confusion about different participant names between instruments (e.g. a nickname on a teacher raking form lacking a clear association with a birth name on the TCAP). Such cases of missing data were also removed from the data set. Finally, special education students specifically included in the teacher ranking may have been disaggregated from TCAP class reports if tested under alternative conditions, resulting in missing data and deletion from the data set. If occurring, this is not perceived as an overt threat to the study as the population of interest (high performing students) is not likely to have been provided with testing accommodations; or, if so, these students are not likely to exist in numbers significant enough to affect outcomes. Another 31 cases were deleted under advice from the authors of the MIR:R (see below). The resulting dataset contains 372 student cases; 191 (51.3%) are *female*, 181 (48.7%) *male*. No demographic data about the teachers were collected.

Participant ethnicity was consistent with the ethnic diversity within the district, which was predominantly White. To facilitate analysis, ethnic categories were collapsed resulting in two categories; *White* (*n* = 346; 93%) and *non-White* (*n* = 26; 7%) comprising African American, Hispanic, Asian, and Native American. The majority of the non-White population was located in a single school within the district that reports a total of 22.6% of its population as non-white; this percentage is atypical for the district as a whole whose combined non-White populations were less than 6% of students (See Table 3.1). The state reports that 52% of the target district was economically disadvantaged, with a

total of 57.7% of students qualifying for free (50.3%) and reduced fee (7.4%) lunch status. However, individually, the individual school's percentage of free and reduced lunch was a rather wide range from 45.1% to 79.4%.

Table 4. Participant School Demographics

ID	GS	TS	W	AA	H	A	NA	M	F	FRL
1	K-5	193	97.9	1.6	0.5	-	-	56.5	43.5	59.4
2	K-5	359	93.6	3.6	2.2	0.3	0.3	51.5	48.5	73.3
3	K-5	746	94.8	3.5	0.7	0.7	0.4	55.8	44.2	45.1
4	K-5	349	95.4	3.4	0.3	0.9	-	50.4	49.6	54.0
5	PK-5	433	97.7	0.2	1.6	0.2	0.2	52.9	47.1	65.9
6	PK-8	638	98.1	1.4	-	0.5	-	50.6	49.4	53.9
7	K-5	712	95.4	3.1	0.8	0.3	0.4	52.4	47.6	72.2
8	K-5	296	77.4	19.6	2	1	-	55.4	44.6	79.4

Notes: ID-School ID#; GS- Grades served; TS-Total Students; W- White (%); AA- African American (%); H- Hispanic (%); A- Asian (%); NA- Native American (%); M- Male (%); F- Female (%); FRL- Free reduced lunch (%). Source: http://www.tn.gov/education/research

Participant Achievement

On state-mandated TCAP testing to assess annual yearly progress (AYP), third-grade students within the target district performed in a manner consistent with that of the state as a whole. It should be restated that teacher rank and universal screening data were collected in September, 2010, and TCAP data were collected in March of 2011. Of the four TCAP achievement levels designated by the state (*below basic*, *basic*, *proficient*, and

advanced), the target district generally had slightly more students performing at a *basic* level in reading and math when compared to the state performance, but fewer at the *below basic* level. State-wide, more than half the students failed to make AYP in reading (56.1%) and math (48.6%) when the two lower achievement levels, basic and below basic, are summed. Rates for students in the target district were higher, with inadequate progress made by 57.6% of students in reading and 54.1% in math. The percentage of students in the target district who were passing in reading (42.4%), the sum of the proficient and advanced level percentages, was only negligibly lower than the state average (43.9%). In math performance, however, the percentage of students in the target district who were passing math (45.9%) was much lower than that of the state as a whole (51.4%).

Table 5. Comparison of Target District (TD) and Tennessee Comprehensive Assessment Program Performance Reading and Math (%)

TD/STATE	BB	B	P	A	P/A
TD Reading	8.6	49.0	33.9	8.4	42.4
State Reading	13.4	42.7	33.0	10.9	43.9
TD Math	8.8	45.3	36.5	9.4	45.9
State Math	9.2	39.4	38.1	13.3	51.4

Notes: BB- Below Basic; B- Basic; P- Proficient; A- Advanced Source: http://tn.gov/education/data/tcap_2011.shtml

Within-district student performance in reading and math achievement were commensurate; TCAP reading achievement and math achievement were very similar at

all achievement levels. Thus, individually discrepant performance between reading and math achievement is notably important. Also noteworthy is the percentage of students who attain at the highest level in the target school district relative to the state average percentage. The target district had 8.4% at the advanced level in reading compared to the state rate of 10.9%; while in math performance the percentage of students who attained the highest level is only 9.4% compared to the state rate of 13.3%.

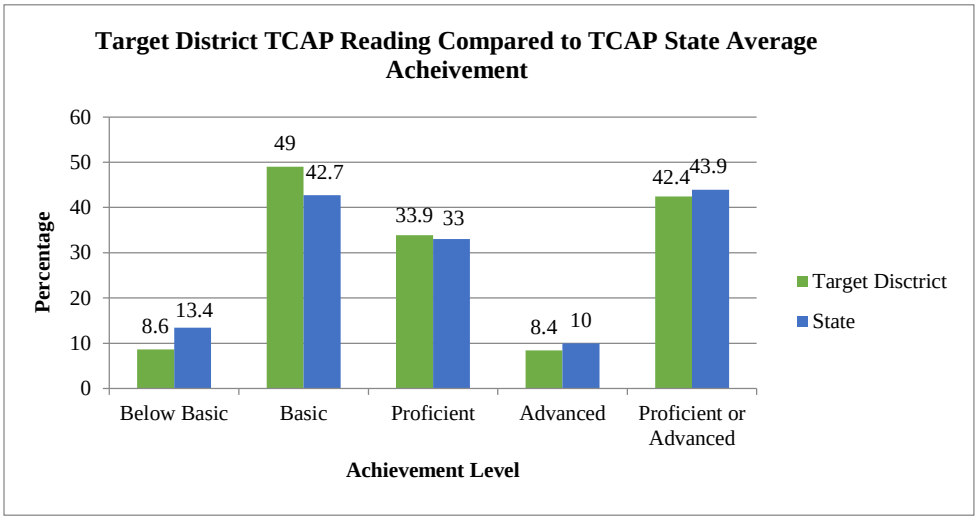

Figure 1. Comparison of Target District and Tennessee Comprehensive Assessment Program Performance Reading Achievement Levels

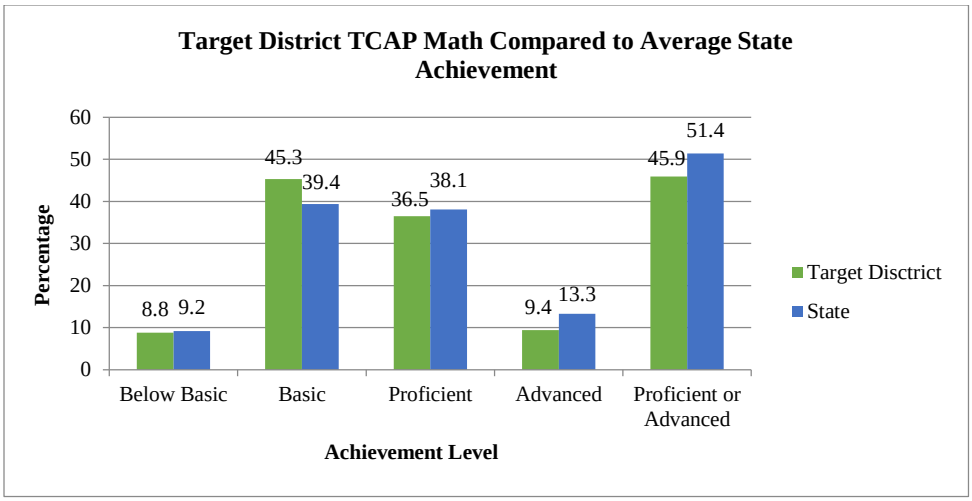

Figure 2. Comparison of Target District and Tennessee Comprehensive Assessment Program Performance Math Achievement Levels

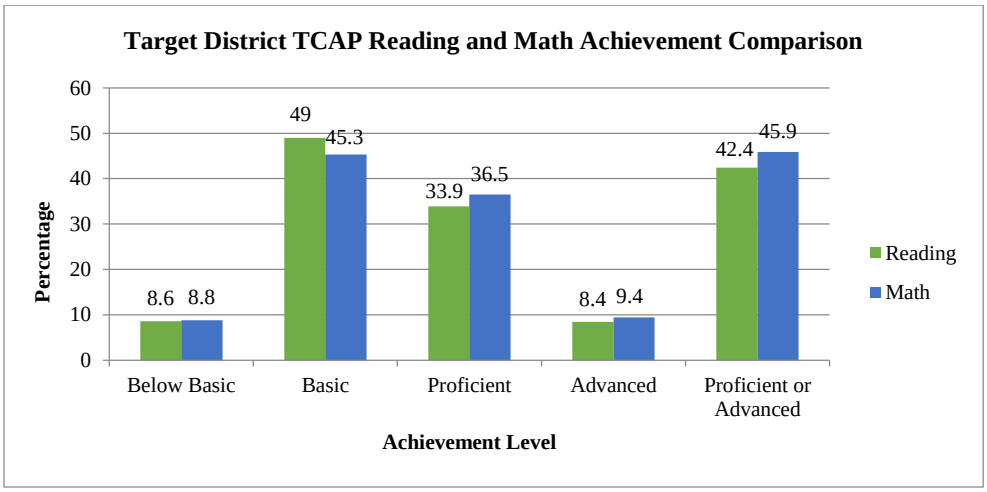

Figure 3. Comparison of Target District Tennessee Comprehensive Assessment Program Performance Reading and Math Achievement Levels

Instrumentation

The analysis proceeds from data collected through three different instruments, two collected early in the academic year, and one at year's end.

Early Measures

Monitoring Instructional Responsiveness: Reading and Math

The *Monitoring Instructional Response: Reading* (MIR:R) and *Math* (MIR:M) were developed by Bell & McCallum and colleagues starting in 2010; the present research was conducted as a part of the ongoing development of this instrumentation. Monitoring Instructional Responsiveness: Reading (MIR:R; Bell, Hilton-Prillhart, McCallum, Hopkins, 2012) assessment probes and the Monitoring Instructional Responsiveness: Math (MIR:M; Bell, Hilton-Prillhart, McCallum, Hopkins, 2012) assessment probes are experimental and intended as CBM of reading and math skills for grades K-5 and within the natural classroom setting. A feature of both is the inclusion of three universal screeners. Although the probes can detail specific skills within reading (rate and comprehension) and math (calculation and problem solving), only composite scores were utilized in the following analyses.

MIR:R and MIR:M have at present no explicit ability to assess students at above-grade level or for gifted attainment. Both were originally conceived as CBM and progress monitoring tools for use within general education classroom and RTI settings, where they have previously been shown to have some degree of utility when identifying at-risk students; though, after probe development for general education classrooms, the authors

hope applications for gifted students would be explored. This research extends previous research to examine the utility of MIR instrumentation as gifted screening and identification tools. No claim is made that any students were identified or received services as a result of MIR testing. Both instruments are described as indicated below and samples are provided (attached).

Monitoring Instructional Responsiveness: Reading (MIR:R; Bell, Hilton-Prillhart, McCallum, & Hopkins, 2012)

The MIR:R is a series of four universal screeners and 18 alternate forms at each grade level for grades 1-5 that was developed to assess reading *comprehension* and *fluency* and is a group administered, ecologically sound, and efficient measure of both. Researchers developed the probes in collaboration with district teachers and other specialists such as literacy leaders, special education personnel, curriculum specialists, and school psychologists from the school district. Probe sources included *Dolch Word Lists*, grade-level curriculum vocabulary lists provided by the district, word lists from the *Qualitative Reading Inventory*-IV (QRI-IV; Leslie & Caldwell, 2006), and state-approved basal texts in order to ensure content validity for each grade level. The probes contain a combination of narrative and informational passages based on the Tennessee Learning Standards for science and social studies. MIR:R was developed using the *Spache* reading difficulty formulas and was piloted and refined in classroom settings through a lengthy process.

Following standardized testing protocols, classroom teachers conducted the administration using scripted directions. Teachers were trained in administration protocols by literacy leaders trained by the researchers. A test-administration checklist to ensure procedural fidelity was included for teachers. This checklist was also used by literacy leaders to conduct fidelity checks to ensure proper test-administration procedures. Prior to the probe administration students received test-taking instructions with opportunities for guided and independent practice; time was also provided for responding to student questions.

In the first grade, students read strings of words within connected text. MIR:R probes for students at the second through fifth grade levels provide students with short reading passages of coherent and meaningful paragraphs. Test text is formatted using all lower case letters and without ending sentence punctuation. The examinee is asked to make vertical slashes between *words* (first grade) or *complete ideas* (second through fifth grades). As with all CBM the difficulty level of the passages reflects end-of-grade performance standard, with a consistent level of difficulty maintained from passage to passage. By targeting end of grade performance, test data inform development of student progress throughout the year (Fuchs et al., 1988; Shinn, 1989; Shinn 1998). All students, regardless of ability, began with the first section of the probe. Entry levels are not a feature of the test. The testing set contains three universal screeners that are intended to be administered periodically throughout the academic year, as well as progress

monitoring probes. This study uses data collected from the September administration of the first, and earliest, screener.

The MIR:R was designed to provide an assessment of both reading *fluency* and reading *comprehension* in a single administration. Group administration requires three minutes, once students are familiar with testing protocols. The three-minute timed administration of the probes allows the calculation of the number of words read correctly per minute, a fluency measure, by dividing the Total Words Read score by three. A comprehension percentage score is derived by dividing the number of ideas a student correctly identified by the number of ideas a student attempted to identify, multiplied by 100. The comprehension percentage score can also be divided by three to indicate the number of ideas identified correctly per minute. These two scores (*Total Words Read* and *Comprehension Percentage*) can be multiplied to create a *Reading Total* score, a composite of fluency (number of words read silently) and comprehension (number of ideas correctly identified) within a three-minute time period. The Reading Total score was used in this study.

Correlation data (e.g., alternate-form reliabilities) have been reported to help establish the psychometric integrity of the MIR:R. Adjacent probe correlation coefficients were calculated; the average reliability was found to be high (.80, $p < .001$). In addition, concurrent validity estimates between MIR:R and Aimsweb© Maze range from .43-.55; the concurrent validity estimate between MIR:R and the STAR Reading Assessment (Renaissance Learning Systems, 1997) is .67 (Hilton-Prillhart, 2011). Hilton-

Prillhart (2011) compared the predictive utility of MIR:R and Aimsweb© Maze scores to estimate end-of-year STAR scores, and, using a step-wise multiple regression, found that MIR:R scores predicted 37% of the variance in the STAR scores and was the most powerful predictor; AIMSweb© scores failed to produce additional predictive variance. Test-retest reliability for the MIR:R indicated a high degree of stability for grades 1-3 (first grade=.90, second-grade=.84, third-grade=.89). These correlations for grades 1-3 meet or exceed the .80 standard established by some experts for psychometric testing (Sattler, 2008).

Miller, Bell, and McCallum (in press) found a zero-order correlation coefficient of .58 ($p < .01$) between the MIR:R Comprehension Rate score and TCAP performance. This moderately strong correlation provides evidence that the MIR:R predicts high-stakes, end-of-year scores reasonably well, and its predictive power is comparable to most other CBM-type measures (e.g., DIBELS Next, AIMSweb, and independently-created measures) in the literature (Crawford et al., 2001; Reschly et al., 2009; Shapiro et al., 2006; Silberglitt & Hintze, 2005).

The mean MIR:R Reading Total score for this sample is 108.68 (*SD* = 63.11). When the sample is evaluated for skewness and kurtosis, scores approximate a normal distribution (skewness = .751; (*SE* = .126); kurtosis = .020 (*SE* = .252). The cutoff score closest to the recommended 85th percentile for gifted screening is 179 (85.8%); there are 58 cases at or above the cutoff score, a G/HA screening rate of 15.32%. For reference, perfect distribution skew = 0 and kurtosis = 3.When the scores of the 58 students scoring

at or above the 85th percentile were examined for skewness and kurtosis, skewness = .73 (*SE*.31); kurtosis = -.53 (*SE* .62). Because this sample was selected based on atypical (i.e., high performance), a non-normal distribution was expected.

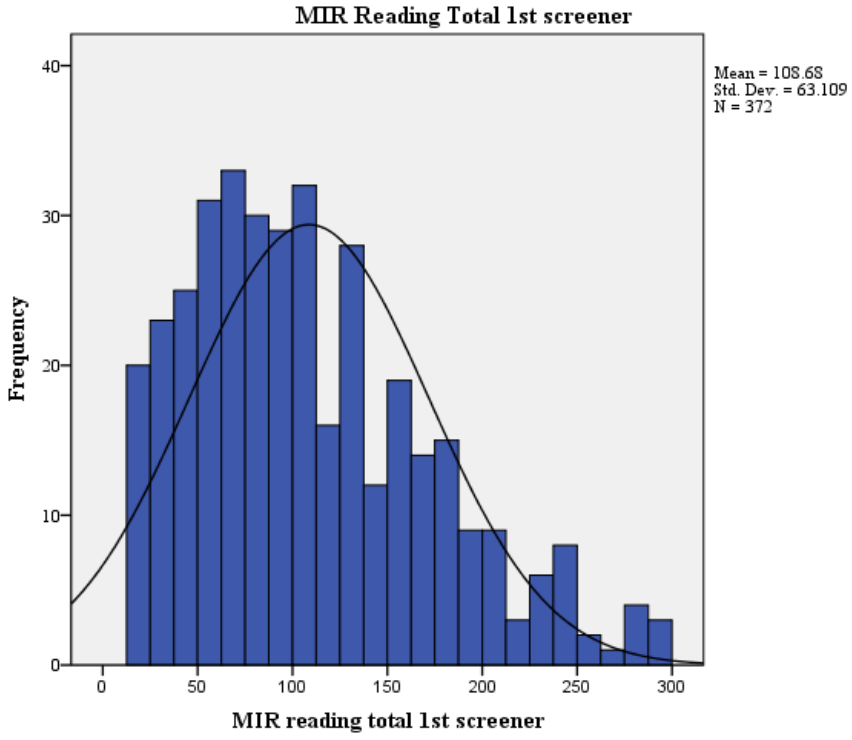

Figure 4. Distribution of Monitoring Instructional Responsiveness: Reading Scores

Researchers have established validity and reliability measures of the MIR:R throughout the ongoing development process. Validity and reliability of the MIR:R as a

screener for identifying struggling readers is promising. However, the test authors have determined that the *total words read* score can be confounded by low *comprehension* scores. As a result, cases with Comprehension scores contributing to Total Reading scores less than or equal to a raw score of 20 were dropped from the data set as invalid measures (31 cases).

Monitoring Instructional Responsiveness: Math (MIR:M; Bell, Hilton-Prillhart, McCallum, Hopkins, 2012)

The MIR:M Universal Screeners and Monitoring Probes is a brief, psychometrically strong multi-faceted set of probes designed to assess the math performance of elementary students. The probes are administered in a group setting and require three minutes once students understand testing protocols. As a measure of math fluency and item problem solving, the authors' intention was to create a non-verbal math assessment; that is, an instrument that is not confounded by an indirect measure of a student's reading ability. The probes were developed by test authors and collaborating school system personnel including literacy leaders, mathematics consultants, principals, and teachers using the Tennessee State Curriculum Standards (Tennessee Department of Education, 2009) and the Saxon Math Curriculum (Larson, 2004), National Council of Teachers of Mathematics Curriculum Standards (National Council of Teachers of Mathematics, 2000). These sources were analyzed to determine the item type and item difficulty appropriate for each grade level. The development process included several

pilot studies and a full-scale, one-year implementation in a school district in northeast Tennessee.

The MIR:M follows standardized testing protocols including the use of scripted instructions to be read by test administrators. Students complete guided practice items and are provided with an opportunity to ask questions for clarification. A timed three-minute administration follows. Fidelity checklists are also included to assure proper administration.

The MIR: M assesses four math skill areas: *Number Sentence-Quantity Discrimination* (NSQD), *Number Pattern* (NP), *Shape Pattern* (SP), and *Computation* (COMP).

Number Sentence-Quantity Discrimination (NSQD) items combined math facts and quantity discrimination items. The NSQD task consists of horizontally-presented number sentences immediately followed by a quantity discrimination task. Vertically arranged symbols (i.e., <, > , =) separate the number sentence answer from a randomly assigned number. Examinees solve the number sentence, record the response, and then circle the symbol representing the relation between the examinee's calculated answer and the random number. *Number Patterns* (NP) items consist of five numbers, presented horizontally and ordered from least to greatest and including an omission. Items are randomly assigned according to grade-specific criteria. Examinees are required to write the missing numbers in the sequence.

Shape Patterns (SPs) are presented horizontally from left to right with one shape missing from each pattern. Each item includes four possible shape choices to the right of the sequence. Examinees are asked to circle the shape that completes the sequence.

Computation (COMP) items require examinees to solve *2 x 2* or *3 x 3* addition and subtraction items. The COMP items do not require regrouping. Values for the 2 x 2 items range from 12 to 99, with any addition by one eliminated from the items due to low difficulty and discrimination indices.

For scoring, each item has a series of boxes equal to the item's number of possible correct responses; teachers tic the boxes to reflect a correct response. Examinees receive credit for each part of an item answered correctly (e.g., digits written correctly, correct item circled). The SP items require one response, whereas the NSQD, COMP, NP items allow two or more possible responses. When scoring the NSQD response, though the calculation may have been incorrect, the quantity discrimination task may still reflect a correct relation between the two numbers; or alternatively, the calculation may be correct with an incorrect response recorded for the discrimination task. In these conditions, each item is scored. When scoring the COMP and NP, though the response may be incorrect as a whole, partial credit is given for each place value with a correct digit. The total number of correct responses is tallied to calculate a single composite score.

Hopkins (2010) established concurrent validity between the MIR:M and the *Monitoring Basic Skills Progress* (MBSP; Fuchs, Hamlett, & Fuchs, 1999) and found median correlations of .66 for grade 1, .41 for grade 2, and .52 for grade 3. Hopkins also

found that the MIR:M was more predictive of end-of-the-year tests scores measured by Star Math (Renaissance Learning, 2002) than the MBSP.

Testing items on the MIR:M and TCAP place different task demands on examinees; specifically, the TCAP items place a much higher demand on reading skills, requiring students to read as many as five sentences. MIR:M items require no reading; consequently, reading ability may be a significant and uncontrolled confound during data analyses. A recent study provided preliminary evidence that reading skills impact TCAP math scores. Using the MIR:M and the MIR:R to identify third-grade students with strong math skills but significantly weaker reading skills, Bell, Taylor, McCallum, Coles, and Hays (2015) found that students with reading weakness scored significantly lower on the math portions of the TCAP than their peers. Additionally, Bell et al. reported that MIR:R scores yield a slightly stronger correlation with TCAP math scores than MIR:M, further evidence that reading skills can be a significant predictor of the TCAP math, or at the very least, that reading skills moderate the association between the MIR:M and the TCAP.

Coles (2014) found that using the Total MIR:M score resulted in the most powerful predictive model accounting for 27% of the variance in TCAP scores for a group (n = 262) in fifth grade. Importantly the initial probe designated as the first universal screener, produced the weakest correlation to the TCAP when compared with the later administrations as reported by Coles. This is believed to be related to the novel nature of the initial administration. Although practice administrations were provided,

these may have inadequately simulated the testing administration format. Changes in administration protocols, such as increased practice, may ameliorate this unexpected finding.

The MIR:M data in this study are skewed right and may have been affected by the presence of outliers that are 18 points above the next closest value (69 and 51 respectively). This score had a corresponding z-score of 4.96; outliers are usually characterized as z-score values of ± 3.29 and higher (Tabachnick & Fidell, 2007, p. 73). The cases were retained as they were of interest to the research. However, to prevent these scores from artificially increasing the slope of the regression line during multiple regressions, these scores were changed such that they remain deviant, but less so to minimize their impact on the analyses. In a process detailed by Tabachnick and Fidell (2007, pg. 77), the raw scores were given a value one unit higher (raw score = 52) than the next most extreme score in the distribution (raw score = 51). After data cleaning described, the mean MIR:M composite score for this sample was 25.62 (*SD* = 8.36). When this sample was evaluated for skewness and kurtosis, scores approximated a normal distribution; skewness = .66 (*SE* .13); kurtosis = .53 (*SE* .25). The cutoff score closest to the recommended 85th percentile for gifted screening was 34 (85.8%); there were 59 cases at or above the cutoff score, a G/HA screening rate of 15.86%. For reference, perfect distribution skew = 0 and kurtosis = 3.When the scores of the 59 students scoring at or above the 85th percentile were examined for skewness and kurtosis,

skewness = 2.34 (*SE* .32); kurtosis = 7.53 (*SE* .61). Because this sample was selected based on atypical (i.e., high performance), a non-normal distribution was expected.

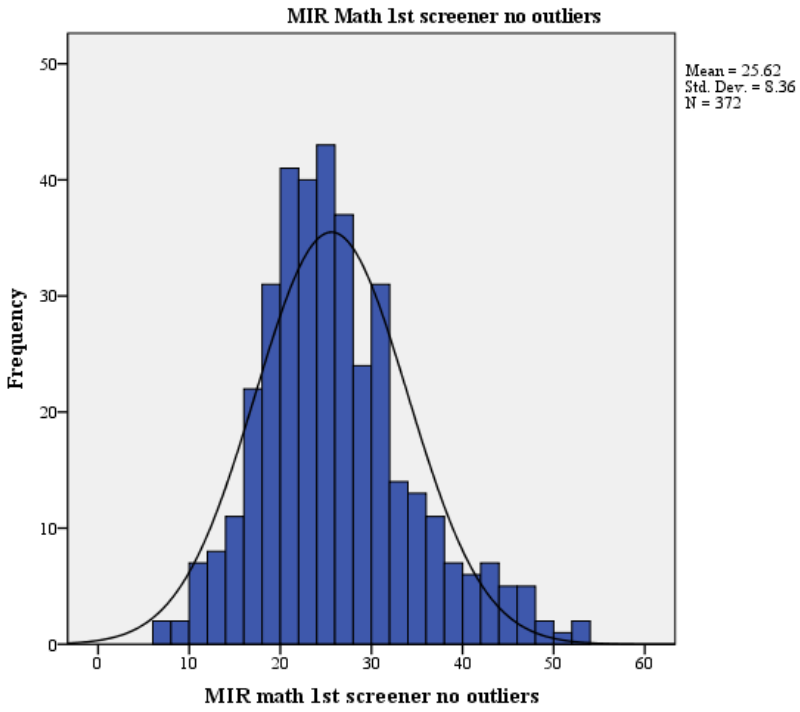

Figure 5. Distribution of Monitoring Instructional Responsiveness: Math Scores

MIR:M is a series of four universal screeners and 18 alternate forms at each grade level for grades 1-3. As with other forms of CBM, the MIR:M targets end-of-grade learning objectives. MIR:M demonstrates partially convincing evidence (as defined by the *National Center on Response to Intervention*) as a reliable and valid brief multi-

operational, curriculum-based measure of math. For more information regarding the psychometrics of MIR:M see Hopkins (2010).

Teacher Ranking of Students in Reading and Math

The *teacher ranking* (TR) forms were provided to teachers participating in the research (n = 28); all participating teachers returned the ranking information on their class. As per the instructions, each teacher was requested to rank order the students in his or her class (1 as highest) according to the students' performance. Teachers ranked students separately in reading (TR:R) and math (TR:M), generating a class rank in each domain. Teacher participants were assigned a code number to maintain confidentiality and to facilitate analyses.

The instructions on the teacher ranking form directed teachers to "base your rating on your own experiences with each student, regardless of whether he or she receives extra tiers of instruction [RTI] and/or special education." The form further directed that teachers should use their best judgment when completing the ranking of their students. Teachers were asked to consider each student's performance and achievement including daily work, assignments, class activities, projects, and tests.

Anecdotally, a few teachers included a handwritten note on their completed forms indicating that the rank order reflected a *quantitative ranking* derived from current grade books. It cannot be said, however, that this method was used by all teachers completing forms or in every case. Additionally, it should not be inferred that all teachers used the same criteria in determining rank; teachers value differing aspects of student profiles

when ordering students. Consequently, this instrument must be considered as a *qualitative ranking* and the more subjective metric of the data set. As such, it becomes an overall measure of teacher perception of a range of student traits including both academic performance and other affective considerations as may be deemed important by the teacher (e.g., motivation, personality, reliability, neatness, etc.). Significantly, the TR was taken as an early measure to assess teacher perception of student ability. Its accuracy will be assessed through comparisons with other early and later measures. Finally, with the use of the TR, no claim is made concerning teacher nomination of top-ranked students to gifted programming or, indeed, about student giftedness. Form instructions made no mention of gifted-level performance. It is important to explicitly state that there is no assumption that any students were screened and/or identified as gifted, or received any interventions based upon gifted identification. Equally, no conclusions should be made concerning gifted programming provided by the district. The analyses proceed from the logic that a teacher, if inclined to nominate students to gifted programs at all, would be *more likely* to nominate to gifted programs the students they assigned to the top ranks. Thus, gifted screening cutoff scores will be limited to the top two assigned ranks in each domain. This method identifies 66 students in each domain.

End-of-Year Measure

Tennessee Comprehensive Assessment Program (TCAP; 2011)

Tennessee Comprehensive Assessment Program (TCAP) is a state-mandated, criterion-referenced high-stakes test given in the spring each year in grades 3 through 12

(Tennessee Department of Education, http://tn.gov/education/assessment/achievement.shtml). A criterion-referenced test measures a student's performance against specific content standards or criteria, rather than comparing the performance of test takers to each other. The test is designed to assess student attainment of state learning goals and is used to document annual yearly progress (AYP). The test is divided into three sections each containing several subtests. In this study results from reading and language arts and math subtests are used, as described below. Scores are obtained in other domains by subtests in science and social studies that are not included in the present research.

On the TCAP Achievement Test, each test item is directly linked to a performance indicator. These indicators were designed by panelists using additional reference data provided by Tennessee student's performance on 4th and 8th grade NAEP and 8th grade Explore, 10th grade PLAN, and 11th grade ACT national assessments. Performance indicators are clustered into reporting categories:

- *Advanced* – Students who perform at this level demonstrate superior mastery in academic performance, thinking abilities, and application of understandings that reflect the knowledge and skill specified by the grade/course level content standards and are significantly prepared for the next level of study.
- *Proficient* – Students who perform at this level demonstrate mastery in academic performance, thinking abilities, and application of understandings that reflect the

knowledge and skill specified by the grade/course level content standards and are prepared for the next level of study.

- *Basic* – Students who perform at this level demonstrate partial mastery in academic performance, thinking abilities, and application of understandings that reflect the knowledge and skill specified by the grade/course level content standards and are minimally prepared for the next level of study.
- *Below Basic* – Students who perform at this level have not demonstrated mastery in academic performance, thinking abilities, and application of understandings that reflect the knowledge and skill specified by the grade/course level content standards and are not prepared for the next level of study.

TCAP Reading

A reading composite scale score was derived from three subtests (Critical Reading, Grammar and Spelling, and Word Usage) and was used in the present research. TCAP tests have a highest obtainable score of 900. The TCAP:R data may have been affected by the presence of one outlier that was 45 points above the next closest value (raw scores = 879 and 834 respectively). This score had a corresponding *z*-score of 4.43; outliers are usually characterized as *z*-score values of ±3.29 and higher (Tabachnick & Fidell, 2007, p. 73). The case was retained as it was of interest to the research. However, to prevent this score from artificially increasing the slope of the regression line during multiple regressions, the score was changed such that it remained deviant, but less so to minimize its impact on the analyses. In a process detailed by Tabachnik and Fidell (2007,

pg. 77), the raw score was given a value one unit higher (raw score = 835; *z*-score = 2.74) than the next most extreme score in the distribution (raw score = 834). The mean TCAP:R composite scale score for this sample was 760.62 (*SD* = 26.28) after data cleaning. When data this sample was evaluated for skewness and kurtosis, scores approximated a normal distribution; skewness = -.06 (*SE* .13); kurtosis = .52, (*SE* .25). The cutoff score closest to the recommended gifted screening 85th percentile was 784 (84.7%); there were 70 cases at or above the cutoff score, a G/HA screening rate of 18.82%. Statewide, advanced status represents the top 10.9%. For reference, perfect distribution skewness = 0 and kurtosis = 3. When the scores of the 70 students scoring at or above the 85th percentile were examined for skewness and kurtosis, skewness = 2.28 (*SE* .29); kurtosis = 8.06 (*SE* .57). Because this sample was selected based on atypical (i.e., high performance), a non-normal distribution was expected.

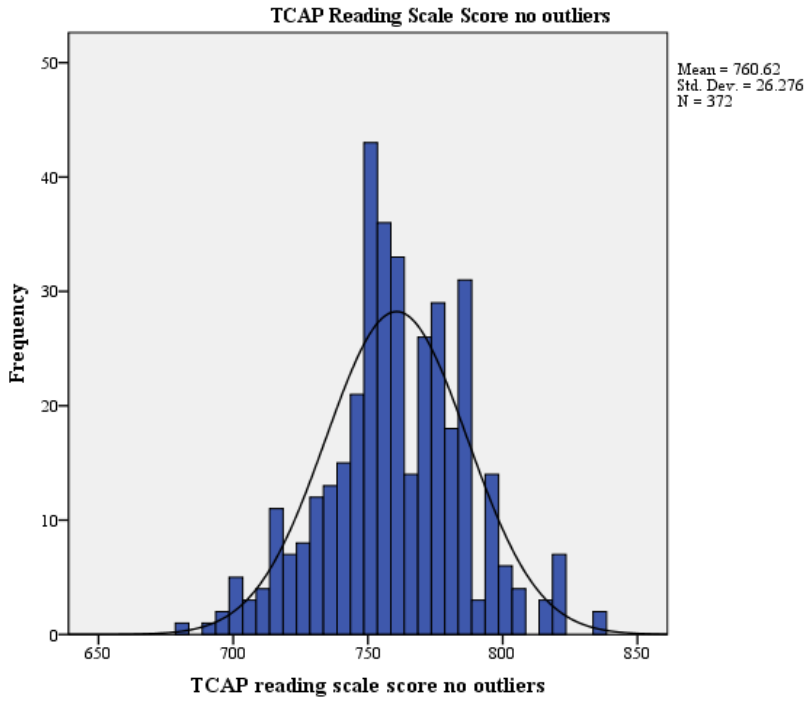

Figure 6. Distribution of Tennessee Comprehensive Assessment Program Performance Reading Scores

TCAP Math

A composite score in mathematics was derived from two subtests (Quantitative Reasoning and Calculation); this math scale score was used in the present research. The TCAP:M data may have been affected by the presence of three outliers that were 67 points above the next closest value (833 and 900 respectively); there were three perfect 900 scores. These scores had a corresponding *z*-score of 4.52; outliers are usually characterized as *z*-score values of ± 3.29 and higher (Tabachnick & Fidell, 2007, p. 73). The cases were retained as they were of interest to the research. However, to prevent

these scores from artificially increasing the slope of the regression line during multiple regressions, these scores were changed such that they remain deviant, but less so to minimize their impact on the analyses. As described above, the raw scores were given a value one unit higher (raw score = 834; *z*-score = 2.61) than the next most extreme score in the distribution (raw score = 833). Similarly, an atypical low score (raw score = 641; *z*-score = 3.75) was raised to a value one unit lower (raw score = 666) than the next extreme score (raw score = 665). The mean TCAP:M composite scale score for this sample was 758.02 (*SD* = 29.16) after data cleaning. When this sample was evaluated for skewness and kurtosis, scores approximated a normal distribution (skewness = -.02 (*SE* .13); kurtosis = .53 (*SE* .25). The cutoff score closest to the 85th percentile was 784 (86.8%); there were 64 cases at or above the cutoff score, with a G/HA screening rate of 17.20%. Statewide, advanced status represents the top 13.3%. For reference, perfect distribution skew = 0 and kurtosis = 3. When the scores of the 64 students scoring at or above the 85th percentile were examined for skewness and kurtosis, skewness = 2.57 (*SE* .30); kurtosis = 7.45 (*SE* .59). Because this sample was selected based on atypical (i.e., high performance), a non-normal distribution was expected.

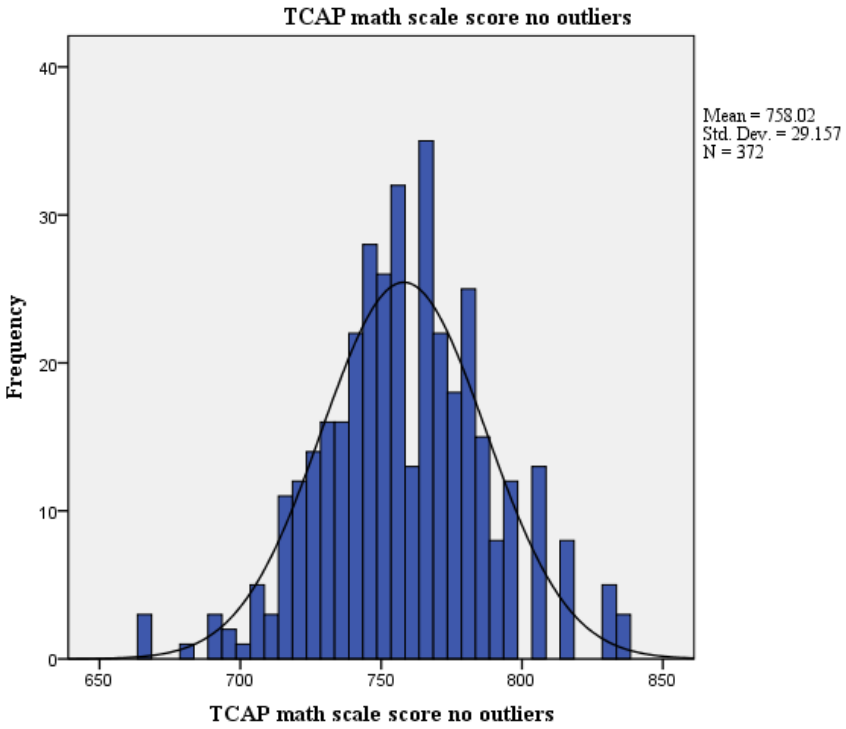

Figure 7. Distribution of Tennessee Comprehensive Assessment Program Performance Math Scores

With the inclusion of TCAP scores on the *TN K-12 Intellectually Gifted Assessment Scoring Grid* these scores are specifically intended for use as a measure of gifted performance and may be used for both screening and identification of gifted students. The matrix uses TCAP performance as a primary marker of academic performance and uses cutoff metrics of 95% and 90% in one or several content domains, respectively, as sufficient indices for identification. As seen above, the advanced cutoff

metric (TCAP:R 10.9%; TCAP:M 13.3%) aligned with other identification metrics but was well above the average Tennessee gifted identification rate of 2.02%. As a standardized, criterion-referenced measure, the TCAP was compared to both the TR and MIR measures to help provide insights concerning the accuracy of measures taken early in the year.

Table 6. Comparison of Instrumentation Descriptive Statistics

	Range	Min	Max	*M* (*SD*)	Skewness (*SE*)	Kurtosis (*SE*)
MIR:R	277	21	297	108.68 (63.11)	.75 (.13)	.02 (.25)
MIR:M	45	7	52	25.62 (8.36)	.66 (.13)	.50 (.25)
TCAP:RSS	154	681	835	760.62 (26.28)	-.06 (.13)	.25 (.25)
TCAP:MSS	168	666	834	758.02 (29.16)	.66 (.13)	2.53 (.25)

Notes: MIR:R- Monitoring Instructional Responsiveness: Reading, MIR:M- Monitoring Instructional Responsiveness: Math, TCAP:RSS- Tennessee Comprehensive Assessment Program: Reading Scale Score, TCAP:MSS- Tennessee Comprehensive Assessment Program: Math Scale Score; N=372

Data Analyses

To answer the research questions, a correlational study was conducted using non-parametric statistics. The analyses for each question is addressed in turn.

Research Questions

*What is the efficacy of early-in-year curriculum-based measures and measures of teacher perception in screening for gifted status in reading and math? T*o answer this question, several subordinate questions were developed.

Question 1: Adequate Ceiling of CBM

1. Do curriculum-based measures of reading and math (as measured by the MIR:R and MIR:M) provide sufficient ceiling to serve as screeners for gifted and high ability students (G/HA) in a general education classroom sample?

To determine if the MIR:R and MIR:R have a sufficient ceiling for use as a screening instrument for gifted and high ability children, the raw MIR:R and MIR:M data were converted to z-scores using SPSS© software. It is predicted that CBM as measured by the MIR:R and MIR:M can yield z-scores that provide evidence of adequate ceiling to screen for students performing at gifted levels. A test *ceiling* is the topmost performance limit assessed by an intelligence or achievement. The accurate assessment of gifted children is frequently confounded by c*eiling effects*, the inability of instrumentation to adequately assess the upper limits of student ability. This effect should be viewed as a limitation of the instrument, rather than the examinee's ability. Frequently, gifted students are not allowed access to out-of-level testing, which might mitigate the problem of ceiling effects. Consequently, achievement scores used to measure gifted students' performance are likely to be clustered at the top levels of performance metrics; thus, the performance of this population exhibits a lack of heterogeneity that reduces variability

and "…leads to attenuated reliability coefficients" known as *restriction of range* (Kieffer et al., 2010).

Generally, *adequate ceiling* for gifted identification is defined as instrumentation that allows for z-scores equal to or higher than 2 standards deviation (*SD*s) above the mean. A *z-score*, also known as a *standard score*, indicates how many standard deviations an individual case is from the mean score for the entire data set (i.e., mean equals zero). Of interest was the presence of z-scores +2 standards of deviation (*SD*s) above the mean, a traditional cutoff score for advanced attainment. Student performance on the MIR probes at this level might be a potential indication of gifted performance, meaning that the MIR:R and MIR:M allow for a ceiling sufficient for use as a gifted screening instrument. This is especially important as the MIR is a measure of on-grade-level performance containing no prompts with content above the grade level for which it was designed. The utility of the MIR as a screening instrument for giftedness increases its value.

For this study (focused on screening for giftedness), z-score distributions and frequencies are examined at levels at or above the 85th (Renzulli, 2010), 90th, 95th (TN Gifted Identification Matrix, 2011), and 97.8th (Bracken, 1987) percentiles, standard cutoff scores from the literature. Adequate ceiling is also evaluated through examination of item gradient and nature of the z-scores; when evaluating sufficient item gradient adjacent raw score items should not convert to adjacent z-score values having an interval of more than .33 (though intervals may be may larger at higher and lower ends of scale

distribution (Bracken, 1987). MIR scores were converted to z-scores and examined at the following levels: 85th% z = 1.00, 90th% z = 1.28, 95th% z = 1.65, 97.5th% z = 1.96, 97.8th% z = 2.0. Also examined were the intervals of item gradients for scores at the top of the distribution.

Performance assessments of gifted and high ability children can be restricted in two ways; the failure to meet the assumptions of statistical analyses though a lack of variability, and the limitations of testing instruments to adequately evaluate the upper limits of G/HA students' ability. This can be problematic in the gifted population because, "Without access to a highest point of potential performance, it is hard not only to judge the effectiveness of an intervention, but also to determine what might be appropriate services for such individuals" (Subotnik & Thompson, 2010). A grade-level measure that allows sufficient ceiling to measure at gifted levels can be a valuable tool in screening and identification for gifted abilities.

Question 2: Intra-Domain Correlations

2a) To what degree or extent are the domain-specific MIR (reading and math) scores related to each other for the entire sample; for students in the G/HA group? [MIR:R X MIR:M]

Correlation between sets of data refers to a measure of how well the sets are related to each other. The most common measure of correlation is the *Pearson Correlation* (*Pearson Product Moment Correlation* or PPMC), reported as an "*r*" value ("Pearson *r*"). This value examines the linear correlation between two sets of data with

results between -1 and 1. The closer the value of *r* to 1, the less the variation in the data points around the line of best fit. *High correlations* are correlations of .5 to 1.0 (or -.5 to -1.0); *medium* correlations .3 to .5 (or -.3 to -.5); and *low* correlations .1 to .3 (or -.1 to -.3; Cohen, 1988; Sattler, 2008). The PPMC, however, it is not able to differentiate between dependent and independent variables, or provide any information about the slope of the line of best fit; it only indicates a correlation. Domain-specific correlations between the raw scores (i.e., MIR:R and MIR:M; TCAP:R and TCAP:M) were obtained and evaluated. A Pearson Correlation (Pearson Product Moment Correlation or PPMC) was used in this study to report and interpret *r* values. It is anticipated that the correlation between domain specific MIR scores is significant at a confidence interval of <.05 (95% surety) and is medium in magnitude (.3 to -.3) or larger as defined by Cohen (1988) for both the entire sample and the G/HA group. For the latter analysis, the G/HA group will be defined as those scoring at or above the 85th %ile on TCAP Reading and TCAP Math composite scores.

2b) To what degree or extent are the domain-specific TR (reading and math) scale scores related to each other for the entire sample?

A Pearson Correlation was used to compare domain-specific TR data. Other questions use a Kendall's tau for correlation with TR instruments, a preference for testing that allows for tied scores that may occur when converting raw score data to rank data for comparisons between instruments. The teacher rank inherently lacks tied scores, or the possibility, so the more popular PPMC was used.

2c) To what degree or extent are the domain-specific TCAP (reading and math) scale scores related to each other for the entire sample? [TCAP:R X TCAP:M]

It is expected that the correlation between domain-specific TCAP scale scores is significant at the <.05 level and is medium in magnitude (.5 to -.5) or larger as defined by Cohen (1988). A Pearson Correlation (Pearson Product Moment Correlation or PPMC) was used in this study to report and interpret *r* values.

2d) To what degree or extent is the magnitude of the MIR correlations comparable to those of the TCAP correlations for the entire sample? [(MIR:R X MIR:M) X (TCAP:R X TCAP:M)]

The magnitude of the difference between the MIR intra-correlation coefficient when compared to the TCAP intra-correlation coefficient is non-significant at $p > .05$ (Hinkle, Wiersma, & Jurs, 1988). Steiger's *z*-test for "correlated correlations" within a population (as described by Meng, Rosenthal, & Rubin, 1992) is used instead of Hotelling's *t* which can overestimate the *t*-value, resulting in a Type I error. Hotelling's *t* uses actual correlation values, even though *r*-values are not normally distributed. Instead, use Fisher's transformation, changing *r* to a *z*-score, and use *z*s in the significance testing formula (which are normally distributed). The *z*-critical values do not depend on degrees of freedom (*df*), and so are consistent for all analyses.

Obtaining additional evidence of psychometric qualities of MIR:R and MIR:M was important to more fully establishing their validity and generalizability for use for students at various skill levels. It is expected that the MIR:R and MIR:M measures,

though developed independently, will yield highly correlated scores because the skill sets share etiology and develop at commensurate rates. These data taken concurrently and administered to the same population allowed for enhanced assessment of the instrumentation. Additional strength would accrue to the validity of the MIR if the correlation coefficient (between academic domains) is similar to that of the TCAP correlation, which has been more extensively evaluated. These data also provide insights into the generalizability (transferability) of the MIR instrumentation.

Question 3: Inter-instrument Correlations of Early Instruments

3. To what degree or extent are the MIR-R and MIR-M correlated with TR as a measure of teacher perception (TR:R, TR:M) for the entire sample? [MIR:R X TR:R] [MIR:M X TR:M]

This question was answered by performing correlations using Kendall's tau (τ), a rank correlation coefficient that specifically measures rank correlation by the similarity of the orderings in the data sets establishing whether two variables may be regarded as statistically dependent. A *rank correlation* is a statistic that measures the relation either between rankings of different ordinal variables, or different discrete rankings of the same variable. A *ranking* is the assignment of the labels "*first* (1)", "*second* (2)", "*third* (3)", etc. A *rank correlation coefficient*, such as *Spearman's* ρ, *Kendall's* τ, and *Goodman* and *Kruskal's* γ measures the degree of similarity between two rankings producing a single coefficient as measure of the statistical dependence between two variables. The coefficient measures between ±1.

As indicated above, a higher-rank correlation coefficient implies increasing agreement between rankings, such that perfect agreement between the two rankings has a value of 1, perfect disagreement -1 (one ranking is the reverse of the other), and zero if the rankings are completely independent (i.e., no relationship). The sign of the correlation indicates the direction of association between the independent variable (X) and the dependent variable (Y). If the value of the DV tends to increase as IV increases, the correlation coefficient is positive; conversely, if the value of the DV decreases as the IV decreases, the correlation coefficient is negative. A correlation of zero indicates that there is no tendency for the DV to either increase or decrease when IV increases.

Correlations between the instruments were performed using SPSS©. The nature of the TR data as ordinal data limited the availability of many statistical tests. To facilitate analyses, MIR data were re-coded by teacher identification number into rank order data using SPSS© functions. The Kendall rank coefficient is *non-parametric*, as it does not rely on any assumptions about the distributions of the independent variable (X) or dependent variable (Y) or the linear relation between X and Y. Kendall tau-b allows for ties in ranked data which may occur when converting the scale data to rank order. *Kendall's tau* (τ) is appropriate when the IV (X) and the DV (Y) are not related by a linear function, or when this may be in question. When hypothesis testing, the coefficient has an expected value of zero. Non-zero coefficients indicate the strength and direction of the correlation. The *tau-b* statistic was used in the present research as it allowed adjustments for tied values which arose when re-coding the MIR data. Kendall's tau-b

value was reported as a range from -1 to +1 and interpreted as with other correlation values. An approximate 95% confidence interval (*CI*) and two sided (H1 dependence) *p*-value with significance at the $p < .05$ level was used as above. It is expected that the correlation between domain-specific MIR scores and the domain-specific TR scores is significant at the $p < .05$ level or less and is medium in magnitude (.5 to -.5) or larger as defined by Cohen (1988).

Question 4: Inter-instrument Correlations (Early to Late)

4a) To what degree or extent are early-in-year CBM (as measured by MIR: R and MIR:M) correlated with the TCAP as an example of end-of-year measure? [MIR:R X TCAP:R] [MIR:M X TCAP:M]

It is expected that the correlation between domain-specific MIR scores and the domain-specific TCAP scores is significant at the $p < .05$ level or less and is medium in magnitude (.5 to -.5) or larger as defined by Cohen (1988). A Pearson Correlation (Pearson Product Moment Correlation or PPMC) as described above was used to report and interpret *r* values as significant at the <.05 level and as small, medium, or large as defined by Cohen (1988).

4b) To what degree or extent are TRs of reading and math as examples of early-in-year measures correlated with the TCAP as an example of end-of-year measure? [TR:R X TCAP:R] [TR:M X TCAP:M]

To answer this question TCAP data were converted to rank order data using SPSS© functions and then correlated to TR data using Kendall's tau-b (τ), which allows

for ties in ranked data that may occur when converting the scale data to rank order. The tau-b value ranges from -1 to +1 and was interpreted as with other correlation values. An approximate 95% confidence interval (*CI*) and two sided (H1 dependence) *p*-value with significance at the $p < .05$ level were used as above. The correlation between domain-specific TR scores and the domain-specific TCAP (recoded) scores was expected to be significant at the $p < .05$ level or less and medium in magnitude (.5 to -.5) or larger as defined by Cohen (1988).

4c) To what degree or extent can the MIR and TR (in reading and math) collectively predict TCAP scores? [MR: DV= MIR:R, TR:R, IV= TCAP:R] [MR: DV= MIR:M, TR:M, IV= TCAP:M]

This question was answered using a multiple regression. It was anticipated that the combined effects of the MIR and TR significantly predict TCAP scores as demonstrated by the percentage of variance accounted for using multiple regression analyses. A Pearson's Bivariate Correlation among all independent variables was used; the correlation coefficients need to be smaller than .08. Report *f*(df), *p*-value as significant at $p < .05$, r^2, confidence interval (CI), and correlations between variables as *r*-values, and beta (β) values.

Question 5: Screening Rates and Group Assignment

5. Do the MIR, TR, and TCAP identify the same cases of G/HA students based on dichotomous gifted group assignment; assignment is defined as at or above the 85^{Th} percentile for MIR and TCAP, and as the top two ranks for the TR?

It was predicted that the rate of agreement in gifted group assignment between the MIR, TR, and TCAP is significantly greater than chance; significance is at $p < .05$ and is medium in magnitude (.3 to -.3) or larger as defined by Cohen (1988). This question will be answered using a non-parametric statistical test, Cochran's Q test, an extension to the McNemar test for related samples. McNemar's test assesses the significance of the difference between dichotomous dependent variables, between two related groups, or two correlated proportions, such as when two measures are taken from the same population sample using a repeated measure, such as pretest/posttest study designs. The test is similar to a *paired-samples t-test*, but uses dichotomous rather than continuous dependent variables. Only three assumptions must be met for its use: 1) there must be a categorical dependent variable with two dichotomous categories and one categorical independent variable with two related groups such as a pretest-posttest, matched pairs or case-control study design; 2) dependent variable categories must be mutually exclusive, i.e., a case cannot be assigned both conditions of the dichotomous state; 3) The cases (participants) should be a random sample from the population; however, in practice, this assumption is not always met. The statistical significance level is a single coefficient reported as a *p-value*. Significance was reported in a manner similar to other *p*-value statistics, and is indicated when $p < .05$.

In a similar manner, Cochran's Q test is a procedure for testing whether the proportions of 3 or more dichotomous variables are equal in the sample. The domain specific TR, MIR, and TCAP variables were re-coded into dummy variables for both

domains of each instrument using gifted screening cutoff scores as described above to establish group assignments; values assigned were 0 = non-gifted group assignment, 1 = screening levels for gifted group assignment. Domain specific correlations were made to establish rates of agreement in identification of gifted students.

Chapter 4

Analyses and Results

Results and Discussion

To complete the intended analyses, some variables had to be recoded, transformed or calculated creating several new variables. This process is detailed here:

- *z-scores*- to answer questions about the test ceiling of the MIR:R and MIR:M, data were converted to *z*-scores during the initial examination of frequency and central tendency; this was completed automatically using the feature in SPSS.
- *Item gradient*- to answer questions about the item gradient of MIR:R and MIR:M, cases in each variable were sorted in SPSS© by *z*-scores. The interval between each *z*-score was hand coded starting with the *z*-score of zero and working toward the termini of the distribution. Adjacent *z*-scores that were equivalent to the preceding score were coded as no change, i.e., zero (0).
- *Rank order*- to compare the MIR and TCAP instruments with the TR instrument, the former were converted to rank order data. Using the feature in SPSS to recode these values automatically, the domain-specific MIR and TCAP were converted to rank order by teacher code; each class had its own ranking of each instrument in each domain as was the case for the TR data. Visual inspection of this variable confirmed the anticipated presence of tied scores, occurring as a result of two or more cases of MIR or TCAP having the same raw score. MIR:R had fewer tied values than MIR:M, attributable, perhaps, to the much smaller range of values for

MIR:M. Allowing tied values to retain the integrity of the data set was deemed preferable to assigning a unique value to each rank. Tied values in the ranking require the use of Kendall's tau for analyzing the relations.

- *Reverse coding*- For the multiple regression in Question 4, TR variables (TR:R and TR:M) were reversed coded using an automatic function in SPSS to create a new variable used only in the regression analyses. This recoding, by teacher code (i.e., by class) reassigned the value of 1, previously high, to a low position. This was done for convenience when interpreting and reporting regression results. Otherwise, all TR values related to the regression would have been negative.
- *Group assignment*- to compare performance on the three instruments using non-parametric correlations, the cases had to be recoded into new dichotomous variables for the Cochran's Q and McNemar's tests. This was completed using the automatic recode feature of SPSS. For the TR:R and TR:M, only the first two ranking placements ("1" and "2") were assigned gifted group membership; thus, each of the classes was represented by two cases. Scores below the gifted screening cutoff scores determined per instrument at values closest to the 85th percentile for MIR and TCAP were assigned a value of zero ("0" non-group assignment) and scores equal to or above the cutoff were coded as "1" (gifted group assignment). It should be noted that these variables were created using cutoff scores particular to the instrument. Group membership for MIR and TCAP was determined by the previously established cutoff scores as described in the

Method and was irrespective of the newly created rank order variable described on the previous page. Thus, in some classes, no student attained MIR or TCAP scores above the screening cutoff; that is, despite presence of class rankings of "1" or "2" in the rank order conversion of MIR and TCAP scores, some cases are not represented in the filtered variable based on the cutoff values of the raw scores. As a result, some classes are not represented at all in the dichotomous variable, and some classes may be represented by as many as five cases.

Research Questions and Findings

Question 1: Adequate Ceiling of CBM

1) Do curriculum-based measures of reading and math (as measured by the MIR:R and MIR:M) provide sufficient ceiling to serve as screeners for gifted and high ability students (G/HA) in a general education classroom sample?

Adequate ceiling for gifted identification is defined as instrumentation that allows for *z*-scores equal to or higher than 2 standard deviations (*SD*s) above the mean (Steiger, 1980). For this study (focused on screening for giftedness), *z*-score distributions and frequencies were examined at levels using standard cutoff scores from the literature or as indicated by state policies; at or above the 85th (Renzulli, 1990), 90th and 95th (TN Gifted Identification Matrix, 2011), and 97.8th (Bracken, 1987) percentiles. Ceiling adequacy was also evaluated through the examination of item gradient and the nature of the *z*-scores; when evaluating for sufficient item gradient, adjacent raw score items

should not convert to adjacent *z*-score values having an interval of more than .33, though intervals may be larger at higher and lower ends of scale distribution (Bracken, 1987).

MIR:R Test Ceiling and *z*-scores

The MIR:R sample yielded a *z*-score range from -1.40 to 2.99. At the 85th percentile (85.8%) *z*-scores equaled 1.12, representing 58 cases or a screening rate of 15.59%; at the 90th percentile (90.1%), *z*-score = 1.35; at the 95th percentile (95.2%), *z*-score = 1.98; at the 97.8th percentile (97.8%), *z*-score = 2.26. Eighteen cases had a *z*-score at or above 2.00 at the 95.4 percentile or higher and equivalent to a gifted screening rate of 4.83%.

Table 7. MIR:R *z*-score Frequency and Percent at Target Percentiles

Raw Score	*z*-score	Cumulative Percent
179	1.12	85.8
194	1.35	90.1
233	1.98	95.2
251	2.25	97.6
251	2.26	97.8

Note: MIR:R- Monitoring Instructional Response: Reading

MIR:R Item Gradient

To examine the item gradient for scores at the top of the distribution, the interval between each *z*-score was calculated, working from $z = 0$ toward each end of the *z*-score continuum, as described above. The average interval for the MIR:R scores was .01; the interval distances ranged from .01 (lowest non-zero value) to .90. Larger

intervals occurred at the higher levels of the distribution above a raw score of 202, *z*-score = 1.48 (MIR:R highest score in sample = 297); however, Bracken (1987) allows for increased tolerance for higher gradient values toward the top of the distribution. Larger intervals also represent the largest gaps between non-consecutive raw scores, as expected. However, distributing the interval range over the score range, even in the case of the largest gradient interval (.90 item gradient interval / 6 point score range [202 to 208]) the average gradient interval was an acceptable .15.

Table 8. MIR:R *z*-score Item Gradient Distances

Gradient Interval	Raw Scores	Interval Range	*z*-score Interval
.11	233 to 241	8 points	1.98 to 2.09
.14	288 to 297	9 points	2.84 to 2.98
.16	277 to 286	9 points	2.67 to 2.81
.17	213 to 224	11 points	1.66 to 1.83
.34	251 to 273	21 points	2.26 to 2.60
.90	202 to 208	6 points	1.48 to 1.57

Note: MIR:R- Monitoring Instructional Response: Reading

MIR:M Test Ceiling and z-scores

The MIR:M yielded a *z*-score range from -2.23 to 3.16. At the 85th percentile (85.8%) *z*-score = 1.00, representing 59 cases or a screening rate of 15.86%; at the 90th percentile (90.6%), *z*-score = 1.36; at the 95th percentile (95.4%), *z*-score = 1.96; at the

97.8th percentile (98.1%), *z*-score = 2.44. Seventeen cases had a *z*-score at or above 2.00, 96th percentile or higher and equivalent to a gifted screening rate of 4.57%.

Table 9. MIR:M *z*-score Frequency and Percent at Target Percentiles

Raw Score	z-score	Cumulative Percent
34	1.00	85.8
37	1.36	90.6
42	1.96	95.4
45	2.32	97.3
46	2.44	98.1

Note: MIR:M- Monitoring Instructional Response: Math

MIR:M Item Gradient

To examine the item gradient for scores at the top of the distribution, the interval between each *z*-score was calculated as described for the MIR:R. The average interval for the MIR:M scores was .02; the interval distances ranged from .05 (lowest non-zero value) to .24. The largest interval occurred at the higher levels of the distribution above a raw score of 49, *z*-score = 2.80 (MIR:M highest score in sample = 52). The largest interval also represents the largest gap between non-consecutive raw scores, as expected.

Table 10. MIR:M *z*-score Item Gradient Distances

Gradient Interval	Raw Scores	Interval Range	*z*-score Interval
.24	49 to 51	2 points	2.80 to 3.04

Note: MIR:M- Monitoring Instructional Response: Math

Question 1: Interpretation and Discussion

In this sample of data from the MIR instruments, reading and math probes both show the requisite psychometric properties for use as gifted screening instruments in settings allowed by its authors as adjudged by the examination of *z*-scores and item gradients. The MIR probes have sufficient ceilings to allow high performing students to attain *z*-score levels at or above two standard deviations above the mean score (MIR:R = 18 students and MIR:M = 17 students at or above this cutoff), as defined by Steiger (1980). Another evaluation of the MIR:R and MIR:M scores examining the item gradient intervals above the 85th percentile indicates that adjacent raw score items do not convert to adjacent *z*-score intervals >.33. As described by Bracken (1987), this characteristic of the adjacent *z*-score intervals indicates that the raw score distribution has acceptable psychometric properties. The MIR probes have been shown by other authors to have measurement capabilities comparable to other available instrumentation when used with general education school populations, and to have moderate correlations to high-stakes, end-of-year testing. This study extends the application of the MIR probes to use as a screening tool for above grade-level populations as part of a gifted identification process.

Establishing the suitability of the MIR universal screeners as valid screening instruments for gifted students is fundamental to the hypothesis testing of the remaining questions.

Question 2: Intra-Domain Correlations

2a) To what degree or extent are the domain-specific MIR (reading and math) scores related to each other for the entire sample and for students in the G/HA group? For the latter analysis, the G/HA group was defined as those scoring at or above the 85th percentile on TCAP:R and TCAP:M scores.

H_0- There is no significant correlation between domain-specific MIR scores as defined by Cohen (1988) for both the entire sample and the G/HA group.

H_1- The correlation between domain-specific MIR scores is significant at the p <.05 level and is *medium* in magnitude (−0.5 to −0.3/0.3 to 0.5) or larger as defined by Cohen (1988) for both the entire sample and the G/HA group.

A Pearson Correlation (Pearson Product Moment Correlation or PPMC) was used to compare domain-specific MIR data. The PPMC yields a single correlation coefficient (r) to represent the extent to which two variables are related; r values are interpreted as significant at the p <.05 level and as *no correlation* (−0.09 to 0.0/0.0 to 0.09); *small* (−0.3 to −0.1/ 0.1 to 0.3); *medium* (−0.5 to −0.3/0.3 to 0.5); or *large* (−1.0 to −0.5/0.5 to 1.0) as defined by Cohen (1988).

2a) MIR:R X MIR:M Correlation

The correlation coefficient between MIR:R scores and the MIR:M scores for the entire sample is significant (p <.000; 2-tailed) and *small* (r = .28; −0.3 to −0.1/ 0.1 to 0.3) as defined by Cohen (1988).

Pearson correlations were also used to compare domain-specific MIR data after screening for gifted-group assignment by TCAP performance at 85th percentile and above (Renzulli, 1990). The populations identified by TCAP:R (n = 70) and TCAP:M (n = 64) were distinct; two tests were conducted comparing MIR:R to MIR:M at the 85th percentile of each TCAP; that is, MIR:R and MIR:M were correlated twice, once at the 85th percentile of TCAP:R and again for TCAP:M.

The correlation coefficient between MIR:R scores and the MIR:M scores screened by TCAP:R scale scores above the gifted screening level of 85$^{\%}$ is not significant (p > .05;) and *small* in magnitude (r = .15; −0.3 to −0.1/ 0.1 to 0.3) as defined by Cohen (1988) for the G/HA group.

The correlation coefficient between MIR:R scores and the MIR:M scores screened by TCAP:M scale scores above the gifted screening level of 85% is not significant (p > .05; r = .19), and *small* (−0.3 to −0.1/ 0.1 to 0.3) as defined by Cohen (1988) for the G/HA group.

Question 2a: Interpretation and Discussion

Theory suggests that commensurate development of reading and math skills should be anticipated. It was hypothesized that performance on the MIR:R and MIR:M

probes would be significantly and moderately correlated (*medium*, −0.5 to −0.3/0.3 to 0.5) being coincidental measures on the same population. Mitigating factors associated with the analysis of data sets comprised by the upper percentiles include problems arising from a lack of variability to parse when performing the analyses, i.e., the scores are closely clustered at the top lead to a restriction of range (Thompson & Subotnik, 2010). Correlation coefficients on MIR tests were significant, but the magnitude is *small*. After TCAP screening at the 85th percentile, the correlations remain *small* in magnitude and lack significance. To accept the alternative hypothesis, correlations must evidence significance at a moderate (*medium*) level. Though the MIR:R by MIR:M correlates were significant, the correlation was insufficiently strong for acceptance; correlates for gifted-screened data lacked the requisite significance and magnitude. The null hypothesis was accepted.

2b) TR:R X TR:M

2b) To what degree or extent are the domain-specific TR (reading and math) scale scores related to each other for the entire sample?

H_0- There is no significant correlation between domain-specific TR scores as defined by Cohen (1988).

H_1- The correlation between domain-specific TR scores is significant at the $p < .05$ level and is *medium* in magnitude (−0.5 to −0.3/0.3 to 0.5) or larger as defined by Cohen (1988).

A Spearman rho Correlation was used to compare domain-specific TR data. Other questions use a Kendall's tau for correlation with TR instruments, a preference for testing that allows for tied scores that may occur when converting raw score data to rank data for comparisons between instruments. The teacher rank inherently lacks tied scores, or the possibility, so the Spearman was used. The correlation coefficient between TR:R scores and TR:M scores is significant (p <.01, 2-tailed) and *large* in magnitude (*rho* = .71; −1.0 to −0.5/0.5 to 1.0) as defined by Cohen (1988).

2b) Interpretation and Discussion

The magnitude of the TR correlation coefficient supports theories that anticipate commensurate development of reading and math skills. It was hypothesized that the TR:R and TR:M probes would be significantly correlated being coincidental measures on the same population. The correlation coefficient is both significant (p < .000) and *large* in magnitude (−1.0 to −0.5/0.5 to 1.0) as defined by Cohen (1988). Thus, the null hypothesis (no significant correlation between TR tests) was rejected.

2c) TCAP:R X TCAP:M

2c) To what degree or extent are the domain-specific TCAP (reading and math) scale scores related to each other for the entire sample?

H_0- There is no significant correlation between domain-specific TCAP scores as defined by Cohen (1988).

H_1- The correlation between domain-specific TCAP scores is significant at the $p < .05$ level and is *medium* in magnitude (−0.5 to −0.3/0.3 to 0.5) or larger as defined by Cohen (1988).

The correlation coefficient between TCAP:R scores and TCAP:M scores is significant ($p < .01$, 2-tailed) and *large* in magnitude ($r = .71$ −1.0 to −0.5/0.5 to 1.0) as defined by Cohen (1988).

2c) Interpretation and Discussion

The magnitude of the TCAP correlation coefficient supports theories that anticipate commensurate development of reading and math skills. It was hypothesized that the TCAP:R and TCAP:M probes would be significantly correlated being coincidental measures on the same population. As standardized state tests, the TCAP tests have undergone extensive evaluation of their psychometric properties. The correlation coefficient is both significant ($p < .000$) and *large* in magnitude (−1.0 to −0.5/0.5 to 1.0) as defined by Cohen (1988). Thus, the null hypothesis (no significant correlation between TCAP tests) was rejected.

2d) (MIR:R X MIR:M) X (TACP:R X TCAP:M)

2d) To what degree or extent is the magnitude of the MIR correlation comparable to that of the TCAP correlation for the entire sample?

H_0- There is no significant difference in the magnitude of the domain-specific MIR correlation (MIR:R X MIR:M *z*-scores) and the magnitude of the domain-specific TCAP correlation (TCAP:R X TCAP:M *z*-scores) as defined by Cohen (1988).

H_1- The correlation between the magnitude of the domain-specific MIR correlation and the magnitude of the domain-specific TCAP scores is significant at the p <.05 level and is *medium* in magnitude (−0.5 to −0.3/0.3 to 0.5) or larger as defined by Cohen (1988).

Steiger's *z*-test for "correlated correlations" within a population (as described by Meng, Rosenthal, & Rubin, 1992) was used. The *z*-scores (after a Fisher's transformation to convert *r*-scores to *z*-scores which are normally distributed) were used in the significance testing formula (Hinkle, Wiersma, & Jurs 1988). The *z*-critical values do not depend on degrees of freedom (*df*), and so are consistent for all analyses.

After calculating Pearson correlations comparing MIR:R *z*-scores to MIR:M *z*-scores ($r_{(1)}$ = .28, n = 372) and comparing TCAP:R *z*-scores to TCAP:M *z*-scores ($r_{(2)}$ = .71, n = 372), an online calculator was used to compare the correlations (at http://vassarstats.net/rdiff.html), producing a new *z*-value to assess the significance of the difference in magnitude of the two correlation coefficients. Because the MIR correlation (r_1) which entered first is smaller than the TCAP correlation (r_2), the sign of *z* is negative. Steiger (1980) demonstrated that *correlation X correlation* *z*-values greater than |1.96| are considered significant when a 2-tailed test is performed. The magnitude of the difference between the MIR intra-correlation coefficient when compared to the TCAP intra-correlation coefficient is significant at p <.05 with a *z*-score value of -8.07.

2d) Interpretation and Discussion

The *correlation X correlation* *z*-value (z = -8.07) is greater than the requisite |1.96| standard for significance (Steiger, 1980). The null hypothesis (no significant difference in

the magnitude of the domain-specific MIR correlation [MIR:R X MIR:M *z*-scores] and the magnitude of the domain-specific TCAP correlation [TCAP:R X TCAP:M *z*-scores]) was rejected. The significance of the correlation indicates there is more content overlap in the two TCAP measures than between the two MIR measures.

Question 3: Inter-instrument Correlations of Early Instruments

3) To what degree or extent are the domain-specific MIR-R and MIR-M correlated with domain specific TR:R and TR:M as a measure of teacher perception for the entire sample?

H_0- There is no significant correlation (as defined by Cohen, 1988) for the entire sample between domain-specific MIR-R and MIR-M when correlated with domain-specific TR:R and TR:M as a measure of teacher perception.

H_1- The correlation for the entire sample between the domain-specific MIR-R and MIR-M when correlated with domain-specific TR:R and TR:M as a measure of teacher perception is significant at the $p < .05$ level and is *medium* in magnitude (−0.5 to −0.3/0.3 to 0.5) or larger as defined by Cohen (1988).

This question was answered using a Kendall's tau-b (τ). The MIR data were converted as described at the beginning of this section to rank order data using SPSS functions. Rank order for the MIR data was coded by teacher code (i.e. by class) with the first position (1) as the highest rank. Rank order ties occur when two or more cases in the sample share the same raw score. Kendall tau-b allows for ties in ranked data which may occur when converting the scale data to rank order. As other correlation coefficients,

Kendall's tau-b values range from -1 to +1 and are interpreted in a manner similar to other correlation values established by Cohen (1988). Tau- values are interpreted by the same significance scale used for *p*-values ($p < .05$, *small*, *medium*, *large*; two-tailed) and using a similar confidence interval (.05)

MIR:R X TR:R

The MIR:R (rank conversion) and the TR:R are significantly ($p < .000$) and positively correlated ($\tau = .37$), and the coefficient is *medium* in magnitude.

MIR:M X TR:M

The MIR:M (rank conversion) and the TR:M are significantly ($p < .000$) and positively correlated ($\tau = .25$); the coefficient is *small* in magnitude.

3) Interpretation and Discussion

For this question, the null hypothesis was rejected for the reading domain, but accepted for the math domain. The magnitude of the reading domain coefficient is *medium* (−0.5 to −0.3/0.3 to 0.5), while the math domain correlation coefficient is *small* (−0.3 to −0.1/ 0.1 to 0.3) as defined by Cohen (1988). The significance ($p < .000$) in the tau coefficients for the MIR (rank conversion) by TR provides some evidentiary support to the validity of the psychometric characteristics of the MIR probes and informs the process of measuring teacher perception; corroboration between the measures adds to the construct validity of each. In this sample, there is a stronger correlation between the measures of the reading domain. However, the mechanism of the correlation may be a function of the validity of either the MIR probes and/or measurements of teacher

perception (i.e., related to the instrumentation) or in the ability of either MIR probes and/or teachers to assess reading skills compared to math skills (i.e., related to the domain). Multiple interpretations are possible; these may influence each other in such ways as may not be mutually exclusive. Important here, however, is the degree of concurrence in two early, domain-specific measures assessing levels of student attainment. This agreement inspires some confidence, more so in reading, in the validity of early measures for improved comparisons to later measures.

Question 4: Inter-instrument Correlations (Early to Late) and Predictability

4a) MIR X TCAP

4a) To what degree or extent are early-in-year CBM (as measured by MIR: R and MIR:M) correlated with the TCAP as an example of end-of-year measure?

H_0- There is no significant correlation (as defined by Cohen, 1988) for the entire sample between domain-specific MIR-R and MIR-M as examples of early curriculum-based measures when correlated with the domain-specific TCAP:R and TCAP:M as examples of later measures.

H_1- The correlation for the entire sample between the domain-specific MIR-R and MIR-M as examples of early measures with domain-specific TCAP:R and TCAP:M as examples of later measures is significant at the $p < .05$ level and is *medium* in magnitude (−0.5 to −0.3/0.3 to 0.5) or larger as defined by Cohen (1988). Pearson correlations were used to compare domain-specific MIR data to domain-specific TCAP data.

Reading

The domain-specific MIR:R score and the domain-specific TCAP:R score are significantly, positively correlated (p <.01, 2-tailed) and is *large* in magnitude (r = .51, −1.0 to −0.5/0.5 to 1.0) as defined by Cohen (1988).

Math

The domain-specific MIR:M score and the domain-specific TCAP:M score are significantly, positively correlated (p <.01, 2-tailed) and the correlation is *medium* in magnitude (r = .38, −0.5 to −0.3/0.3 to 0.5) as defined by Cohen (1988).

4a) Interpretation and Discussion

The null hypothesis was rejected when answering this question. The significant correlations between domain-specific MIR scores and the domain-specific TCAP scores provides evidence that the MIR probes have some validity when used as early curriculum-based measures to assess student progress toward attainment of TCAP learning goals. As above, the correlation in reading was stronger (*large* at r = .51) when compared to that of math (*medium* at r = .38). This level of agreement helps support the validity of the use of MIR probes as curriculum-based measures to evaluate student progress toward learning goals as assessed by the TCAP scores. The strength of these correlations also improves the construct validity of the MIR probes as the psychometric properties TCAP measures have been more extensively evaluated. Moreover, the correlations at this magnitude support the use of the MIR tests as indicative of future performance.

4b) TR X TCAP

4b) To what degree or extent are TRs of reading and math as examples of early-in-year measures correlated with the TCAP as an example of end-of-year measure?

H_0- There is no significant correlation (as defined by Cohen, 1988) for the entire sample between domain-specific TR-R and TR-M as examples of early measures when correlated with the domain-specific TCAP:R and TCAP:M (rank conversion) as examples of later measures.

H_1- The correlation of the entire sample between the domain-specific TR-R and TR-M as examples of early measures with domain-specific TCAP:R and TCAP:M (rank conversion) as examples of later measures is significant at the p <.05 level and is *medium* in magnitude (−0.5 to −0.3/0.3 to 0.5) or larger as defined by Cohen (1988).

This question was answered using a Kendall's tau-b (τ). TCAP data were converted to rank order data using SPSS function through a process described in question 3. Rank order was coded by teacher code (i.e. by class) with the first position (1) as the highest rank.

Reading

The TR:R and the TCAP:R (rank conversion) are significantly, positively correlated (p < .000, 2-tailed; τ = .53); the correlation is *large* in magnitude.

Math

The TR:M and the TCAP:M (rank conversion) are significantly, positively correlated (p < .000, 2-tailed; τ = .42); the correlation is *medium* in magnitude.

4b) Interpretation and Discussion

The null hypothesis was rejected in both domains when answering this question. The significance ($p < .000$) of the tau-b coefficient between domain-specific TR scores and the domain-specific TCAP scores provides evidence that a TR has some validity when used to assess student progress toward attainment of TCAP learning goals. As suggested by Gagné (1994), teacher ability to assess student attainment may be better than was once thought to be the case. Once again, the correlation in reading was stronger (*large* at $\tau = .53$) when compared to math (*medium* at $\tau = .42$). However, the magnitude of these correlations helps support the validity of the use of early measures of teacher perception to evaluate student progress toward learning goals as assessed by the TCAP scores and indicative of future performance.

4c) (MIR X TR) X TCAP

4c) To what degree or extent can the MIR and TR (in reading and math) collectively predict TCAP scores?

H_0: The combined effects of the MIR and TR cannot significantly predict TCAP scores as demonstrated by the percentage of variance accounted for when using multiple regression analyses.

H_1: The combined effects of the MIR and TR significantly predict TCAP scores as demonstrated by the percentage of variance accounted for when using multiple regression analyses.

This question was answered using a multiple regression, a multi-step process. *Multiple Regression* (MR) is a statistical technique to assess the relations between one continuous dependent variable (DV) and several independent variables (IVs; continuous or dichotomous). A regression details the amount of variance accounted for in a DV (criterion, y) based on the IVs (predictors, x). The result is an equation that represents the best prediction of a DV from the IVs. Using SPSS, a standard (simultaneous) MR was used in which all IVs enter the regression equation at once. Each IV is assigned only its unique contribution to the relation with the DV. No IV is assigned the overlapping variance. Several coefficients are reported with the results including:

- *R*- the multiple correlation between the obtained and the predicted *y* (DV) values.
- R^{2-} a squared multiple correlation as the proportion of variance in the DV that is predictable from the best linear combination of the IVS. This is also known as the effect size.

By convention, the regression coefficient is represented as *R* in a non-parametric regression. The R^2 (also called the coefficient of determination) reflects the "goodness of fit" for the model and is a percentage of the variance that is explained by the regression. This coefficient represents how well the regression predicts the value of *y* (the criterion). Cohen's table of effect sizes (1988) was used for interpretation of R^2:

- Small effect size: r^2, R^2 = .01
- Medium effect size r^2, R^2 = .09
- Large effect size: r^2, R^2 = .25

Other reporting values include:

- Adjusted R^2- an R adjusted for the size of the sample, as R^2 can overestimate of the relationship.
- *B* weights- the *unstandardized* regression coefficients representing the level of change in the DV associated with a one unit change in an IV while all other IVs are held constant. These weights are in the same metric as the original data.
- β weights (Beta weights)- the *standardized* regression coefficients representing the level of change in the standard of deviation of a DV associated with a one unit change in the standard deviation in an IV while all other IVs are held constant. This is indicative of the strength of the relation between an IV and the DV; the relation is parallel, larger beta weights indicate stronger relations.
- Squared semi-partial correlations (sr_i^2)- the unique contribution of an IV to the total variance of the DV, that is the amount of variance accounted for by the individual IV.

It is necessary to assure the data provide support for the use of a multiple regression by meeting certain assumptions. Regression requires adequate sample size with outliers eliminated or converted as these can greatly impact the regression equation; in this research, the cleaned data meet this requirement (N = 372) for all variables.

The test begins by assessing the relation of bivariate correlations between the domain-specific variables, or the *collinearity*. Collinearity refers to the linear relation between two variables. Perfect collinear correlations equal /1/, indicating an exact linear

relation between the two. Initial bivariate correlations were assessed using a PPMC among all independent variables; regression is most effective when the variables are related (>/.3/) but not overly so (>/.6/). Pearson coefficients remain denoted as *r*. Unlike the previous Kendall's tau tests using rank conversion scores, for this correlation raw scores were used. TR:R and TR:M were reverse coded (from 1 as high to 1 as low) as described above to provide positive coefficients that facilitate interpretation. Other assumptions were met through an analysis and reporting of other test results.

Multicollinearity, a threat to regression, is reported using two statistics. There is a tendency in the literature to use the terms *collinearity* and *multicollinearity* interchangeably, which, strictly speaking, is not the case. Collinearity is a measure of two variables as they are correlated with each other (Pedhazur, 1997). Multicollinearity is a function of two (or more) IVs *in combination* that predict a substantial percentage of variance in another IV. High multicollinearity lowers both the t-value and the level of significance in the predictor, and destabilizes the B and beta coefficients. Technically, multicollinearity is not a problem in this regression model as there are only two IVs. The statistical metrics to assess multicollinearity are nonetheless provided.

Variance inflation factor (VIF) is a statistic of the severity of multicollinearity in the multiple regression model. This assesses the degree to which the standard error is inflated due to multicollinearity. The value is parallel to the degree of inflation of the standard error; higher *VIF* values indicate a larger threat of multicollinearity; a score of 1 indicates no multicollinearity. Another value, *tolerance* (T) also assesses

multicollinearity. The T value is calculated as an inverse of *VIF*. Recommendations for acceptable levels of tolerance vary in the published literature. Commonly, a value of .10 is recommended as the minimum level of tolerance (Tabachnick & Fidell, 2001). There is no consensus on maximum values of T, as factors such as number of IVs and sample size must be considered.

Independence of error terms are another threat to be assessed. Errors of prediction should be independent of one another with a lack of autocorrelation. Assessed from residual statistics, the Durbin Watson statistic tests for autocorrelation in the regression analysis. The value is always between 0 and 4, with a desired value of 2 indicating no autocorrelation in the sample. Values approaching 0 are interpreted as a positive autocorrelation, and values toward 4 indicate a negative autocorrelation.

The regression itself is similar to an Analysis of Variance (ANOVA) test, but allows for continuous variables such as the MIR and TCAP data; the DV must be continuous. The regression is reported using a value of *F* with degrees of freedom, and significance reported as a *p*-value evaluated by the same parameters above. Finally, individual t-tests are conducted to determine the unique contribution of each IV to the combined total effect. Reported from the t-test are R^2, *Adj.* R^2, *B*, β, and sr_i^2. Threat statistics reported are *VIF*, *T*, and the Durbin-Watson. The regressions are domain-specific. Only the relationship is assessed, no causality should be inferred.

4c) Multiple Regression

Reading

A simultaneous (standard) multiple regression was conducted using MIR:R and TR:R as the IVs and TCAP:R as the DV. As noted at the beginning of this chapter, the TR variable was reverse-coded. It should also be noted that the initial correlations of the regression are Pearson correlates; these correlations will differ from the tau-correlations used in question 3.

The IVs MIR:R and TR:R are significantly ($p < .000$, 1-tailed), moderately and positively correlated with each other ($r = .49$). The correlation is >/.3/ and </.6/. Both the IVs (MIR:R and TR:R) are significantly ($p < .000$, 1-tailed) related to the DV (TCAP:R); MIR:R ($r = .51$) and TR:R ($r = .61$) are moderately (*medium*) positively correlation to TCAP:R. The minimal magnitude of the correlation is >/.3/ for both MIR:R and TR:R. The maximum magnitude of the correlation for the MIR:R is </.6/. The magnitude of the TR:R ($r = .61$) value is allowed being only slightly (.01) over the recommendation.

Table 11. Correlations between Monitoring Instructional Response: Reading (MIR:R), Teacher Ranking: Reading (TR:R), and Tennessee Comprehensive Assessment Profile: Reading (TCAP:R)

		TCAP:R	MIR:R	TR:R
Pearson Correlation	TCAP:R	1.00	.51**	.61**
	MIR:R	.51**	1.00	.49**

Note: ** $p < .01$, $N = 372$

Standard Multiple Regression:

IV: MIR:R, TR:R
DV: TCAP:R
$F(2,369) = 139.78$, $p < .000$, $R^2 = .43$, Adj. $R^2 = .43$

MIR:R: $\beta = .28$, $p < .01$, $sr_i^2 = .06$
TR:R: $\beta = .48$, $p < .01$, $sr_i^2 = .17$

Both variable IVs (MIR:R, TR:R) entered the model. The ANOVA for the test is significant ($p < .000$). The *R* is significantly different from zero (0). This set of IVs significantly predicts (or is related to) the DV, though it cannot be said which IVs were the significant predictors. Tolerance (.76) and the variance inflation factor (VIF; 1.32) are the same for MIR:R and TR:R. *VIF* was acceptable, i.e., close to 1; however, the tolerance was weak (<1), the assumption for multicollinearity may be threatened. The Durbin-Watson (1.91) is also acceptable, as it is >1.5 but <2.5. TR:R is the better predictor of TCAP:R (MIR:R $sri^2 = .06$; TR:R $sri^2 = .17$). The effect size R^2 (.43) indicates that in this model the combination of variables accounts for 43% of the variability in the TCAP:R, which is *large*. For every one-unit change in the TR:R, there is a predicted increase of .48 points ($\beta = 48$) in TCAP:R.

Table 12. The Effect of Monitoring Instructional Response: Reading and Teacher Ranking: Reading on Tennessee Comprehensive Assessment Profile: Reading

Variable	B	β	sr_i^2
MIR:R:	.12	.28***	.06
TR:R:	3.06	.48***	.17

Note: ***$p < .000$; $R^2 = .43$, Adj. $R^2 = .43$, $N = 372$.

Math

A simultaneous (standard) multiple regression was conducted using MIR:M and TR:M as the IVs and TCAP:M as the DV. As noted at the beginning of this chapter, the TR variable was reverse-coded. It should also be noted that the initial correlations of the regression are Pearson correlates; these correlations will differ from the tau- correlations used in question 3.

The IVs MIR:M and TR:M are significantly ($p < .000$, 1-tailed), moderately and positively correlated with each other ($r(369) = .32$). The correlation is >/.3/ and </.6/.

Both the IVs (MIR:M and TR:M) are significantly ($p < .000$, 1-tailed) related to the DV (TCAP:M); MIR:M ($r(369) = .38$) is moderately (*medium*) positively correlated to TCAP:M; TR:M ($r(369) = .51$) has a stronger (*large*) positive correlation to TCAP:M. The magnitude of the correlation is between >/.3/ and </.6/ for both MIR:R and TR:R.

Table 13. Correlations between Monitoring Instructional Response: Math (MIR:M), Teacher Ranking: Math (TR:M), and Tennessee Comprehensive Assessment Profile: Math (TCAP:M)

		TCAP:M	MIR:M	TR
Pearson Correlation	TCAP:M	1.00	.38***	.51***
	MIR:M	.38***	1.00	.32***

Note: *** $p < .000$, $N = 372$

Standard Multiple Regression:
IV: MIR:M, TR:M
DV: TCAP:MSS

$F(2,369) = 84.73$, $p < .000$, $R^2 = .32$, Adj. $R^2 = .31$

MIR:M: $\beta = .24$, $p < .000$, $sr_i^2 = .05$
TR:M: $\beta = .44$, $p < .000$. $sr_i^2 = .17$

Table 14. The Effect Monitoring Instructional Response: Math (MIR:M) and Teacher Ranking: Math (TR:M) on Tennessee Comprehensive Assessment Profile: Math (TCAP:M)

Variable	B	β	sr_i^2
MIR:M:	.83	.24***	.05
TR:M:	3.11	.44***	.17

Note: ***$p < .000$, $R^2 = .32$, Adj. $R^2 = .31$, $N = 372$.

Both variables (MIR:M and TR:M) entered the model. The ANOVA for the test is significant ($p < .000$). The *r* was significantly different from zero (0). This set of IVs significantly predicts (or is related to) the DV, though it cannot be said which IVs were the significant predictors. Tolerance (.90) and the *VIF* (1.11) are the same for MIR:M and TR:M. *VIF* is acceptable, i.e., close to 1; however, the tolerance is weak (<1), the assumption for multicollinearity may be threatened. The Durbin-Watson (1.83) is also acceptable, as it was >1.5 but <2.5. The effect size R^2 (.32) indicates that in this model the combination of variables accounts for 32% of the variability in the TCAP:M, which is *large*. TR:M was the better predictor of TCAP:M (MIR:M $sri^2 = .05$; TR:M $sri^2 = .17$). For every one-unit change in the TR:M, there is a predicted increase of .44 points ($\beta = .44$) in TCAP:M.

4c) Interpretation and Discussion

Assumptions and Issues that must be addressed in MR:

- *Causation* is not in question, though causation is difficult to prove using MR. The regression was used to evaluate *prediction* of TCAP scores through a combination of MIR and TR data.
- Correlations- the IVs in the reading regression are significantly ($p < .000$) and moderately positively (*medium*) correlated with each other: MIR:R & TR:R ($r = .49$). Both the MIR:R ($r = .51$) and TR:R ($r = .61$) are significantly related to the DV (TCAP:R) having *large*, positive correlations. The correlations are >/.3/ and </.6/; except, the TR:R X TCAP:R value of .61 is only very slightly over the recommendation. For the math regression, the IVs MIR:M and TR:M are significantly ($p < .000$) moderately ($r = .32$), and positively correlated with each other. Both MIR:M ($r = .38$) and TR:M ($r = .51$) are significantly ($p < .000$) related to the DV (TCAP:M) having a moderately positive (*medium*) correlation, >/.3/ and </.6/.
- Ratio of cases to IVs- adequate sample size was evaluated using the following formula; for testing R ($N \geq 50+8m$, where m is number of IVs) and for testing individual predictors ($N \geq 104 + m$). The cleaned data set had 372 cases meeting the requirement for testing ($372 \geq 50+8(2) = 372 \geq 50+16 = 372 \geq 66$) and individual predictors ($372 \geq 104 + 2 = 372 \geq 106$).

- Absence of outliers- the presence of outliers can negatively impact the regression equation by artificially increasing the slope, and as a result affect the precision of the estimation of regression weights. Outliers assessed through *z*-scores were changed to a value of 3 SDs + 1 during initial data cleaning.
- Absence of multicollinearity and singularity- these were assessed through evaluation of *r*, T and *VIF* values. No correlations were > |.6|. In the reading regression T =.76 and the *VIF* = 1.32 were the same for MIR:R and TCAP:R and TR:R and TCAP:R. *VIF* was acceptable, i.e., close to 1; however, the tolerance statistic was weak (<1). For the math regression, T = .90 and the *VIF* = 1.11 and were the same for MIR:M and TCAP:M and TR:M & TCAP:M. *VIF* was acceptable, i.e., close to 1; however, the tolerance statistic was weak (<1). The assumption for multicollinearity may be threatened.
- Independence of errors- a Durbin-Watson was used to evaluate errors of prediction and establish that error terms were independent of one another. A value of 2 shows perfect independence. The Durbin-Watson for both the reading (1.91) and math (1.83) regressions was acceptable, being >1.5 but <2.5.
- Normality, linearity, and homoscedasticity of residuals – there is no assumption that the IVs must be normally distributed but the prediction is enhanced if they are.

For this question, the null was rejected. In this model, significant ($p < .000$) and *large* effect sizes were reported for both regressions. The reading R^2 (.43) indicates that

the combination of variables accounted for 43% of the variability in the TCAP:R, in math the R^2 (.32) indicates that the combination of variables accounts for 32% of the variability in the TCAP:M. It is also the case that in both regressions the TR is the better predictor of TCAP scores.

In rejecting the null hypothesis, the claim is made that early-in-year measures of student ability as exampled by a curriculum-based measure MIR) and a teacher ranking (TR) have strong predictive value with large effect sizes toward late-in-year measures as exampled by the TCAP. This effect is stronger in the reading domain (R^2 = .43) than in the math domain (R^2 = .32). Though the combination of MIR and TR was shown to be an effective early predictor, the better, single predictor in both domains is the TR.

Question 5: Screening Rates and Group Assignment

5a) MIR X TR X TCAP: Cochran's Q

5a) Is there a significant difference in the rate MIR, TR, and TCAP identify G/HA students based on dichotomous gifted group assignment? Group assignment is defined as attainment at or above the 85th percentile for MIR and TCAP and as the top two ranks for the TR.

H_0: There is no significant difference in the proportion of identified G/HA students between the MIR, TR, and TCAP based on dichotomous gifted group assignment.

H_1: There is a significant difference in the proportion of identified G/HA students between the MIR, TR, and TCAP based on dichotomous gifted group assignment.

This question was answered using a Cochran's Q test. Recoded, dichotomous, dummy variables were created for both domains (reading and math) of each instrument (MIR, TR, TCAP) for gifted group assignment as defined by gifted screening cutoff standards particular to each instrument (0 = non-gifted group assignment, 1 = gifted group assignment), as described above. Cochran's Q test extends the McNemar test for two related samples, for use of three or more sets of proportions from the same population sample (or matched from similar populations). Cochran's Q is reported as a chi-square value (χ^2), but is also referred to as a Q-value, as is the case here to avoid confusion. The null hypothesis for the Cochran's Q test is that there are no differences between the variables (Sheskin, 2004). If the calculated probability is low, i.e., *p* is less than the selected significance level, the null-hypothesis is rejected indicating that the proportions in at least two of the variables are significantly different from each other. The tests are again divided by domain specificity, reading and math.

Reading

A Cochran's Q test was used to determine the significance of the relation between the three reading instruments (MIR:R, TR:R, TCAP:R) when used as screening tools for gifted-group assignment. Dummy-coding of the variables for gifted group assignment resulted in the following gifted-group assignments: MIR:R identified 58 gifted students, the TR:R 54 students, and the TCAP:R 70 students. The *Q*-value of 3.85(2), is not significant ($p > .05$, $N = 372$) indicating there is no significant difference in the three

instruments of the proportion of identification in this sample when used for gifted screening. MIR:R, TR:R, and TCAP:R screen for giftedness at about the same rate.

Table 15. Cochran's Q correlation between Monitoring Instructional Response: Reading (MIR:R) dichotomously screened for gifted group assignment; Teacher Rank: Reading (TR:R) dichotomously screened for gifted group assignment; and Tennessee Comprehensive Assessment Profile: Reading (TCAP:R) dichotomously screened for gifted group assignment

Cochran's Q = 3.85(2) Asymp. Sig. = .15	0	1
MIR:R screened for gifted group assignment dichotomous	314	58
TR:R screened for gifted group assignment dichotomous	318	54
TCAP:R screened for gifted group assignment dichotomous	302	70

Note: $p = n/s$, $N = 372$, 0 = non-gifted group assignment, 1 = gifted group assignment

Math

A Cochran's Q test was used to determine the significance of the relation between the three math instruments (MIR:M, TR:M, TCAP:M) when used as screening tools for gifted group assignment. The dummy-coding for gifted group assignment had the following results: MIR:M identified 59 gifted students, the TR:M 55 students, and the TCAP:M 64 students. The *Q*-value of 1.15 (2), is not significant ($p > .05$, n = 372) indicating there is no significant difference in this sample among the proportions of the three instruments when used for gifted screening.

Table 16. Cochran's Q correlation between Monitoring Instructional Response: Math (MIR:M) dichotomously screened for gifted group assignment; Teacher Rank: Math (TR:M) dichotomously screened for gifted group assignment; and Tennessee Comprehensive Assessment Profile: Math (TCAP:M) dichotomously screened for gifted group assignment

Cochran's Q = 1.51(2) Asymp. Sig. = .56	0	1
MIR:M screened for gifted group assignment dichotomous	313	59
TR:M screened for gifted group assignment dichotomous	317	55
TCAP:M screened for gifted group assignment dichotomous	308	64

Note: $p = n/s$, $N = 372$, 0 = non-gifted group assignment, 1 = gifted group assignment

5a) Interpretation and Discussion

The null hypothesis is accepted in both domains. A lack of significance in a Cochran's Q test is interpreted as meaning there is no significant difference in the proportion of identified cases in dichotomous variables. In this case, a negative Q-test result is desirable as this indicates no significant difference in the proportions of gifted-group assignment between the three measures. It can be inferred that the MIR, TR, and TCAP each identify cases for gifted-group assignment at approximately the same rate; from which it can be induced, then, that in the domains of reading and math, these early measures (MIR and TR) identify gifted-group assignment at a rate comparable to a later measure (TCAP).

5b) Chi-Square and McNemar Tests

5b) Do the MIR, TR, and TCAP identify the same cases of G/HA students based on dichotomous gifted-group assignment? Group assignment is defined as attainment at or above the 85th percentile for MIR and TCAP and as the top two ranks for the TR.

To further elucidate the findings based on the Cochran's Q test, the Crosstabs feature of SPSS was used to examine the rates of agreement in gifted-group assignment between the instruments. A Chi-Square Test of Independence/Crosstabs is a test of categorical association between the variables and was used to examine the relation among the cases (cells) of the two target variables. A McNemar's test was then used to assess the significance of the difference between the two correlated proportions (lists). In both tests the domain-specific, dichotomous, categorical variables for gifted group assignment were used. As the temporal sequence of the two measures was relevant (early compared to late measures), MIR and TR tests were defined as *before* measures and the TCAP tests as *after* measures. The results are coded as "1" for those subjects who attain gifted-group assignment and as "0" for those who do not (non-gifted group assignment).

Reading

MIR:R X TCAP:R

The results of a chi-square test of independence with MIR:R (dichotomously screened for gifted group assignment) by TCAP:R (also screened for gifted group assignment) showed that these two variables are significantly related, $x^2(1) = 39.03$, $p <$ .000, $N = 372$. Values of the McNemar test lack significance ($p > .05$; Cohen, 1988) indicating that the proportion of gifted students screened by MIR:R and TCAP:R is not statistically significantly different, $p = .20$ (2-tailed; $N = 372$), as anticipated by the Cochran's Q above.

Table 17. Chi-square correlation between Monitoring Instructional Response: Reading (MIR:R) dichotomously screened for gifted group assignment and Tennessee Comprehensive Assessment Profile: Reading (TCAP:R) dichotomously screened for gifted group assignment

Chi-Square Tests	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)
Pearson Chi-Square	39.03[a]	1	.00	
McNemar Test				.20[b]

Table 18. Chi-square correlation percentages of group assignment between Monitoring Instructional Response: Reading (MIR:R) dichotomously screened for gifted group assignment and Tennessee Comprehensive Assessment Profile: Reading (TCAP:R) dichotomously screened for gifted group assignment

			TCAP:R screened		Total
			0	1	
MIR:R screened	0	Count	272	42	314
		% within MIR:R screened	86.6%	13.4%	100.0%
		% within TCAP:R screened	90.1%	60.0%	84.4%
		% of Total	73.1%	11.3%	84.4%
	1	Count	30	28	58
		% within MIR:R screened	51.7%	48.3%	100.0%
		% within TCAP:R screened	9.9%	40.0%	15.6%
		% of Total	8.1%	7.5%	15.6%
Total		Count	302	70	372
		% within MIR:R screened	81.2%	18.8%	100.0%
		% within TCAP:R screened	100.0%	100.0%	100.0%
		% of Total	81.2%	18.8%	100.0%

Note: *p = n/s, N* = 372, 0 = non-gifted group assignment, 1 = gifted group assignment

- Specificity/True negative (not identified by either test) = 272; 73.1%
- Type II error/False negative (not identified by MIR:R but by TCAP:R) = 42; 11.3%
- Type I error/False positive (identified by MIR:R but not by TCAP:R) = 30; 8.1%
- Sensitivity/True positive (identified by both) = 28; 7.5%

TR:R X TCAP:R

The results of a chi-square test of independence with TR:R (dichotomously screened for gifted-group assignment) by TCAP:R (also screened for gifted-group assignment) shows that these two variables are significantly related, $x^2(1) = 40.21$, $p <$ .000. McNemar test lacks significance ($p > .05$; Cohen, 1988) indicating that the proportion of gifted students screened TR:R and TCAP:R is not statistically significantly different, $p = .07$ (2-tailed; $N = 372$).

Table 19. Chi-square correlation between Teacher Rank: Reading (TR:R) dichotomously screened for gifted group assignment and Tennessee Comprehensive Assessment Profile: Reading (TCAP:R) dichotomously screened for gifted group assignment

Chi-Square Tests	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)
Pearson Chi-Square	40.21[a]	1	.000	
McNemar Test				.07[b]

Note: a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.16. b. Binomial distribution used; $N = 372$

Table 20. Chi-square correlation percentages between Teacher Rank: Reading (TR:R) dichotomously screened for gifted group assignment and Tennessee Comprehensive Assessment Profile: Reading (TCAP:R) dichotomously screened for gifted group assignment

			TCAP:R screened		
			0	1	Total
TR:R screened	0	Count	275	43	318
		% within TR:R screened	86.5%	13.5%	100.0%
		% within TCAP:R screened	91.1%	61.4%	85.5%
		% of Total	73.9%	11.6%	85.5%
	1	Count	27	27	54
		% within TR:R screened	50.0%	50.0%	100.0%
		% within TCAP:R screened	8.9%	38.6%	14.5%
		% of Total	7.3%	7.3%	14.5%
Total		Count	302	70	372
		% within TR:R screened	81.2%	18.8%	100.0%
		% within TCAP:R screened	100.0%	100.0%	100.0%
		% of Total	81.2%	18.8%	100.0%

Note: $p = n/s$, $N = 372$, 0 = non-gifted group assignment, 1 = gifted group assignment

- Specificity/True negative (not identified by either test) = 275; 73.9%
- Type II error/False negative (not identified by TR:R but by TCAP:R) = 43; 11.6%
- Type I error/False positive (identified by TR:R but not by TCAP:R) = 27; 7.3%
- Sensitivity/True positive (identified by both) = 27; 7.3%

Math

MIR:M X TCAP:M

The results of a chi-square test of independence with MIR:M (dichotomously screened for gifted-group assignment) by TCAP:M (also screened for gifted-group assignment) shows that these two variables are significantly related, $x^2(1) = 31.18$, $p <$

.000. McNemar test lacks significance ($p > .05$; Cohen, 1988) indicating that the proportion of gifted students screened MIR:M and TCAP:M is not statistically significantly different, $p = .64$ (2-tailed; n = 372).

Table 21. Chi-square correlation between Monitoring Instructional Response: Math (MIR:M) dichotomously screened for gifted group assignment and Tennessee Comprehensive Assessment Profile: Math (TCAP:M) dichotomously screened for gifted group assignment

Chi-Square Tests	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)
Pearson Chi-Square	31.18[a]	1	.000	
McNemar Test				.64[b]

Note: a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.16. b. Binomial distribution used; *N* = 372

Table 22. Chi-square correlation percentages between Monitoring Instructional Response: Math (MIR:M) dichotomously screened for gifted group assignment and Tennessee Comprehensive Assessment Profile: Math (TCAP:M) dichotomously screened for gifted group assignment

			TCAP:M screened		
			0	1	Total
MIR:M screened	0	Count	274	39	313
		% within MIR:M screened	87.5%	12.5%	100.0%
		% within TCAP:M screened	89.0%	60.9%	84.1%
		% of Total	73.7%	10.5%	84.1%
	1	Count	34	25	59
		% within MIR:M screened	57.6%	42.4%	100.0%
		% within TCAP:M screened	11.0%	39.1%	15.9%
		% of Total	9.1%	6.7%	15.9%
Total		Count	308	64	372
		% within MIR:M screened	82.8%	17.2%	100.0%
		% within TCAP:M screened	100.0%	100.0%	100.0%
		% of Total	82.8%	17.2%	100.0%

Note: *p = n/s*, *N* = 372, 0 = non-gifted group assignment, 1 = gifted group assignment

- Specificity/True negative (not identified by either test) = 274; 73.7%
- Type II error/False negative (not identified by MIR:M but by TCAP:M) = 39; 10.5%
- Type I error/False positive (identified by MIR:M but not by TCAP:M) = 34; 9.1%
- Sensitivity/True positive (identified by both) = 25; 6.7%

TR:M X TCAP:M

The results of a chi-square test of independence with TR:M (dichotomously screened for gifted-group assignment) by TCAP:M (also screened for gifted-group

assignment) shows that these two variables are significantly related, $x^2(1) = 46.07$, $p <$ .000. McNemar test lacks significance ($p > .05$; Cohen, 1988) indicating that the proportion of gifted students screened TR:M and TCAP:M is not statistically significantly different, $p = .32$ (2-tailed; $N = 372$).

Table 23. Chi-square correlation between Teacher Rank: Math (TR:M) dichotomously screened for gifted group assignment and Tennessee Comprehensive Assessment Profile: Math (TCAP:M) dichotomously screened for gifted group assignment

Chi-Square Tests	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)
Pearson Chi-Square	46.07[a]	1	.000	
McNemar Test				.32[b]

Note: a. 0 cells (0.0%) have expected count less than 5. The minimum expected count is 10.16. b. Binomial distribution used; $N = 372$

Table 24. Chi-square correlation percentages between Teacher Rank: Math (TR:M) dichotomously screened for gifted group assignment and Tennessee Comprehensive Assessment Profile: Math (TCAP:M) dichotomously screened for gifted group assignment

			TCAP:M screened		
			0	1	Total
TR:M screened	0	Count	280	37	317
		% within TR:M screened	88.3%	11.7%	100.0%
		% within TCAP:M screened	90.9%	57.8%	85.2%
		% of Total	75.3%	9.9%	85.2%
	1	Count	28	27	55
		% within TR:M screened	50.9%	49.1%	100.0%
		% within TCAP:M screened	9.1%	42.2%	14.8%
		% of Total	7.5%	7.3%	14.8%
Total		Count	308	64	372
		% within TR:M screened	82.8%	17.2%	100.0%
		% within TCAP:M screened	100.0%	100.0%	100.0%
		% of Total	82.8%	17.2%	100.0%

Note: *p = n/s*, *N* = 372, 0 = non-gifted group assignment, 1 = gifted group assignment

- Specificity/True negative (not identified by either test) = 280; 75.3%
- Type II error/False negative (not identified by TR:M but by TCAP:M) = 37; 9.9%
- Type I error/False positive (identified by TR:M but not by TCAP:M) = 28; 7.5%
- Sensitivity/True positive (identified by both) = 27; 7.3%

5b) Interpretation and Discussion

Reading

Using the dichotomous TCAP:R metric as a standard for screening for gifted-group assignment and a late measure, the MIR:R and the TR:R in this sample as early measures show little difference when selecting for gifted screening as evidenced by the

lack of significance in the McNemar tests comparing MIR:R and TCAP:R and TR:R and TCAP:R ($p > .05$ for both MIR:R and TR:R). A chi-square test of independence can be interpreted as indicating that MIR:R (73.1%) and TR:R (73.9%) accurately identified *non-gifted* group assignment (specificity/true negative) as measured by TCAP gifted-group assignment at a rate of about 73%. Similarly, there is little difference in the ability of the early measures to screen for *gifted-group* assignment (sensitivity/true positive); both identified at a rate of about 7% (MIR:R 7.5%; TR:R 7.3%) in this sample. The Type I error rate (false positive) is less than 10% (MIR:R 8.1%; TR:R 7.3%); meaning that less than 10% of those identified by MIR:R or TR:R were ultimately *not* identified by TCAP:R for gifted screening. The Type II error (false negative), those identified *only* by TCAP:R and *not* by MIR:R or TR:R, is at a rate of about 11% (MIR:R 11.3%; TR:R 11.6%) for this sample. In assessing rates of accurate group assignment (gifted into gifted group [true positive, sensitivity] and non-gifted into non-gifted group [true negative, specificity]) the early measures of reading identified correctly at a rate of about 80% (MIR:R 80.6%, TR:R 81.2%).

Math

Using the dichotomous TCAP:M metric as a standard for screening for gifted group assignment and a late measure, the MIR:M and the TR:M in this sample as early measures show little difference in the proportions of the samples when selecting for gifted screening as evidenced by the lack of significance in the McNemar tests ($p > .05$ for both MIR:M and TR:M). MIR:M (73.7%) and TR:M (75.3%) accurately identified

non-gifted group assignment (specificity/true negative) as measured by TCAP *gifted-group* assignment at a rate of about 74%. Similarly, there is little difference in the ability of the early measures to screen for *gifted group* assignment (sensitivity/true positive); both identified at a rate of about 7% (MIR:M 6.7%; TR:M 7.3%) in this sample. The Type I error rate (false positive) is less than 10% (MIR:M 9.1%; TR:M 7.5%); that is, less than 10% of those identified by MIR:R or TR:R were ultimately *not* identified by TCAP:M for *gifted-group* screening. The Type II error (false negative), those identified *only* by TCAP:M and *not* by MIR:M or TR:M, is about 10% (MIR:M 10.5%; TR:M 9.9%) for this sample. In assessing rates of accurate group assignment (gifted into gifted group [true positive, sensitivity] and non-gifted into non-gifted group [true negative, specificity]) the early measures of math identified correctly at a rate of about 80% (MIR:R 80.4%, TR:R 82.6%).

Chapter 5

Conclusions, Significance, Implications

Conclusions and Summary

Conclusions from the Review of Literature

Research in the literature of the gifted education field generates many questions with few obvious answers. The reification of giftedness results in broad, omnibus definitions too convoluted to be of practical use in research, or more narrow definitions that allow for discovery but yet fail to be comprehensive. Both can be negatively viewed. Giftedness, a real phenomenon manifest in certain individuals and characterized by advanced, atypical performance of some nature socially and culturally important, may be identified and developed for benefits that accrue to society or to the individual for reasons of social, artistic, or technical advancement, or only for the egalitarian motivations of equal and appropriate treatment.

Though it seems intuitive to suggest that students who consistently perform toward the upper levels of assessment metrics should be considered as candidates for gifted identification, NCLB requires quantification of practices and instruments used to screen, identify, remediate, or otherwise serve special needs students, a category that in many states subsumes giftedness. It is necessary, then, to specifically re-examine valuations of student performance and teacher accountability when newly extending these practices for above-grade level applications. The language of NCLB compels increased scientific rigor in the quantification of the interventions used to screen, identify, and/or

serve students in RTI settings and of the instrumentation utilized to track student attainment toward end-of-year goals. Such rigor is typified by concise operationalizing of definitions and terms (a practice already known to be problematic in gifted education) and the use of methods adopted and adapted from other social sciences, such as hypothesis testing.

A review of federal and state policies for gifted education was of questionable benefit. Broad definitions prevail at a federal level, but this level lacks both mandates and funding. IDEA fails to include giftedness, but provides frameworks for screening, identification, and programming used for other disabilities that may be adapted for gifted populations. NCLB provides a definition of giftedness and requires progress monitoring of all students, however, the law fails to explicitly parse at a federal level any protocols for gifted students, leaving this to the states. With few concrete provisions made for gifted students it was, then, the dearth of definitions and actionable policies and procedures that led to this study.

At a state level, policies often provide more specific definitions that are limited most frequently, and unsurprisingly, to intelligence and academic achievement. Though many experts in the field advocate broad definitions of giftedness, such restricted definitions were used in this study. Vagaries in definition and policy naturally also lead to a range of gifted identification rates at the state level. Consequently, this study follows recommendations from experts in the field encouraging the use of local norms and a screening cutoff at the 85th percentile to include potentially gifted and high ability

students who might attain at gifted levels if provided with services (Renzulli, 2010; Subotnik, 2010; Adams, 2010). Methods used by states for gifted identification also have great variance. Thus, when documenting student outcomes, student performance on high-stakes testing and on smaller, more frequent curriculum-based measures often generated in RTI settings becomes an important, sometimes overriding, evidentiary base for a wide range of concerns in short- and long-term, even daily decision making.

This study was implemented using three types of instruments, adhering to best practice recommendations by the use of a *multiple criteria method* and *multiple sources* (NACG, 2014) of data. A developing universal screener (MIR) with applications in general education and RTI settings was evaluated as an early screening tool for gifted students. Another early measure derives from a qualitative measure of teacher perception of to-date student performance by a teacher's ranking of students from highest achievement to lowest (TR). These were compared to each other with inter-instrument and intra-domain correlations. Investigation was limited to two academic domains, reading and math, based on required domain testing parameters at the federal level as mandated by NAEP requirements.

Data in this study were collected from third graders enrolled in eight public elementary schools during the 2010-2011 academic school year from a small, rural school district located in the southeast of the United States. The cleaned dataset yielded 372 student cases, 51.3% *female*, 48.7% *male*, and comprising two ethnic categories (collapsed for convenience) *White* (93%) and *non-White* (7%). The data set also includes

results from a teacher ranking of these students collected from the 28 third grade teachers at the schools. No demographic data about the teachers were collected.

As justified in chapters 1, 2, and 3 above, this study focused on the use of early-in-year measures of student performance with a goal of quantifying the predictability of early measures to late measures, and establishing a process for gifted screening when other measures may be unavailable. Early measures were compared to a late-in-year, high-stakes test used by the state of Tennessee (TCAP) as a standard of determination for gifted status; TCAP performance is a heavily-weighted criterion for gifted identification using the Gifted Identification Matrix developed by state personnel as a protocol for identification. Applications of instrumentation to above grade level performance screening were also explored.

Conclusions from the Hypothesis Testing

Question 1: MIR as US for Giftedness

1) Do CBM of reading and math (as measured by the MIR:R and MIR:M) provide sufficient ceiling to serve as screeners for gifted and high ability students (G/HA) in a general education classroom sample?

The first question considers the efficacy of the MIR probes for use in G/HA screening, seeking to validate this untested application. The probes of both domains were shown to have psychometric properties adequate for the purpose. Through an examination of *z*-score conversions and item gradients, results from this sample provide evidence that the MIR probes in both domains have an adequate test ceiling and

sufficiently small item gradients at and above 2 *SDs* to support use of the MIR as a screener for G/HA students. These results support the validity of the MIR for use in gifted screenings and lay the foundation for increased confidence in the results of the subsequent hypothesis testing.

Question 2: Domain Inter-correlations (Reading to Math)

2a) To what degree or extent are the domain-specific MIR (reading and math) scores related to each other for the entire sample; for students in the G/HA group? For the latter analysis, the G/HA group was defined as those scoring at or above the 85th percentile on TCAP Reading and TCAP Math composite scores.

2b) To what degree or extent are the domain-specific TR (reading and math) scale scores related to each other for the entire sample?

2c) To what degree or extent are the domain-specific TCAP (reading and math) scale scores related to each other for the entire sample?

2d) To what degree or extent is the magnitude of the MIR correlations comparable to those of the TCAP correlations for the entire sample?

As explained in the literature review (Chapter 2), researchers have shown that cognitive development in reading and math follows the same developmental pattern from use of small fact-based units with concrete representations to increasing fluency and processing speed by elaborating, synthesizing, and abstracting the smaller fact-based units into more complex procedures (e.g., for reading see Duncan & Seymour, 2000; Ehri & McCormick, 1998; Seymour , Aro, & Erskine, 2003; for math, see Gersten et al., 2005;

Okamoto & Case, 1996). Early performance indicators of phonemic awareness, generally considered a reading skill, strongly predict future performance in *both* reading and math, suggesting that reading and math development is mediated by the same skill set and should thus, in general, display commensurate development. It was hypothesized in this question that the performance measures in each domain (reading and math) for MIR would be significantly correlated and the magnitude of the correlation would be *medium* or higher. In domain-specific, single test comparisons the correlation coefficient between MIR:R scores and the MIR:M scores for the entire sample is significant but *small*. Correlations of MIR:R and MIR:M for those performing above the TCAP 85th percentile in reading and math lack significance and are also *small*. It was further hypothesized that the TR:R and TR:M measures and the TCAP:R and TCAP:M probes would also be significantly correlated with a magnitude of *medium* or higher being coincidental measures on the same population, and, in fact, there is a *large* and significant degree of correlation in both. Researchers have repeatedly found strong inter-correlations between literacy and math competencies reporting coefficients ranging between $r = .40$ and $r = .60$ (e.g., Berg, 2008; Lee, Ng, Ng, & Lim, 2004; Koponen et al., 2007; Schneider, 2009). The inter-domain correlation of both the TR and TCAP are consistent with these findings.

Question 3: Early Measure Inter-correlations (MIR to TR)

3) To what degree or extent are the domain-specific MIR-R and MIR-M correlated with domain specific TR:R and TR:M as a measure of teacher perception for the entire sample?

It was hypothesized in this question that the domain-specific MIR and TR probes as examples of coincidental, early measures would be significantly correlated and the magnitude of the correlation would be *medium* or higher. When comparing the MIR probes to the TR, the correlation in each domain (reading and math) is significant. However, the relation between the two measures is stronger in the reading domain (*medium*) than in math (*small*). Subjective measures, such as of teacher perception, may seem the least trustworthy lacking some of the characteristics associated with extensive psychometric analysis. Important here, however, is the degree of concurrence in two early, domain-specific measures assessing levels of student attainment. This agreement inspires some confidence, more so in reading, in the validity of early measures for improved comparisons to later measures.

Question 4: Predictability; Early Measure (MIR, TR) to Late Measure (TCAP) Correlations and Regressions

4a) To what degree or extent are early-in-year CBM (as measured by MIR: R and MIR:M) correlated with the TCAP as an example of end-of-year measure?

4b) To what degree or extent are TRs of reading and math as examples of early-in-year measures correlated with the TCAP as an example of end-of-year measure?

4c) To what degree or extent can the MIR and TR (in reading and math) collectively predict TCAP scores?

A pattern of stronger correlations in reading compared to those of math held for the relation between the MIR and TCAP and the TR and TCAP; the correlations between

both the domain-specific MIR and TR and the domain-specific TCAP are significant, with a stronger (*large*) correlation in the reading domain than in math (*medium*). *Large* effect sizes were reported for both regressions. The reading R^2 (.43) indicates that the combination of variables accounts for 43% of the variability in the TCAP:R, in math the R^2 (.32) indicates that the combination of variables accounts for 32% of the variability in the TCAP:M. It is also the case that in both regressions the TR is the better predictor of TCAP scores.

Validity accrues to instrumentation in various ways. Face validity derives from the seeming appropriateness of an instrument to measure as it purports; reading skill, for example should be measured by tests requiring reading, math skills by tests with math problems, etc. This, the weakest form of validity, is apparent in all three instruments of this study. Strong correlations between developing instrumentation and other instrumentation that is already known to be valid allows for increased confidence in the validity of both. Strong inter-instrument correlations between MIR and TCAP and TR and TCAP performance measures are expected if each instrument reliably and validly measures reading and math as designed (convergent validity); or, conversely, unintentionally measures the same non-reading or non-math characteristic. The *large* inter-correlations between these instruments in the reading domain suggest that reading performance is actually being measured by each (construct validity), and the similarities in assessment of student attainment by each increases the reliability of all three, as each measures performance outcomes in a similar manner.

As noted, both domain-specific inter-correlations between MIR and TR, and the TCAP are stronger (*large*) in reading when compared to math (*medium*). Though increased confidence in the reliability and validity of instrumentation can derive from *medium* inter-correlations, the comparison of the *large* inter-correlations of the reading domain to the *medium* inter-correlations of the math domain may lead to diminished optimism relative to the psychometric merits of the math instrumentation. Indeed, in a similar study correlating MIR to TCAP, Bell et al. (2015) have shown that results of the TCAP math tests are confounded by a lack of discriminant validity, or the surety that the outcome measures derive solely from the skills ostensibly assessed. Specifically, in this context, Bell et al. found that the TCAP math test as a measure of student performance in math is confounded by an indirect, unintended measure of reading by virtue of the reading skills required to take the test. Thus, the lower (*medium*) math correlations compared to reading correlations (*large*) between MIR, TR, and TCAP, while still helpful in determining the suitability of early measures in predicting later performance, may not necessarily stem from the reliability or validity of the MIR or the TR, but from that of the TCAP:M. It should be noted that the MIR:M was intentionally designed as a non-verbal math assessment for this reason, i.e., to be a "pure" measure of math skills without the confound of reading skill.

Question 5: Rates and Accuracy of Identification

5a) Is there a significant difference in the rate MIR, TR, and TCAP identify G/HA students based on dichotomous gifted group assignment? Group assignment is defined as

attainment at or above the 85th percentile for MIR and TCAP and as the top two ranks for the TR.

5b) Do the MIR, TR, and TCAP identify the same cases of G/HA students based on dichotomous gifted-group assignment? Group assignment is defined as attainment at or above the 85th percentile for MIR and TCAP and as the top two ranks for the TR.

The scores were converted into dichotomous group assignments (gifted, non-gifted) based on cutoff scores for each instrument, and then examined for consistent rates of identification. Through a comparison of the proportion of domain-specific group assignment in MIR, TR, and TCAP, it was established that there is no meaningful difference between the three instruments in either domain of this sample in the proportional rates of gifted identification when used for gifted screening. This finding was reinforced when completing dyadic crosstabs correlations between the instruments (noting again the difference in testing between Questions 3 and 5). It can be induced from the fact that the MIR, TR, and TCAP each identify cases for gifted-group assignment at a comparable rate, that no single instrument tends to over- or under-identify either group or non-group assignments when screening for giftedness. Moreover, the MIR and TR in both domains as early measures were shown to have successful rates of gifted- and non-gifted group assignment when compared to the screening standard of the later TCAP. Group assignment by the MIR and TR was accurate approximately 80% of the time in both reading and math, with levels of Type I and Type II error at around 10% each for each instrument.

The goal of this study was to answer two main questions regarding the psychometric properties of two early measures (MIR and TR) and their predictive characteristics as early measures compared to later measures. Evaluating single question responses has much merit, though it is also important to synthesize the findings.

The use of CBM such as MIR reading and math tests to make early-in-year decisions relative to giftedness is a practice that may have some validity when used in combination with other measures such as the TR. In both reading and math, the MIR in combination with the TR contributes to the ~80% success rate in group assignment when predicting future TCAP performance, with TR as the stronger indicator. The level of agreement in gifted-group assignment between the MIR and TCAP (Question 5) is important in helping to support the validity of the MIR probes as curriculum-based measures to evaluate student progress toward learning goals as assessed by the TCAP scores, and supports the construct validity of the MIR probes as the psychometric properties of TCAP measures have been more extensively evaluated.

The *small* correlation between reading and math domains of the MIR probes may be an area of concern (Question 2). As has been stated, researchers have repeatedly found substantial inter-correlations between literacy and math competencies (*medium* to *large*; e.g., Berg, 2008; Lee et al., 2004; Koponen et al., 2007; Schneider, 2009), indicating that similar cognitive competencies influence performance and development in these two disparate areas of school achievement. Performance outcomes on the domain-specific MIR probes ought to manifest higher magnitudes of correlation as they were taken from

the same population and at the same time. The correlations between the TR and TCAP domain-specific measures, both *large*, are further evidence by comparison, of this limitation.

Findings in this study are not intended to discourage the use of the MIR as a curriculum-based measure. Through analysis of one data point (the first administration of the US), the MIR was shown to have significant but *small* correlations in both domains to the TCAP standards (reading to reading, math to math). Other researchers (Miller, 2012; Hilton-Prillhart, 2011) however, have shown that multiple data points and progress slopes collected through complete administration of the MIR probes evidence more robust, *moderate* correlations to the TCAP and support the use of MIR as a curriculum-based measure.

Overall, early indicators as measured by the MIR in combination with TR, in a manner consistent with the use of multiple measures for screening and identification, can help to inform decisions about group assignment and gifted screening, placement decisions that would eventually be corroborated as correct by future TCAP performance in about 80% of the cases. The results of this hypothesis testing using correlations, regressions, crosstabs, etc., can be helpful in providing evidentiary support for the use of these instruments in terms of their validity and reliability. MIR and TR can be accurate in gifted group assignment when making early-in-year educational decisions for such interventions for G/HA learners as homogeneous grouping, suitability of forms of

acceleration, and/or lesson differentiation in presentation of content, student product, or assessment.

The teacher ranking proved to be an interesting instrument. While the domain-specific rankings have a *large* and significant correlation (Question 2), as early measures, the relation between MIR and TR is inconsistent. TR:R and MIR:R have a significant and *medium* correlation. However, the TR:M and MIR:M correlation is significant, but *small* (Question 3). A pattern of stronger reading correlations holds when examining the relation between TR and TCAP; the reading correlates are significant and *large*, with correlations in math being significant and *medium* (Question 4). Both are more robust than the MIR/TCAP correlates, however, thus it is not surprising that the TR contributes the larger share of the combined effect found when examining the relation between early measures and late measures and gifted group assignment (Question 5).

Significance

MIR as a Gifted Screener

The strength of the MIR may be its efficacy when used as universal screener for giftedness. The Association for the Gifted, a division of Council for Exceptional Children (CEC-TAG) now recommends that the "RTI model be expanded in its implementation to include the needs of gifted children" (CEC, the Association for Gifted, 2009, p. 1) and recognizes the potential of adapting the RTI framework for gifted learners. Establishing the reliability, validity, and generalizability of CBM instruments such as MIR is essential before the intention to identify and serve gifted students becomes actionable. New

applications of existing processes for screening and identifying special needs students such as G/HA learners and instrumentation that has yet to be tested in these applications seem to require explicit validation by the language of NCLB. Though developed well before formal implementation of RTI, CBMs (Deno, 1985) are now an integral part of the RTI framework (established by IDEA) and included as a part of NCLB legislation, meaning that quantification of MIR as a CBM suitable for gifted screening is requisite in terms both specific to the MIR probes and to CBM in general. CBM can be used as a method of *screening* (Ardoin et al., 2004), *identifying*, or *monitoring* (Fuchs & Fuchs, 1999; Fuchs, Fuchs, Hamlett, Walz, & Germann, 1993; Hosp & Hosp, 2003; Stecker & Fuchs, 2000). This study contributes to the ongoing investigation of the validity of CBM used for the *screening* process as advocated by such authors as Burns, Jacob, and Wagner (2008) by providing evidence that the MIR as an example of a CBM can be acceptable and useful as a screening tool for G/HA students.

Use of TR in gifted Screening

Best practice in gifted identification includes the use of nominations from teachers, parents, or peers (Gallagher, 1994; Sternberg, 1998). A seminal study with negative implications of teacher perceptions of giftedness and teacher nomination written by Pegnato and Birch (1959) still retains some resonance despite being soundly refuted by Hoge and Cudmore (1986) and Gagné (1994). In this study, significant and *large* coefficient effect sizes were reported for regressions in both domains (accounting for 43% of the variability in the TCAP:R and 32% of the variability in the TCAP:M).

Additionally, the predictive influence of the TR and the level of agreement (approximately 80% success rate) in group assignment in both reading and math (Question 5) is relevant because, as suggested by Gagné (1994), teachers actually do seem to apprehend levels of student competency fairly well. Peters and Gentry (2012) state that teacher ratings of giftedness are more successful when teachers are provided with explicit behaviors to observe, and that without concrete parameters, the resulting teacher nominations did not appear to be especially accurate. It would be specious, however, to suggest these findings might be contradictory to this tenet, remembering that though the hypothesis testing with TR generally demonstrated significant results and robust correlations, the data may have little to do with teacher evaluations of gifted characteristics. Only five states require all teachers to receive pre-service training in gifted and talented education (Delcourt, Cornell, & Goldberg, 2007), and such is not the case in Tennessee. In this sample, teachers' understanding of giftedness is very much in question.

Utility of Early Measures for Decision Making

The combination effect of MIR and TR in predicting TCAP performance supports the credibility of early measures to evaluate student progress toward learning goals as assessed by the TCAP scores and as indicative of future TCAP performance. Performance indicators measured across time can be confounded by many variables (e.g., maturity, changes in interest or motivation, etc.) creating difficulties in comparisons

between early and later measures, so interpretation of these comparisons should be cautious.

Differences between domain-specific correlations may be attributable to mechanisms within the instrumentation; readings probes may be better at assessing reading skills than math probes are at assessing math skills, for example. It may also be the case that the discrepancies in the domain-specific outcomes could result from actual discrepancies in student performance, or perhaps reflect enhanced teacher abilities in teaching and assessing reading when compared to their abilities in math. However, in comparing early and late measures, the strength of the contribution of the TR in predicting TCAP performance may more likely indicate that teachers, who presumably have a deeper understanding of the end-of-year testing goals as assessed by TCAP, may simply be better able to predict eventual student TCAP performance. If this is the case, early-in-year teacher rankings may be more a function of anticipating end-of-year test performance than a function of demonstrated student ability without regard to high-stakes testing. Notwithstanding, the use of these early measures seems supported by the consistent rates of gifted identification when comparing identification rates of each instrument, *large* effect sizes of predictability when early measures are combined to predict performance on later measures, and accuracy of gifted-group assignment.

Implications

Importantly, it must be stated that the testing results cannot necessarily engender many conclusions about the MIR or TR in terms of gifted education other than to

determine that students who perform well above peers on one instrument, also seem to perform well above peers on the other instruments. It does not, and should not, follow that because the instruments *can* be used for G/HA screening, or are predictive of future performance at gifted levels, the accuracy in gifted- and non-gifted group assignment is uniquely attributable to mechanisms of the instrumentation as intentionally designed for gifted identification, or illustrative of teacher knowledge of or experience working with gifted children. Specifically, these results indicate only that teachers are relatively good at predicting future TCAP performance, but this is at all levels of attainment, which happens to include gifted levels of attainment, and is insufficient for the conclusion that the teacher knows anything about the characteristics of giftedness or gifted learners. Rather, the teacher may simply be able to intuit, at all levels, future student performance on metrics with which they are familiar. This view is consistent with research findings that many states do not require teaching interns to have coursework pertinent to gifted education.

If the goal is a better understanding of the use of early-in-year teacher rankings as part of a process of gifted identification, it would be important to remove this confound as a variable to create a more stable, and thus accurate, testing condition. The use of a different comprehensive late-in-year measure of reading and math, one with which the teacher participants have little familiarity, would help ameliorate this effect. This will occur naturally as state testing adapts to comply with recent changes in curricula and developments in standards evolving from educational reforms. Another possibility would

be an analysis of gifted performance comparing early performance on valid CBMs to CBM data collected at or toward year's end, an option that would simultaneously increase confidence in the use of CBMs as screening instrumentation if medium to large correlations were found.

However, the most important concern centers on early identification of giftedness. Gifted students can be unengaged in the classroom leading to behavior problems (Winner, 1997), unmotivated to learn (Blaas, 2014), and at an increased risk for dropping out (Neihard, Reis, Robinson, & Moon, 2002) when their learning needs are not met within the classroom environment. Pfeiffer and Stocking (2000) go so far as to claim that, "Gifted children and youth possess a set of personality characteristics that make them uniquely vulnerable" (p. 1). As with all other special education identifications, early identification of gifted students is a crucial first step toward serving this population. The importance of early universal screening has been well established for at-risk and special needs students. Extending this concept to include possible giftedness requires metrics specifically evaluated for this purpose. To this end, the MIR screeners were shown to be highly effective. Best practice in screening indicates that the screener administered should be the most appropriate universal screener for the function it serves (NCRTI, 2007). This study helps validate the use of the MIR probes for gifted screening by establishing appropriate psychometric qualities for the purpose of gifted screening. Gersten et al. (2009) and the NCRTI (2010) recommend screening and benchmarking occur at least twice annually in grades K-8. With three universal screeners in the MIR

probe series, the MIR tests can accommodate this best practice recommendation through grade 5.

In the absence of "gold-standard" metrics, or in districts or states that allow for gifted identification but lack cogent policies and procedures for gifted identification, teacher nomination may be the only path for gifted student identification. Whether seeking a formal identification process or simply motivated to meet individual student needs within the classroom environment, clearly, the role of the teacher as a student advocate becomes a foremost consideration. Robust correlations, as evidenced by this study, between early measures of teacher perception and end-of-year student outcomes indicate that teachers have a clear understanding of student ability relative to learning goals. Ensuring that this understanding is more related to a deep knowledge of individual learning styles, student profiles, and curricular goals and objectives, rather than being reflective of a teacher's familiarity with high-stakes tests would increase the trustworthiness of measures of teacher perception, ultimately bringing benefits to all students.

Future Research

The role of the teacher in early gifted screening has been shown to be highly relevant. Findings of this research, consequently, would be strengthened by more demographic information of teacher characteristics. Years of service; specific training in the characteristics and identification of gifted students; measures of teacher understanding of state-, district-, or school-level gifted identification protocols and

programming options; and documentation of efforts toward or interest in accommodating and differentiating instruction for gifted learners are all relevant demographic data points to consider when examining teacher nomination of students to gifted programs. Similar research using the model established here but including more teacher qualifiers could be very informative. Establishing a clear link between teacher experience and (specifically) early gifted identification could add to the already compelling case for the inclusion of more extensive coursework in gifted education for pre-service teachers that will provide strategies for teaching and assessing the gifted children who will be present in the classroom. Understanding the teacher characteristics that lead to effective, accurate gifted screening and identification proceeding from thoughtful intent, as opposed to being merely a facet of overall, class-wide ability estimation, will also lead to a deeper understanding of needs for curricular content leading to improved coursework for teacher preparation.

The MIR probes should be re-examined for a stronger inter-domain correlation. Apparently developed independently, re-evaluation of the testing content to improve them as a set will increase their efficacy and utility. It would first be important to ascertain how this single, data-point correlation might compare to an inter-domain correlation of all data-points from a full administration of the MIR probes. The model of this research could also be applied to data collected from other grades to elaborate or improve these findings.

Data from on-grade level MIR probes were here analyzed from on-grade level students. Best practice as recommended by national organizations such as NACG and researchers such as Subotnik and Thompson (2010) is to allow gifted students access to out-of-grade level testing. Test authors should develop performance-level metrics associated with universal screening scores to facilitate other aspects of standardized testing protocols, such as points of entry, when to administer below- or above-grade level tests, etc. The use of the MIR first universal screener has been shown to be effective in gifted *screening*; this may now necessitate the development of protocols for gifted *identification*, or the use of the probes as ongoing CBM for *monitoring* of gifted students' progress in RTI settings.

Finally, the preponderance of stronger correlations in reading compared to math, and found between all inter-instrument correlations, is noteworthy and interesting. Per Bell et al. (2015) one reason is likely that the end-of-year group achievement test requires some reading while the MIR:M does not. Other reasons may include less elementary teacher knowledge about math than reading; this topic warrants further research.

Limitations

As stated in Chapter 1, these findings and interpretations are not intended to be transferred or generalized to other instrumentation, grades, schools, or states. This investigation was conducted with data collected from specific populations and the conclusions remain pertinent only to this environment. Extension of these findings to other settings, investigations of this instrumentation in other contexts, similar

explorations on different populations or with additional demographic information would elaborate, or perhaps even contradict, the findings presented here. These findings should not be expected in other situations, and similar investigations would require explicit, purposeful intent.

List of References

Ackerman, P. T., & Dykman, R. A. (1995). Reading-disabled students with and without comorbid arithmetic disability. *Developmental Neuropsychology, 11*(3), 351-371. doi: http://dx.doi.org/10.1080/87565649509540625

Adams, C. (2009). Myth 14: Waiting for Santa Claus. Detail Only Available / Part of a special issue: *Demythologizing Gifted Education Gifted Child Quarterly*, 53 (4), 272-273 doi: 10.1177/0016986209346942.

Adelson, J. L., McCoach, D. B., & Gavin, M. K. (2012). Examining the effects of gifted programming in mathematics and reading using the ECLS-K. *Gifted Child Quarterly, 56*(1), 25-39. doi: 10.1177/0016986211431487.

Advantage Learning Systems. (1997). *STAR Reading*. Wisconsin Rapids, WI: ALS, Advantage Learning Systems.

Advantage Learning Systems. (2002). *STAR Math*. Wisconsin Rapids, WI: ALS, Advantage Learning Systems.

American Institute for Research, National Center for Response to Intervention. (2013). Retrieved from http://www.rti4success.org/

Anderson, J. R., Reder, L. M., & Lebiere, C. (1996). Working memory: Activation limitations on retrieval. *Cognitive Psychology*, 30, 221–256.

Antell, S. E., & Keating, D. P. (1983). Perception of numerical invariance in neonates. *Child Development*, 54, 695–701.

Ardoin, S. P., Witt, J. C., Suldo, S. M., Connell, J. E., Koenig, J. L., Resetar, J. L., . . . Williams, K. L. (2004). Examining the incremental benefits of administering a

maze and three versus one curriculum-based measurement reading probes when conducting universal screening. *School Psychology Review, 33*(2), 218-233.

Arvedson, P. J. (2002). Young children with specific language impairment and their numerical cognition. *Journal of Speech, Language, and Hearing Research*, 45, 970-982.

Ash, C., & Huebner, E. S. (1998). Life satisfaction of gifted middle-school children. *School Psychology Quarterly, 13*(4), 310-321.

Askew, M., & Williams, D. (1995). *Recent Research in Mathematics Education 5-16.* London, UK: Her Majesty's Stationery Office, 1995.

Assouline, S. G., & Whiteman, C. S. (2011). Twice-exceptionality: Implications for school psychologists in the post–idea 2004 era. *Journal of Applied School Psychology*, 27(4), 380-402, doi: 10.1080/15377903.2011.616576

Bain, S. K., & Bell, S. M. (2004). Social self-concept, social-attributions, and peer relationships in fourth, fifth, and sixth graders who are gifted compared to high achievers. *Gifted Child Quarterly, 48*(3), 167.

Baker, B. D., & Friedman-Nimz, R. (2003). Gifted children, vertical equity, and state school finance policies and practices. *Journal of Education Finance, 28*(4), 523-556.

Barton, J. M., & Starnes, W. T. (1989). Identifying distinguishing characteristics of gifted and talented/learning disabled students. *Roeper Review, 12*(1), 23-29.

Barrouillet, P., & Fayol, M. (1998). From algorithmic computing to direct retrieval: Evidence from number and alphabetic arithmetic in children and adults. *Memory & Cognition*, 26, 355–368. doi:10.3758/BF03201146.

Baum, S. (2004). *Twice-exceptional and special populations of gifted students*. Thousand Oaks, CA: Corwin Press.

Baum, S. (1994). Meeting the needs of gifted/learning disabled students: How far have we come? *Journal of Secondary Gifted Education, 5*(3), 6-22.

Baum, S., Owens, S. V., & Dixon, J. (1991). *To be gifted and learning disabled*. Mansfield Center, CT: Creative Learning Press.

Bernal, E. M. (2002). Three ways to achieve a more equitable representation of culturally and linguistically different students in gt programs. *Roeper Review, 24*(2), 82-88.

Bell, S.M., Hilton-Prillhart, A., McCallum, R.S., & Hopkins, M. (2010) *Monitoring Instructional Responsiveness: Reading (MIR:R).* Unpublished test, Department of Educational Psychology and Counseling and Department of Theory and Practice in Teacher Education, University of Tennessee, Knoxville, TN.

Bell, S. M., Hilton-Prillhart, A. N., Hopkins, M. B., & McCallum, R. S. (2012). *Monitoring Instructional Responsiveness: Reading (MIR:R).* Unpublished test. University of Tennessee.

Bell, S.M., Taylor, E., McCallum, R.S., Coles, J., & Hays, E. (2015). Comparing prospective twice-exceptional students to high-performing peers on high-stakes tests of achievement. *Journal for Education of the Gifted*.

Berg, D. H. (2008). Working memory and arithmetic calculation in children: The contributory roles of processing speed, short term memory, and reading. *Journal of Experimental Child Psychology*, 99, 288–308.

Bianco, M. (2010). Strength-based RTI: Conceptualizing a multi-tiered system for developing gifted potential. *Theory Into Practice*, *49*(4), 323-330. doi:10.1080/00405841.2010.510763

Bianco, M. (2005). The effects of disability labels on special education and general education teachers' referrals for gifted programs. *Learning Disability Quarterly, 28*, 285-293.

Bijeljac-Babic, R., Bertoncini, J., & Mehler, J. (1993). How do four-day-old infants categorize multisyllabic utterances? *Developmental Psychology*, 29, 711–721.

Bjork, R. A., & Druckman, D. (1994). Editors. *Learning, Remembering, Believing: Enhancing Human Performance.* Washington, DC: National Research Council, 1994.

Blaas, S. (2014). The relationship between social-emotional difficulties and underachievement of gifted students. *Australian Journal of Guidance and Counselling*, Cambridge University Press on behalf of Australian Academic Press Pty Ltd. 24(2) p. 243–255 DOI:10.1017/jgc.2014.1

Bogdan, R. C., & Biklen, S. K. (2007). *Qualitative research for education: An introduction to theory and methods* (5th ed.). New York, NY: Pearson.

Borland, J. H. (1978). Teacher identification of the gifted: A new look. *Journal for the Education of the Gifted*, 2, 22–32.

Borland, J. H. (1989). *Planning and Implementing Programs for the Gifted*. Teachers College Press, New York. NY USA.

Borland, J. H. (1990). Postpositivist inquiry: Implications of the "New Philosophy of Science" for the field of the education of the gifted. *Gifted Child Quarterly 34*(4) 161-167.

Bracken, B. A. (1998). *Examiner's Manual: Bracken Basic Concept Scale - Revised.* San Antonio, TX: The Psychological Corporation.

Brown, E. F., & Garland, R. B. (2015). Reflections on policy in gifted education: James J. Gallagher. *Journal for the Education of the Gifted 38*(1) 90–96, doi:10.1177/0162353214565558.

Brown, E., Avery, L., Van Tassel-Baska, J., Worley, B. B., II, & Stambaugh, T. (2006). A five-state analysis of gifted education policies. *Roeper Review, 29*(1), 11-23.

Bull, R., & Scerif, G. (2001). Executive functioning as a predictor of children's mathematics ability: Inhibition, switching and working memory. *Developmental Neuropsychology 19*(3) 273–93, http://dx.doi.org/10.1207/S15326942DN1903_3.

Bull, R., & Johnston, R. S. (1997). Children's arithmetical difficulties: Contributions from processing speed, item identification, and short-term memory. *Journal of Experimental Child Psychology, 65,* 1–24. doi:10.1006/jecp.1996.2358.

Burns, M. K., Jacob, S., & Wagner, A. R. (2008). Ethical and legal issues associated with using response-to-intervention to assess learning disabilities. *Journal of School Psychology*, 46, 263–279.

Christ, T. W. (2014). Scientific-based research and randomized controlled trials, the "gold" standard? Alternative paradigms and mixed methodologies. *Qualitative Inquiry*, *20*(1), 72-80.

Christ, T., & Silberglitt, B. (2007). Estimates of the standard error of measurement for curriculum-based measures of oral reading fluency. *School Psychology Review*, *36*(1), 130-146.

Cline, S., & Schwartz, D. (1999). *Diverse populations of gifted children*. Englewood Cliffs: NJ: Merrill/Prentice-Hall.

Cohen, L., Manion, L., & Morrison, K. (2011). *Research methods in education (7th ed.)*. London, England: Routledge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Coles, J. T. (2014). Predicting high-stakes tests of math achievement using a group-administered RTI instrument: Validating skills measured by the Monitoring Instructional Responsiveness: Math. Unpublished dissertation.

Compton, D. L. (2003). Modeling the relationship between growth in rapid naming speed and growth in decoding skill in first grade children. *Journal of Educational Psychology*, 95, 225–239.

Council for Exceptional Children Gifted and Talented Division (CEC) Arlington, VA Available online: www.cec.sped.org/gifted

Council for Exceptional Children. (2007). *Position on Response to Intervention (RTI): The unique role of special education and special educators.* Retrieved from http://www.cec.sped.org/AM/Template.cfm?SectionDHome&TemplateD/CM/ContentDisplay.cfm&ContentIDD11769.

Council for Exceptional Children, The Association for Gifted. (2009, April). *Response to intervention for giftedness: A position paper.* Presentation at the Council for Exceptional Children Annual Conference, Seattle, WA.

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches (4th ed.).* Los Angeles, CA: Sage.

Daves, D., & Walker, D. W. (2012). RTI: Court and case law—Confusion by design. *Learning Disabilities Quarterly*, *35*(2), 68–71.

Davidson Institute for Talent Development Available online: www.davidsoninstitute.org

De Brauwer, J., Verguts, T., & Fias, W. (2006). The representation of multiplication facts: Developmental changes in the problem size, five, and tie effects. *Journal of Experimental Child Psychology*, 94, 43–56. doi:10.1016/j.jecp.2005.11.004.

Dehaene, S., Piazza, M., Pinel, P., & Cohen, L. (2003). Three parietal circuits for number processing. *Cognitive Neuropsychology*, 20, 487–506.

de Jong, P. F., & van der Leij, A. (1999). Specific contributions of phonological abilities to early reading acquisition: Results from a Dutch latent variable longitudinal study. *Journal of Educational Psychology*, 91, 450–476.

Delcourt, M. A. B., Cornell, D. G., & Goldberg, M. D. (2007). Cognitive and affective learning outcomes of gifted elementary school students. *Gifted Child Quarterly, 51*(4), 359-381.

Deno, S. L. (2003). Developments in Curriculum-Based Measurement. *The Journal of Special Education, 37*(3), 184–192. doi:10.1177/00224669030370030801.

Deno, S. (1992). The nature and development of curriculum-based measurement. *Preventing School Failure, 36*(2), 5–10.

Deno, S L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children*, *52*(3), 219–32.

Duke University Talent Identification Program Available online: http://www.tip.duke.edu

Duncan, L. G., & Seymour, P. H. K. (2000). Socio-economic differences in foundation-level literacy. *British Journal of Psychology,* 91, 145.

Durbin, J., & Watson, G. S. (1950). Testing for serial correlation in least squares regression, I. *Biometrika, 37(*3–4), 409–428. doi:10.1093/biomet/37.3-4.409. JSTOR 2332391.

Durbin, J., & Watson, G. S. (1951). "Testing for serial correlation in least squares regression, II". *Biometrika* 38(*1–2*) 159–179. doi:10.1093/biomet/38.1-2.159. JSTOR 2332325.

Ehri, L. C., & McCormick, S. (1998). Phases of word learning: Implications for instruction with delayed and disabled readers. *Reading & Writing Quarterly: Overcoming Learning Difficulties, 14*(2), 135-163. doi:http://dx.doi.org/10.1080/1057356980140202.

Elementary and Secondary Education Act (ESEA) Public Law PL 107-110, the No Child Left Behind Act of 2001.

Elliott, S. N., Busse, R. T., & Gresham, F. M. (1993). Behavior rating scales: Issues of use and development. *School Psychology Review*, 22, 313–321.

Elliot, S. N., Gresham, F. M., Frank, J. L. & Beddow, P. A. (2008). Intervention validity of social behavior rating scales. *Assessment for Effective Intervention, 34*(1), 15-24. doi: 10.1177/1534508408314111.

Farkas, S., & Duffett, A. (2008). *High-achieving students in the era of NCLB: Results from a national teacher survey*. Washington, DC: Thomas B. Fordham Institute.

Feigenson, L., Carey, S., & Spelke, E. (2002). Infants' discrimination of number vs continuous extent. *Cognitive Psychology*, 44, 33–66.

Fazio, B. B. (1996). Mathematical abilities of children with specific language impairment: A 2-year follow-up. *Journal of Speech, Language, and Hearing Research,* 39, 1–10.

Feldhusen, J. F., & Heller, K. A. (Ed.). (1986). Introduction. In K.S. Heller & J. F. Feldhusen (Eds.), *Identifying and nurturing the gifted: An international perspective.* (p. 19–31). Toronto, Canada: Hans Huber.

Ford, D. Y., & Grantham, T. C. (2003). Providing access for culturally diverse gifted students: From deficit to dynamic thinking. *Theory Into Practice*, 42, 217–225.

Ford, D. Y., Grantham T. C., & Whiting, G. W. (2008). Culturally and linguistically diverse students in gifted education: Recruitment and retention issues. *Exceptional Children*, 74, 289–306.

Frasier, M. M., Passow, A. H., National Research Center on the Gifted & Talented, S. C. T. (1994). *Towards a New Paradigm for Identifying Talent Potential. Research Monograph 94112*.

Fuchs, L. S., Seethaler, P.M., Fuchs, D. & Hamlett, C. L. (2008). Using curriculum-based measurement to identify the 2% population. *Journal of Disability Policy Studies*, *19*(153).

Fuchs, L. S., & Fuchs, D. (1999). Monitoring student progress toward the development of reading competence: A review of three forms of classroom-based assessment. *School Psychology Review, 28*(4), 659.

Fuchs, L. S., Hamlett, C. L., & Fuchs, D. (1999). *Monitoring basic skills progress basic math manual*. Austin, TX: Pro-ed.

Fuchs, L. S., Fuchs, D., Hamlett, C. L., Walz, L., & Germann, G. (1993). Formative evaluation of academic progress: How much growth can we expect? *School Psychology Review, 22*(1), 27.

Gagné, F. (1994). Are teachers really poor talent detectors? Comments on Pegnato and Birch's (1959) study of the effectiveness and efficiency of various identification techniques. *Gifted Child Quarterly, 38*(3), 124-126.

Gallagher, J. J. (2005, May 25). Commentary: National security and educational excellence. *Education Week, 24*(38), 32–33, 40.

Gallagher, J. J. (1994). A retrospective view: The Javits Program. *Gift Child Quarterly, 38*(2), 95-96.

Gallagher, S., Smith, S., & Merrotsy, P. (2011). Teachers' perceptions of the socioemotional development of intellectually gifted primary aged students and their attitudes towards ability grouping and acceleration. *Gifted and Talented International*, 26(1); and 26(2).

Geary, D. C. (1993). Mathematical disabilities: Cognitive, neuropsychological, and genetic components. *Psychological Bulletin*, 114, 345–362.

Georgiou, G. K., Papadopoulos, T. C., Fella, A., & Parrila, R. (2012). Rapid naming speed components and reading development in a consistent orthography. *Journal of Experimental Child Psychology, 112*(1), 1-17. doi: 10.1016/j.jecp.2011.11.006.

Gersten, R., Compton, D., Connor, C.M., Dimino, J., Santoro, L., Linan-Thompson, S., & Tilly, W.D. (2008). *Assisting students struggling with reading: Response to Intervention and multi-tier intervention for reading in the primary grades. A practice guide.* (NCEE 2009-4045). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S.

Department of Education. Retrieved from http://ies.ed.gov/ncee/wwc/publications/practiceguides/.

Gersten, R., Jordan, N. C., & Flojo, J. R. (2005). Early identification and interventions for students with mathematics difficulties. *Journal of Learning Disabilities*, 38, 293–304.

Gersten, R. e. a. (2009). *Assisting students struggling with reading: response to intervention and multi-tier intervention in the primary grades. washington, d. c.: national center for education evaluation and regional assistance*, Institute of Education Sciences.

Gilger, J. W. & Hynd, G. W. (2008). Neurodevelopmental variation as a framework for thinking about the twice exceptional. *Roeper Review*, *30*(4), 214-228.

Gilmore, C., Attridge, N., Clayton, S., Cragg, L., Johnson, S., Marlow, N., et al. (2013). Individual differences in inhibitory control, not non-verbal number acuity, correlate with mathematics achievement. PLoS One 8(6): e67374, http://dx.doi.org/10.1371/journal.pone.0067374.

Glaser, B. G., & Strauss. A. L. (1967). *The discovery of grounded theory: Strategies for qualitative researc*h. New York: Aldine de Gruyter.

Good, R. H. & Kaminski, R. A. (Eds.) (2002). *Dynamic Indicators of Basic Early Literacy Skills* (6th ed.). Eugene, OR: Institute for the Development of Educational Achievement.

Hecht, S. A., Torgesen, J. K., Wagner, R. K., & Rashotte, C. A. (2001). The relations between phonological abilities and emerging individual differences in mathematical computation skills: A longitudinal study from second to fifth grades. *Journal of Experimental Child Psychology*, 79, 192–227. doi:10.1006/jecp.2000.2586.

Heller, K. A., & Feldhusen, J. F., eds. (1986). *Identifying and Nurturing the Gifted: An International Perspective*. 6th World Conference on Gifted and Talented Children. Hans Huber Publishers Toronto, Ontario. ISBN-0-920887-11-2.

Hilton-Prillhart, A. N. (2011). *Validation of the Monitoring Academic Progress: Reading (MAP: R): Development and investigation of a group-administered comprehension-based tool for RTI.* Unpublished doctoral dissertation, University of Tennessee, Knoxville.

Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C. § 1400 et seq. Stat. (2004).

Hinkle DE, Wiersma W, Jurs SG (1988) *Applied statistics for the behavioral sciences.* 2nd ed. Boston: Houghton Mifflin Company.

Hodge, K. A., & Kemp, C. R. (2006). Recognition of giftedness in the early years of school: perspectives of teachers, parents, and children. *Journal for the Education of the Gifted, 30*(2), 164-204.

Hoge, R. D., & Cudmore, L. (1986). The use of teacher-judgment measures in the identification of gifted pupils. *Teaching and Teacher Education, 2*(2), 181-196.

Hopkins, M. (2011). *A validation of the Monitoring Academic Progress Mathematics: An experimental multidimensional group administered curriculum-based measure of mathematics fluency and problem solving.* Unpublished Dissertation.

Hopkins, M., McCallum, R.S., Bell, S.M., & Prillhart, A. (2011). *Monitoring Academic Progress: Math*. Unpublished Manual.

Hosp, M. K., & Hosp, J. L. (2003). Curriculum-based measurement for reading, spelling, and math: How to do it and why. *Preventing School Failure, 48*(1), 10-17.

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.

Huntley-Fenner, G., & Cannon, E. (2000). Preschoolers' magnitude comparisons are mediated by a preverbal analog mechanism. *Psychological Science*, 11, 147–152.

Individuals with Disabilities Education Improvement Act of 2004, 20 U.S.C. § 1400 et seq. Stat. (2004).

Jenkins, J. R., Schiller, E., Blackorby, J., Thayer, S. K., & Tilly, W. D. (2013). Responsiveness to intervention in reading: Architecture and practices. *Learning Disability Quarterly, 36*(1), 36-46. doi: 10.1177/0731948712464963.

Johnsen, S., & National Association for Gifted Children, S. P. M. N. (2004). *Identifying Gifted Students: A Practical Guide*: National Association for Gifted Children.

Karnes, F. A., & Stephens, K. R. (2000). State definitions for the gifted and talented revisited. *Exceptional Children, 66*(2), 219-238.

Kaufman, A. S., & Harrison, P. L. (1986). Intelligence tests and gifted assessment: What are the positives? *Roeper Review, 8*(3), 154-159.

Kavale, K. A. & Spaulding, L. S. (2008). Is response-to-intervention good policy for specific learning disability? *Faculty Publications and Presentations*. Paper 119. http://digitalcommons.liberty.edu/educ_fac_pubs/119

Kieffer, K. M., Reese, R. J., & Vacha-Haase, T. (2010). Reliability generalization (RG) methods in the context of giftedness research. In B. Thompson & R. F. Subotnik (Eds.), *Methodologies for conducting research on giftedness* (p. 89–111). Washington, DC: American Psychological Association. doi:10.1037/12079-005.

Krajewski, K., & Schneider, W. (2009). Exploring the impact of phonological awareness, visual–spatial working memory, and preschool quantity–number competencies on mathematics achievement in elementary school: Findings from a 3-year longitudinal study. *Journal of Experimental Child Psychology*, 103, 516–531.

Koponen, T., Aunola, K., Ahonen, T., & Nurmi, J. E. (2007). Cognitive predictors of single-digit and procedural calculation and their covariation with reading skill. *Journal of Experimental Child Psychology*, 97, 220–241.

Koponen, T., Salmi, P., Eklund, K., & Aro, T. (2013). Counting and RAN: Predictors of arithmetic calculation and reading fluency. *Journal of Educational Psychology*, *105*(1), 162–175. doi: 10.1037/a0029285.

Kuhn, T. S. (1966). *The Structure of Scientific Revolutions*, 3rd ed. University of Chicago Press. ISBN 0226458121.

LaBerge, D., & Samuels, S. J. (1974). Toward a theory of automatic information processing in reading. *Cognitive Psychology*, *6*, 293–323. doi:10.1016/0010-0285(74)90015-2

Landerl, K., & Moll, K. (2010). Comorbidity of learning disorders: Prevalence and familial transmission. *Journal of Child Psychology & Psychiatry, 51*(3), 287-294. doi: 10.1111/j.1469-7610.2009.02164.x.

Landerl, K., & Wimmer, H. (2008). Development of word reading fluency and spelling in a consistent orthography: An 8-year follow-up. *Journal of Educational Psychology, 100*(1), 150-161. doi: 10.1007/BF01026945

Larson, N. (2004). *Saxon Math*. Norman, OK: Saxon Publishers, Inc.

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. doi: 10.1177/0016986207311152.

Lemaire, P., & Siegler, R. S. (1995). Four aspects of strategic change: Contributions to children's learning of multiplication. *Journal of Experimental Psychology: General*, 124, 83–97. doi:10.1037/0096-3445.124.1.83.

Leppänen, U. (2006). Development of literacy in kindergarten and primary school. J*yväskylä Studies in Education, Psychology and Social Research*, 289.

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, CA: Sage.

Long, H. (2014). An empirical review of research methodologies and methods in creativity studies (2003–2012). *Creativity Research Journal*, 26:4, 427-438, doi: 10.1080/10400419.2014.961781.

Maker, C. J. (1996). Identification of gifted minority students: A national problem, needed changes and a promising solution. *Gifted Child Quarterly, 40*(1), 41-50.

Margolis, H. (2012). Response to intervention: RTI's linchpins. *Reading Psychology*, 33:8–10 doi: 10.1080/02702711.2011.630600.

Marland, S. (1972). *Education of the gifted and talented: Report to the Congress of the United States by the U.S. Commissioner of Education*. Washington, DC: U.S. Government Printing House.

Maslow, A. H. (1971). *The farther reaches of human nature*. New York: Viking.

Maslow, A. H. (1964). *Religions, values, and peak-experiences*. Columbus: The Ohio State Press.

McCallum, R.S., Hopkins, M. Bell, S.M., & Hilton-Prillhart, A. (2011). *Monitoring Instructional Responsiveness: Math (MIR:M).* Unpublished test, Department of Educational Psychology and Counseling and Department of Theory and Practice in Teacher Education, University of Tennessee, Knoxville, TN.

Meng, Rosenthal, & Rubin (1992) Comparing correlated correlation coefficients. *Psychological Bulletin*, 111, 172-175.

Merriam, S. B. (2002). *Qualitative research in practice: Examples for discussion and analysis.* New York, NY: John Wiley Sons.

Miller, K. C. (2012). *Predictive validation of the Monitoring Instructional Responsiveness: Reading (MIR:R): Investigation of a group-administered, comprehension-based tool for RTI implementation.* Unpublished Dissertation.

Miller, K. C., Bell, S. M., & McCallum, R.S. (2015). Using reading rate and comprehension CBM to predict high-stakes achievement. *Journal of Psychoeducational Assessment.*

Mullan, B., Todd, J., Chatzisarantis, N., & Hagger, M. S. (2014). Experimental methods in health psychology in Australia: Implications for applied research. *Australian Psychologist* 49, 104–109. doi:10.1111/ap.12046.

National Association for Gifted Children. (2009) State of the states in gifted education: National policy and practice data 2008-2009 (Report by the Council of State Directors of Programs for the Gifted & the National Association for Gifted Children).Washington, DC: Author.

National Association for Gifted Children. (2014) State of the states in gifted education: National policy and practice data 2013-2014 (Report by the Council of State Directors of Programs for the Gifted & the National Association for Gifted Children).Washington, DC: Author.

National Association for Gifted Children. (2013) State of the states in gifted education: National policy and practice data 2012-2013 (Report by the Council of State Directors of Programs for the Gifted & the National Association for Gifted Children).Washington, DC: Author.

National Association for Gifted Children. (2012) State of the states in gifted education: National policy and practice data 2011-2012 (Report by the Council of State Directors of Programs for the Gifted & the National Association for Gifted Children).Washington, DC: Author.

National Association for Gifted Children. (2007). State of the states in gifted education: 2006-2007 (Report by the Council of State Directors of Programs for the Gifted & the National Association for Gifted Children). Washington, DC: Author.

National Association for Gifted Children. (2003) State of the states in gifted education: National policy and practice data 2002-2003 (Report by the Council of State Directors of Programs for the Gifted & the National Association for Gifted Children).Washington, DC: Author.

National Association for Gifted Children. (1998) State of the states in gifted education: National policy and practice data 1997-1998 (Report by the Council of State Directors of Programs for the Gifted & the National Association for Gifted Children).Washington, DC: Author.

National Association for Gifted Children (2010). *2010 Pre-K-Grade 12 Gifted Programming Standards* (Report by the National Association for Gifted Children).Washington, DC: Author.

National Association for Gifted Children (NAGC). (2014). Students with concomitant gifts and learning disabilities. Position paper. Washington, DC: Author.

National Association for Gifted Children (NAGC). (2012). Students with concomitant gifts and learning disabilities. Position paper. Washington, DC: Author.

National Association for Gifted Children (NAGC). (1998). Students with concomitant gifts and learning disabilities. Position paper. Washington, DC: Author.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel: Teaching Children to Read* (NIH Publication No. 00-4769). Washington, D.C.: Government Printing Office.

Neihard, M., Reis, S., Robinson, N.M., & Moon, S. M. (2002). *The social and emotional development of gifted children: What do we know?* Washington, DC: Prufrock Press, Inc.

No Child Left Behind Act of 2001, 20 U.S.C. 70 § 6301 *et seq.* (Reauthorization of the Elementary and Secondary Education Act of 1965) (2002).

No Child Left Behind Act of 2001, PL 107-110, 115 Stat. 1425 (2002).

No Child Left Behind (NCLB) Act of 2001, 20 U.S.C. § 1411(e)(2)(C)(xi)].

No Child Left Behind Act, Title IX General Provisions, Part A Sec. 9101. Definitions US Department of Education. Retrieved September 30, 2012, from: http://www.ed.gov/policy/elsec/leg/esea02/pgl07.html

Okamoto, Y., & Case, R. (1996). *Exploring the microstructure of children's central conceptual structures in the domain of number*. Monographs of the Society for Research in Child Development, 61, 27–59.

Ostad, S. A. (1999). Developmental progression of subtraction strategies: A comparison of mathematically normal and mathematically disabled children. *European Journal of Special Needs Education*, 14, 21–36. doi:10.1080/0885625990140103.

Passow, A. H., & Rudnitski, R. A. (1994). Transforming policy to enhance educational services for the gifted. *Roeper Review*, *16*(4), 271-75.

Pegnato, C. W., & Birch, J. W. (1959). Locating gifted children in junior high schools: A comparison of methods. *Exceptional Children*, 25, 300–304.

Peters, S. J., & Gentry, M. (2012). Group-specific norms and teacher-rating scales: Implications for underrepresentation. *Journal of Advanced Academics, 23*(2), 125-144. DOI: 10.1177/1932202x12438717.

Pfeiffer, S. I., & Stocking, V. B. (2000). Vulnerabilities of academically gifted students. Duke University Talent Identification Program, *Special Services in the Schools*, The Haworth Press, Inc., *16*(1/2).

Plato & Bloom, A. (trans), (1991). *The Republic*, 2nd Edition. Basic Books; 2 Sub edition. ISBN-10: 0465069347.

Plucker, J. A., Callahan, C. M., & Tomchin, E. M. (1996). Wherefore art thou, multiple intelligences? Alternative assessments for identifying talent in ethnically diverse and low income students. *Gifted Child Quarterly*, *40*(2), 81-91.

Purcell, J. H., Eckert, R. D., & National Association for Gifted Children, W. D. C. (2005). *Designing services and programs for high-ability learners. a guidebook for gifted education*. Corwin Press.

Raghubar, K. P., Barnes, M. A. & Hecht, S. A. (2010). Working memory and mathematics: a review of developmental, individual difference, and cognitive approaches. *Learning and Individual Differences*, *20*(2) 110–22. http://dx.doi.org/10.1016/j.lindif.2009.10.005-285.

Räsänen, P., & Ahonen, T. (1995). Arithmetic disabilities with and without reading difficulties: A comparison of arithmetic errors. *Developmental Neuropsychology*, *11*(3), 275-295. doi:http://dx.doi.org/10.1080/87565649509540620.

Renaissance Learning. (2009). *STAR Early Literacy: Technical manual.* Wisconsin Rapids, WI: Author.

Renzulli, J. S. (1990). A practical system for identifying gifted and talented students. *Early Child Development and Care*, *63*(1), 9-18. DOI: 10.1080/0300443900630103.

Renzulli, J. S., & Delcourt, M. A. B. (1986). The legacy and logic of research on the identification of gifted persons. *Gifted Child Quarterly, 30*(1), 20-23.

Renzulli, J. S., Siegle, D., Reis, S. M., & Gavin, M. K. (2009). An investigation of the reliability and factor structure of four new scales for rating the behavioral characteristics of superior students. *Journal of Advanced Academics, 21*(1), 84-108.

Resnick, L. B. (1989). Developing mathematical knowledge. *American Psychologist*, 44, 162–169.

Resnick, Lauren B. *Education and learning to think.* Washington, DC: National Research Council, 1987.

Rogers, K. B. (2007). Matching needs of gifted learners to school possibilities. *Understanding Our Gifted, 19*(2), 15-20.

Rohrer, J. C. (1995). Primary teacher conceptions of giftedness: Image, evidence, and nonevidence. *Journal for the Education of the Gifted, 18*(3), 269-283.

Rousselle, L., Palmers, E., & Noël, M.-P. (2004). Magnitude comparison in preschoolers: What counts? Influence of perceptual variables. *Journal of Experimental Child Psychology*, 87, 57–84.

Russo, C. J., Harris, J. J., & Ford, D. Y. (1996). Gifted education and the law: A right, privilege, or superfluous? *Roeper Review, 18*(3), 179-182.

Sapon-Shevin, M. (1996). Beyond gifted education: Building a shared agenda for school reform. *Journal for the Education of the Gifted*, 19, 194–214.

Sattler, J. (2008). *Assessment of children: Cognitive foundations*. San Diego: J.M.

Savage, R., & Frederickson, N. (2005). Evidence of a highly specific relationship between rapid automatic naming and text reading speed. *Brain and Language*, 93, 152–159.

Schneider, W. (2009). The development of reading and spelling: Relevant precursors, developmental changes, and individual differences. In W. Schneider & M. Bullock (Eds.), *Human development from early childhood to early adulthood:*

Findings from a 20-year longitudinal study (pp. 199–220). Mahwah, NJ: Lawrence Erlbaum.

Schroth, S. T., & Helfer, J. A. (2008). Identifying gifted students: Educator beliefs regarding various policies, processes, and procedures. *Journal for the Education of the Gifted, 32*(2), 155-179.

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology, 94*(2), 143.

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*(3), 337-360.

Siegle, D., & Powell, T. (2004). Exploring teacher biases when nominating students for gifted programs. *Gifted Child Quarterly, 48*(1), 21.

Siegler, R. S. (1987). Strategy choices in subtraction. In J. Sloboda & D. Rogers (Eds.), *Cognitive process in mathematics* (pp. 81–106). Oxford, England: Oxford University Press.

Siegler, R. S., & Shrager, J. (1984). Strategy choices in addition and subtraction: How do children know what to do? In C. Sophian (Ed.), *Origins of cognitive skills* (pp. 229–293). Hillsdale, NJ: Erlbaum.

Simmons, F., & Singleton, C. (2008). Do weak phonological representations impact on arithmetic development? A review of research into arithmetic and dyslexia. *Dyslexia*, *14*(2), 77–94.

Shinn, M.R., and Shinn, M.M. (2002). *AIMSweb training workbook: Administration and scoring of reading-curriculum based measurement (R-CBM) for use in general outcome measurement.* Eden Prairie, MN: Edformation.

Shinn, M. R. (1989). *Curriculum-based measurement: Assessing special children.* New York. Guilford.

Speirs Neumeister, K. L., 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*(4), 479-499.

Starch, D. (1915). The measurement of efficiency in reading. *Journal of Educational Psychology, 6*, 1–24. doi:10.1037/h0073433.

Stecker, P. M., Fuchs, L. S., & Fuchs, D. (2005). Using curriculum based measurement to improve student achievement: Review of research. *Psychology in the Schools*, *42*, 795–820.

Stecker, P. M., & Fuchs, L. S. (2000). Effecting superior achievement using curriculum-based measurement: The importance of individual progress monitoring. *Learning Disabilities Research & Practice, 15*(3), 128 - 134.

Steen, L. A. (1999). Stiff, L. (ed).Twenty questions about mathematical reasoning. *Developing Mathematical Reasoning in Grades K-12*. Reston, VA: National Council of Teachers of Mathematics, 1999, pp. 270.

Sternberg, R. J. (1998). Ability testing, instruction, and assessment of achievement: Breaking out of the vicious cycle. *NASSP Bulletin*, *82*(595), 4-10.

Subotnik, R., Olszewski-Kubilius, P., & Worrell, F. (2012). A proposed direction forward for gifted education based on psychological science. *Gifted Child Quarterly*, 56(176). doi: 10.1177/0016986212456079

Subotnik, R., & Thompson, B., eds. (2010). *Methodologies for Conducting Research on Giftedness.* American Psychological Association, Washington, DC, USA.

Swanson, L., & Kim, K. (2007). Working memory, short-term, and naming speed as predictors of children's mathematical performance. *Intelligence, 35,* 151–168. doi:10.1016/j.intell.2006.07.001.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics* (3rd ed.). New York, NY: Harper Collins.

Tennessee Comprehensive Academic Program (TCAP, 2013). Achievement Test and Modified Academic Achievement Standards (MAAS) Assessment Grades 3 - 8, Tennessee Department of Education, Office of Assessment Logistics; http://www.state.tn.us/education/assessment/achievement.shtml.

Title IV, Part A. [Jacob K. Javits Gifted and Talented Students Education Act of 1988], Elementary and Secondary Education Act of 1988, 20 U.S.C. section 3061 et seq.

Title IV, Part B. [Jacob K. Javits Gifted and Talented Students Education Act of 1988], Elementary and Secondary Education Act of 1988, 20 U.S.C. section 3061 et seq.

Tennessee State Department of Education. (2010). *Tennessee state plan for the education of intellectually gifted students*. Revised August 2010.

Tennessee State Department of Education. *Tennessee Comprehensive Assessment Program*. Published test. State of Tennessee. Retrieved October 1, 2013, from http://tn.gov/education/assessment/achievement.shtml.

Tennessee State Department of Education. (2009). *User's guide to the Tennessee mathematics curriculum framework*. Tennessee Department of Education (Ed.) Retrieved from http://www.state.tn.us/education/ci/curriculum.shtml

Tennessee State Department of Education. (2010). TN Gifted Identification Matrix. Tennessee Department of Education (Ed.) Retrieved from http://www.tn.gov/education/article/special-education-evaluation-eligibility

Tomlinson, C. (2005). Quality curriculum and instruction for highly able students. *Theory Into Practice*, *44*(2), 160-166. doi: 10.1207/s15430421tip4402_10.

Tomlinson, C., Kaplan, S., Purcell, J., Leppien, J., Burns, D., & Strickland, C. (2005). *The Parallel Curriculum in the classroom: Essays for application across the content area, K–12.* Thousand Oaks, CA: Corwin.

Trochim, W. M. (2006). The Research Methods Knowledge Base, 2nd Edition. Internet WWW page, at URL: <http://www.socialresearchmethods.net/kb/> (version current as of October 20, 2006).

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP) 2011

Reading and Mathematics Assessments. http://nces.ed.gov/nationsreportcard/naepdata/\

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2000, 2003, 2005, 2007, 2009, 2011 and 2013 Mathematics Assessments.

U.S. Department of Education, Institute of Education Sciences, National Center for Education Statistics, National Assessment of Educational Progress (NAEP), 2005, 2007, 2009, 2011 and 2013 Reading Assessments.

U.S. Department of Health and Human Services, National Institutes of Health, National Institute of Child Health and Human Development. (2000a). *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00–4769). Retrieved from http://www.nichd.nih.gov/publications/nrp/smallbook.htm.

U.S. Department of Health and Human Services, National Institutes of Health, National Institute of Child Health and Human Development. (2000b). *Report of the National Reading Panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction: Reports of the subgroups* (NIH Publication No. 00–4754). Washington, DC: Government Printing Office.

Valdés, G. (2003). *Expanding definitions of giftedness*. Mahwah, NJ: Lawrence Erlbaum Associates.

Volker, M. A., Lopata, C., & Cook-Cottone, C. (2006). Assessment of children with intellectual giftedness and reading disabilities. *Psychology in the Schools*, *43*(8), 855-869. doi: 10.1002/pits.20193.

Wang, C., Algozzine, B., Ma, W., & Porfeli, E. (2011). Oral reading rates of second-grade students. *Journal of Educational Psychology, 103*(2), 442-454. doi: 10.1037/a0023029.

Weaver, B., & Wuensch, K. L. (2013). SPSS and SAS programs for comparing Pearson correlations and OLS regression coefficients. *Behavior Research Methods*, *45*(3), 880-895. Published online: 24 January 2013, Psychonomic Society, Inc. 2013.

Winner, E. (1997). Exceptionally high intelligence and schooling. *American Psychologist*, *52*, 1070-1081.

Wu, S. C., & Elliott, R. T. (2008). A study of reward preference in Taiwanese gifted and nongifted students with differential locus of control. *Journal for the Education of the Gifted*, *32*(2), 230-244.

Xu, F., Spelke, E. S., & Goddard, S. (2005). Number sense in human infants. *Developmental Science*, 8, 88–101.

Yeniad, N., Malda, M., Mesman, J., van IJzendoorn, M. H. & Pieper, S. (2013). Shifting ability predicts math and reading performance in children: A meta-analytical

study. *Learning and Individual Differences,* 23(0), 1–9, doi:10.1016/j.lindif.2012.10.004.

Zirkel, P. A. (2011a). RTI and the law. *West's Education Law Reporter*, 268(1), 1–16.

Zirkel, P. A. (2011b). State laws and guidelines for RTI: Additional implementation features. *Communiqué*, 39(7), 30–32.

Zirkel, P. (2005). *The Law on Gifted Education*. National Research Center on the Gifted and Talented (NRC/GT).

Vita

After an extensive career as a performing artist, Bruce Alan Ewing returned to the University of Tennessee, Knoxville, to finish a Bachelor of Arts degree in Art History. A successful certification process for teaching credentials from the Royal Academy of Dancing, London, UK, sparked an interest in curriculum design and teaching methods that led completion of a Master of Science in Education program, also at the University of Tennessee, Knoxville. Upon graduation, Mr. Ewing taught third grade at a small, independent school for five years. However, thirty years of artistic and creative collaborations inevitably led to an interest in giftedness and creativity, resulting in re-enrollment at the university as a candidate for a Doctor of Philosophy degree in Special Education, with a concentration in gifted education. During this course of study, he had the opportunity to teach undergraduate and graduate courses in special education as a graduate teaching assistant, and to mentor and supervise preservice teachers as a clinical supervisor. Under department supervision, he also developed and taught coursework designed to lead to employment-standard certification in gifted education, or as part of a STEM degree, for the distance education programs offered by the Education Department at the university. His research and professional interests include *all* levels of student learning, the neuroscience of learning, and creativity (as distinct from intellectual giftedness). As an advocate for children, he is especially interested in teacher training programs.