2008-07-17

Predicting Performance on Criterion-Referenced Reading Tests with Benchmark Assessments

Kaitlyn Nicole Dyson
Brigham Young University - Provo

PREDICTING PERFORMANCE ON CRITERION-REFERENCED READING TESTS WITH BENCHMARK ASSESSMENTS

by

Kaitie Dyson

A thesis submitted to the faculty of

Brigham Young University

in partial fulfillment of the requirements for the degree of

Educational Specialist

Department of Counseling Psychology and Special Education

Brigham Young University

August 2008

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a thesis submitted by

Kaitie Dyson

This thesis has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

_____	_____
Date	Gordon Gibb, Chair
_____	_____
Date	Timothy B. Smith
_____	_____
Date	Melissa Allen Heath

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the thesis of Kaitie Dyson in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

_____	_____
Date	Gordon Gibb, Chair

Accepted for the Department

_____	_____
Date	Ellie L. Young Graduate Coordinator

Accepted for the College

_____	_____
Date	Barbara Culatta Associate Dean, School of Education

ABSTRACT

PREDICTING PERFORMANCE ON CRITERION-REFERENCED READING TESTS WITH BENCHMARK ASSESSMENTS

Kaitie Dyson

Department of Counseling Psychology and Special Education

Educational Specialist in School Psychology

The current research study investigates the predictive value of two frequently-used benchmark reading assessments: Developmental Reading Assessment (DRA) and the Dynamic Indicators of Basic Early Literacy Skills (DIBELS). With an increasing emphasis on high-stakes testing to measure reading proficiency, benchmark assessments may assist in predicting end-of-year performance on high-stakes testing. Utah's high-stakes measurement of end-of-year reading achievement is the English Language Arts Criterion-Referenced Test (ELA-CRT). A Utah urban school district provided data for students who completed the DRA, DIBELS, and ELA-CRT in the 2005-2006 school year. The primary purpose of the study was to determine the accuracy to which the Fall administrations of the DRA and the DIBELS predicted performance on the ELA-CRT. Supplementary analysis also included cross-sectional data for the DIBELS. Results indicated that both Fall administrations of the DRA and the DIBELS were statistically

significant in predicting performance on the ELA-CRT. Students who were high risk on the benchmark assessments were less likely to score proficiently on the ELA-CRT. Also, demographic factors did not appear to affect individual performance on the ELA-CRT. Important implications include the utility of data collected from benchmark assessments to address immediate interventions for students at risk of failing end-of-year, high-stakes testing.

ACKNOWLEDGMENTS

I am grateful to several individuals and institutions that have contributed to this project. The Salt Lake City School District made this research project possible by providing a vast database and directing initial stages of the research. Gordon Gibb, my thesis chair, endured many long hours of refining, directing, and clarifying the literature review and outcomes analyses. Melissa Allen Heath facilitated my efforts to theoretically and practically apply this research to current needs in the field of education. Tim Smith helped to articulate the important outcomes and implications of the research. My cohort has sustained tremendous support through the academic and emotional rigors of my graduate career. My experience with the faculty and the campus at BYU has been broadening both academically and spiritually. Finally, I am grateful to my father, who has endorsed my pursuit of higher education as a means for augmenting my talents and abilities. This thesis has enlarged my capacity and appreciation for research and writing.

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

Assessing the achievement of students in the United States has expanded and intensified over several decades, with increasingly higher stakes placed upon state, local, and individual performance when tied to federal funding. The Elementary and Secondary Education Act of 1965 (ESEA) began an era of accountability which has increased measures and standards with each reauthorization. Political interests for American education have also intensified, with increasing influence from the federal government in local and state legislation, and controversy continues to circulate regarding the political and monetary pressures placed upon local education agencies to produce high performing students. Political interests were originally driven by concerns for international comparability evoked by the space race with the Soviet Union (Nichols & Berliner, 2007). The publication of *A Nation at Risk* (1983) catalyzed another shift in accountability for education, and *No Child Left Behind* (NCLB) (2001) increased accountability yet again, causing dilemmas for local and states "to comply or to educate" (McNeil, Coppola, & Radigan, 2008, p. 25).

Bluebello (2000) reports that rigorous measures and assessments have become fundamental tools for education reform in the present day. Assessment for accountability in the NCLB era requires states and districts to measure and report Adequate Yearly Progress (AYP) in order to determine school quality based on student performance. While the ideals of NCLB are well-intentioned and have led to increases in student performance for some urban areas, the greater concern is whether assessment for AYP has students demonstrating broad content knowledge or narrow test-driven facts (International Reading Association, 1999). The high stakes associated with federal

funding are generally accepted by the public and political figures (Afflerbach, 2002), but the related testing is believed by many to provide only a contracted snapshot of the wide range of benchmarks and learning goals for many state standards (Davis, 1998).

Reading achievement is one area of particular interest. While high-stakes testing is utilized increasingly for measuring AYP in reading, no research relates this level of testing to improved reading achievement (Afflerbach, 2005). High-stakes testing at the end of each school year yields little or no useful information to guide teachers in their daily instructional decision making, so the utility of such testing lies only in quantifying school, district, and state performance for the requirements of NCLB. This leaves districts and schools with the task of finding or developing reading assessments that provide more readily useable student achievement data at regular intervals or benchmark periods during the school year.

This study looks at the accuracy of two benchmark assessments used in the Salt Lake City school district to monitor reading skill progress in anticipation of the end-of-year state testing used to determine AYP. The review of literature presents the historical context of NCLB and explains how educational objectives have shifted the function of testing in the last century. This is followed by a description of Utah's test for AYP in reading, a review of formative assessment for reading achievement, and descriptions of the two benchmark assessments used in the Salt Lake City schools.

LITERATURE REVIEW

In 2002, President George W. Bush secured enactment of the No Child Left Behind Act (NCLB), a landmark legislation with an ambitious plan to increase elementary and secondary school quality and performance (U.S. Department of Education, 2005). According to Paige (2006), the philosophical roots of NCLB emerged from decades of concern about the quality of the U.S. educational system and its comparability to other countries. In addition, NCLB provisions derived from concerns about the poor education for children with disabilities and other disadvantages, such as limited English language learners (limited English proficiency, LEP) and the effects of poverty.

Historical Context of No Child Left Behind 2001

NCLB is the most recent federal legislation aimed at improving education outcomes for American children. Federal interest in raising overall achievement was sparked by the Soviet Union's successful launch of the Sputnik spacecraft in 1957. In response to Sputnik, federal provisions were offered to elementary, secondary, and higher education institutions through the National Defense Education Act (1958), specifically targeted at mathematics, science, and foreign language as well as vocational training, school libraries and media centers, and counseling (Paige, 2006).

Following passage of the Civil Rights Act of 1964, Congress passed the Elementary and Secondary Education act (ESEA, 1965) as the education focus of President Johnson's War on Poverty (Schugurensky, 2002). Congress derived ESEA's initiatives from President John F. Kennedy's proposals addressing an American education competitive with other countries and equally accessible to all, objective to

religion, race, and socio-economic background (Jeffrey, 1978). In particular, the link to Johnson's War on Poverty" sought to address the negative impact of poverty on educational opportunity. Johnson envisioned equal opportunities for all children to participate in quality education and to receive necessary support services. Johnson proposed that increased concentration and funding for educating lower-income children and their education would decrease the drop-out rate and inevitably produce more capable adults who would be less likely to perpetuate the cycle of poverty (Schugurenskey, 2002). The Committee on Labor and Public Welfare (1965) identified the pursuit of ESEA to "strengthen and improve educational quality and educational opportunities in the nation's elementary and secondary schools" (p. I). The original programs for the ESEA of 1965 provided funds for educational programs including:

Title I: Education of children of low income families

Title II: School library resources and instructional materials

Title III: Supplementary educational centers and services

Title IV: Educational research and training; Cooperative research act

Title V: State departments of education

The ESEA has since undergone numerous amendments to improve and expand its application and implementation. The original programs implemented in the 1965 legislation served as groundwork for further educational legislation (Spring, 1993).

Spring (1993) stated that the ESEA led to three important consequences for future legislative action. First, the bill linked federal aid to specific concerns of national policy, including poverty and economic growth, and identified specific programs and needs to be met through federal aid. Second, it also linked federal aid to educational programs

directly assisting underprivileged children rather than institutions. Third, ESEA gave the federal government a more direct role in educational initiatives and offered state departments of education some administrative power over federal funds. ESEA was a catalyst for later educational legislation, including the Education for All Handicapped Children Act (1975), subsequently revised as the Individuals with Disabilities Education Act (1990).

Reauthorization and Expansion of ESEA

President George W. Bush described his plan for bipartisan educational reform as "the cornerstone of my administration" and expressed concern that "too many of our neediest children are being left behind" (U.S. Department of Education, n.d.). NCLB indicated that all-inclusive education reform required higher standards and stronger accountability for student performance (U.S. Department of Education, 2004). NCLB has emerged from a backdrop of increasing concern for the "mediocre educational performance" (U.S. Department of Education, 1983, para. 2) of the general U.S. student population, particularly the achievement gap and students at-risk academically and economically. The authorization of NCLB aimed to improve standards of accountability for states, school districts, and individual schools. The authorization raised standards of academic assessment and student performance and required evidence-based methods for teaching core curriculum.

President Bush appointed Rod Paige as the U.S. Secretary of Education to promote and regulate the initial stages of NCLB. Paige was the former superintendent of the Houston Independent School District which Bush endorsed for its significant but controversial performance gains in state-wide testing and significant reduction of the

achievement gap between white and minority students in the urban school district. Paige attributed Houston's success to heightened accountability for school and district administers by linking district and employee monies to student performance. He used the standardized state assessment, the Texas Assessment of Academic Skills (TAAS), as the primary measure of academic achievement and offered state funding to private schools for students attending Houston's lowest performing schools (Steinberg, 2000). Much debate circulates regarding Houston's success; nevertheless, some of NCLB's basic premises are based upon the evidence of Houston's gains under Paige's direction (Schemo & Fessenden, 2003).

Some fundamental components of NCLB based upon Houston's previous success were increased expectations and provisions for accountability and student performance outcomes (Schemo & Fessenden, 2003). Accountability is an ongoing issue in educational reform (Samuels & Edwall, 1975). NCLB seeks to strengthen Title I accountability by requiring all public schools to report student performance on annual statewide progress broken down by race, socio-economic status, disability, and limited English proficiency in order to track progress of all students, specifically disadvantaged students. Under accountability provisions for NCLB, states are required to establish benchmark standards as well as evidenced instructional and assessment tools consistent with federal standards to ensure that children make sufficient gains in reading, mathematics, and science (The Center for Public Education, 2006).

Adequate Yearly Progress (AYP) is a reporting measure which determines the academic achievement of schools, districts, and states. Each state is responsible for determining annual state targets to accomplish reading, math, and science proficiency by

2014 (Fuchs & Fuchs, 2004). AYP is the minimum level of improvement that schools, districts, and states must achieve each year in order to reach 100% proficiency by 2014. Individual schools who do not meet AYP each year face consequences which can result in reduced funding and corrective action on the federal level. In order to monitor academic achievement at the school and district level, each state developed and implemented a statewide accountability system that should be effective in ensuring that all local educational agencies, public elementary schools, and public secondary schools make AYP. This accountability system includes statewide exams aligned directly with state standards that measure whether or not students at each grade level have mastered specific content and skills (The Education Trust, 2004).

While NCLB advocates program flexibility and federal grants for States and Local Education Agencies (LEA), such autonomy is contingent on approved statewide accountability systems. The contingencies apply tremendous pressure on states to produce high-stakes testing measures to meet AYP, and the consequences of failing such standards can result in federal overhaul of individual schools and funding restrictions for school districts (Rothkope, 2007). As a result, the National Council of Teachers of English (2005) reports that the Utah House Representatives voted to reject NCLB implementation “except where there is adequate federal funding” (p. 2). Similarly, legislators in both Minnesota and Arizona have introduced “opt-out” legislation that basically permits them to reject certain NCLB stipulations, and 10 other state legislatures have passed statements “highly critical of the law” (*Lack of Funding section*, para. 2).

School choice is offered to Title I students who live within boundaries of a poorly performing schools. The option of school choice given to parents and students includes

funds for transportation and other supplemental services necessary for students to meet State academic standards. The flexibility of school choice is intended to impress LEAs and educators to provide highly qualified teachers and systematic teaching strategies in order to keep students, moreover, monetary funding at their schools (United States Department of Education, 2002).

Criticism of both ESEA and NCLB involve increased federal participation in state and local educational objectives (Kantor, 1991; McColl, 2005). With NCLB, greater flexibility for federal funding expenditures is exchanged with states for more robust accountability outcomes. A competitive State and Local Flexibility Demonstration Program is based on a performance agreement with the Secretary of State. States are offered flexibility to expend funds in any ESEA-authorized programs which include Teacher Quality State Grants, Educational Technology State Grants, Innovative Programs, and Safe and Drug-Free Schools programs (United States Department of Education, 2002). State and local eligibility for the Flexibility Demonstration Programs are based upon state and local performance in AYP and annual state assessments.

Other ESEA programs were reauthorized and expanded through NCLB. The Eisenhower Professional Development and Class Size Reduction programs were combined to create the Teacher Quality State Grants program. The primary objective of the program is to yield high-quality teachers with training in scientifically and empirically-based teaching methods. States and LEAs are offered increased flexibility to employ strategies which best meet their particular needs contingent with student performance on annual state testing. Programs were also expanded to support bilingual and limited English language learners. A new state formula program has been created to

ensure specific program implementation for limited English proficient learners and provide additional support to meet state and federal standards.

In addition, schools are required to report school safety statistics and offer school choice to children who are victims of persistently dangerous schools. LEAs are mandated to contribute federal funds toward empirically-based drug and school safety curriculums in their schools (United States Department of Education, 2002). The expansion and reauthorization of ESEA initiatives through NCLB has significantly increased accountability and potential rewards with positive outcomes for states and LEAs.

Summary of Reading First and Early Reading First Initiatives

No Child Left Behind places particular emphasis on reading proficiency and claims that every child should read at grade-level by the end of third grade. The Reading First initiative for elementary school and the Early Reading First program for preschool are designed to target students at-risk for reading failure. In 2002, Lyon (as cited in Hare, 2002) reported that 37 percent of United States fourth graders read below grade level that increases to 60 percent among minorities. In addition, 75 percent of 9 year-old children who cannot read, never learn to read. Research denotes that early literacy concepts can predict children's subsequent reading achievement (DeBruin-Parecki, 2004). In addition, Adams (1990) asserts that reading proficiency in first grade appears to be the best indicator of latter school achievement. Reading difficulties become increasingly problematic and proliferate over time (Torgensen, 1998). Remediation is offered through intense intervention (Vaughn & Schumm, 1996); however, students continue to lag as their peers progress at expected reading benchmarks (Rashotte, Torgesen, & Wagner,

1997). Torgesen asserts that early identification and prevention is critical to reduce reading failure.

Impact of Educational Objectives on the Function of Testing

Educational historian and former Department of Education official Diane Ravitch echoed NCLB's sentiment for testing, stating that, "standards are essential both for quality and equal opportunity" (Ravitch, 1995, para. 14). Moreover, homeschool advocate Duffy (2001) stated that NCLB emphasizes "if standards are not tested, they will not be taught" (*Broad Support for Standards*, para. 2). With increased accountability provisions and student performance standards, high-stakes testing has emerged as the primary form of evaluation for AYP (NCES, 2006). High-stakes testing is the use of standardized tests as a measure of academic performance and indicates important consequences for students, school districts, and states (Lagenfeld, Thurlow, & Scott, 1997). These consequences include monetary funding and district reputation (Nichols, Glass, & Berliner, 2005), as well as grade promotion for students in some public schools (Lagenfeld et al.).

National standards requiring states to align student achievement with specific content and skill performance have shifted the function of testing. Early philosophy and utility of testing was primarily intelligence testing, which provided normative information. The function of intelligence testing was to determine students' capability for benefiting from contemporary education, and for military service and employment in industry (Samuells & Edwall, 1975). With political issues evoked by NDEA, ESEA, Civil Rights, and IDEA, the U.S. Department of Education published *A Nation at Risk* (1983), which addressed concerns about American education and stimulated the concerns

for improving standards and investment by educators, parents, and citizens. Previously-accepted reports of teacher qualifications and budget allocations were no longer sufficient. Both the government and the public required schools to address the quality of student education and performance; specifically, how schools were meeting societal needs, through current objectives, methods of assessment, and ameliorating instructional effectiveness (Samuells & Edwall). As a result, the function of testing shifted from comparative data to competency-based information.

The shift from comparative to competency-based assessment was influenced by the limited decision-making effectiveness of comparative data. Tyler (1972) stated that normative testing does not determine what is learned and does not provide reliable information for decisions regarding student ability and progress. Thus, criterion-referenced testing emerged as the form of testing to determine what is learned.

Criterion-referenced, or competency-based testing, is derived from the mastery model of instruction. Bloom (1968) endorsed learning by mastering skill domains rather than normative comparisons. Gagne's *Conditions of Learning* (1965) examined the relationship between learning objectives and appropriate instructional design. He provided research-based instructional design to enhance the learning experience (Gagne & Medsker, 1996). Mastery learning is based on performance as interpreted in terms of defined criteria. The mastery model of instruction assumes that all children are capable of learning and seek to facilitate progress toward various skill/knowledge criterion. Students are taught skill and content domains and, subsequently assessed on that specific material. Once students exhibit the essential elements involved in learning that particular domain, they are promoted to the next domain (Bloom, 1968). Criterion-referenced testing

describes specific behavior expected of a person at a particular level or whether behavior meets a standard of quality. Students' academic performance is located along a continuum of achievement from no proficiency to perfect performance where the standard or criterion by which students' performance is compared determines the degree of proficiency (Glaser, 1963). The primary objective of criterion-referenced testing is to determine the degree to which a student has learned the material taught in the classroom.

Influence of Criterion-Referenced Testing

Criterion-referenced testing (CRT) is used to inform classroom instruction and measure the degree to which students meet state and national standards. Criterion-referenced testing as a form of high-stakes testing is reportedly used among 43 out of 50 states as a measure of AYP (NCES, 2006). Students earn a proficiency score which reflects the degree to which each individual meets state standards. Overall student proficiency reflects the aggregate performance of school districts and states.

Historical Context of Reading Assessment and Criterion-Referenced Testing

The Reading First and Early Reading First initiatives place particular emphasis on reading achievement and curbing literacy-related failure by employing scientifically-based reading instruction programs in early grades, as well as early identification of at-risk readers. The evolution of reading assessment has followed a similar shift to general testing trends with increasingly diagnostic educational objectives.

At the turn of the twentieth century, reading was synonymous with literature appreciation (Smith, 2002) and involved mostly recitation and rote regurgitation (Huey, 1908). Huey asserted that oral reading required no thought retention and/or manipulation. Reading objectives changed when Judd and Parker, predecessors to John Dewey, asserted

that deriving meaning was more important than reciting (Smith). Their theoretical expansion led Monroe to conclude that general reading ability consists of multiple factors (Singer, 1983). While the definition of reading continued to evolve, Thorndike (1917) stated that reading is thinking and requires cognitive manipulation and attention to subtle linguistic rules. Sarroub and Pearson (1998) reported that the earliest systematic efforts to illustrate reading ability by evaluating comprehension appeared during World War I. High rates of illiteracy among World War I and II soldiers fueled research to define important reading fundamentals, such as deriving meaning from text (Smith). However, research has mostly resulted in indirect indicators of the actual process (Sarroub & Pearson).

Purpose of reading assessment. According to Sarroub and Pearson (1998), reading assessment has served a similar purpose since its initial praxis in the early 20th century. Its evaluative functions have provided accountability, and instructional and program placement utility to varying degrees. Both political and fundamental views of the reading process have influenced evolutionary developments in reading assessment. Efforts to characterize more specific elements of reading comprehension resulted in various testing formats including short answer (1910-1930), multiple-choice, answer bubbles (1930s), the essay (after WWII), and oral response in discussion (with expanded emphasis on assessment). Reading readiness, such as letter identification, became a normative measure which compared kindergarten students' proficiency in skills presumed requisite for formal reading instruction.

Predictive studies show that children are disadvantaged when entering primary school grades without basic early literacy skills (Hammill & McNutt, 1980; Scarborough,

1998). While these formats facilitated outcomes of reading comprehension, they failed to penetrate the actual learning process to effectively inform instruction. The philosophical backdrop of behavioral psychology in the 1930's-60's induced standardized, normative testing into reading assessment (Sarroub & Pearson).

Development of criterion-referenced assessment. Criterion-referenced assessments sought to break down learning into elements for any learning domain or process. The number of comprehension sub-skills increased significantly and assessments emphasized skill sets rather than complete passages (i.e. understanding sequential order vs. whole-passage comprehension). Criterion-referenced assessments enabled teachers and parents to articulate the specific areas in which students lagged and the degree to which students were achieving reading proficiency (Sarroub & Pearson, 1998).

Further research for reading comprehension expanded assessment utility (Durrell, 1955, Meyer & Rice, 1985; Ericson & Simon, 1980; Goodman & Burke, 1970; Sarroub & Pearson, 1998; Vygotsky, 1986). Sarroub and Pearson stated that the reauthorization of Title I in 1968 provoked an accountability era with states and districts exchanging performance scores for additional funding to help at-risk readers.

As a result, state assessment systems emerged in the early 1970's and high-stakes testing became a subsequent form of evaluation to meet both state and national standards. While the philosophical elements for the reading process may continue to evolve, criterion-referenced testing remains the most efficacious way to measure reading proficiency and concurrently inform teacher instruction.

Evaluation of Reading Achievement

The Reading First Program requires that assessments must identify students who may be at risk for reading failure or who are already experiencing reading difficulty. A National Institutes of Health study showed that 67 percent of children identified at risk for reading difficulties were able to obtain grade-level reading ability when they received early intervention (Coordinated Campaign for Learning Disabilities, 1997). Assessment is fundamental to guide classroom instruction (Wren, 2004). The Access Center (2005), an organization devoted to improving education for students with disabilities, defined the purpose and benefits of assessment.

Progress monitoring. First, identifying skills that students have or have not mastered enable teachers to know the status of each student's reading ability. Second, monitoring student progress allows teachers to know, both individually and collectively, whether their students have mastered fundamental literacy skills and are prepared to build upon those skills with more difficult content. Third, consistent assessment facilitates informed decision-making in regard to instructional appropriateness for each student. In conjunction with decision-making utility, assessment permits teachers to evaluate their own instructional effectiveness and create the most appropriate instructional environment for their students. Finally, assessment facilitates the improvement of teacher instruction. Therefore, the most useful assessment evaluates both student outcome data and instructional effectiveness to help shape the most efficacious learning environment.

Formative assessment. Black and William (1998) define the key outcome of *formative* assessment as instructional adaptation from feedback gathered from student learning procedures. The instructional adaptations are then used to meet student needs.

According to Cowie and Bell (1996), formative assessment is a two-way process with both teacher and student identifying, increasing, and responding to learning. Nicol and Macfarlane-Dick (2006) have shown how feedback endorses processes of self-regulation. William (2005) stated that formative assessment should support learning and may be used to support summative inferences; evidence collected for summative purposes can rarely be broken down to support learning. In terms of students at-risk for reading failure, formative assessment promotes differentiated instruction for students and individualized intervention where the need for responsive services post-failure may be prevented (Shinn et al., 2002).

Curriculum-based assessment and measurement. Curriculum-based assessment (CBA) and curriculum-based measurement (CBM) are frequently used formative evaluations utilized locally for decision-making purposes including progress monitoring, diagnostic measurement, and planning for interventions for reading achievement (Sibly, Biwer, & Hesch, 2001). CBA is based upon direct observation and student performance data in the local curriculum that provides information for instructional decision-making (Deno, 1987). They can offer both general normative and specific criterion-based information gathered through continuous, research-based assessment (Learning First Alliance, 2000). CBM is a reliable and valid measurement system used for progress monitoring in basic academic skill areas, such as reading proficiency (Deno, 1985; Shinn, 1989). The content of the CBM tests may be drawn from a specific curriculum or represent generalized grade-level outcomes.

More than current student performance data, CBM test content represents projected end-of-year global performance standards (Stecker, n.d.). Progress monitoring

involves an intra-individual framework, with CBM data recorded regularly (weekly or monthly). Based on the data, student scores are graphed, and the slope derived from CBM data quantifies reading improvement. Subsequently, the teacher interprets the outcome data to formulate instructional decisions (Fuchs, Fuchs, & Compton, 2004). Thus, CBM becomes a formative way to assess student progress over time and infer summative outcomes (Stecker).

Benchmarks. Benchmarks are standards or reference points from which normative and criterion-referenced data may be derived. CBM benchmark assessments specify the lowest performance levels linked with prospective reading achievement (Fuchs, Fuchs, & Compton, 2004). More specifically, benchmarks designate a performance level that evidences probability of meeting subsequent objectives (Good, Simmons, & Kame'enui, 2001). At specific points in the year (i.e. quarterly), student performance data is compared to grade-expected criterion in skills areas necessary for reading proficiency, such as phonemic awareness, fluency, and comprehension. Students unable to meet benchmarks at specified periods are candidates for intensive reading instruction (Fuchs, Fuchs, & Compton, 2004). Good et al. (2001) emphasized the necessity of establishing meaningful benchmark systems that adequately account for student population. Benchmark assessment systems have become increasingly utilized to predict student performance on end-of-year accountability assessments (Olsen, 2005).

National, state, and local standards. As a result of NCLB, local curriculum has been aligned with state and national standards. These standards are increasingly emphasized in classrooms, and benchmark testing has become the common method to determine the achievement of standards in the local curricula (O'Shea, 2006). In 2005,

Olsen reported that an estimated 70% of districts utilized benchmark testing and, as many as 80% projected their use the following year. Districts have apparently accepted that early comparison of reading performance to standards will better prepare students for end-of-year assessments which determine AYP. Reeves (as cited in Olsen, 2005) credits feedback to principals and teachers as the most useful feature of benchmark assessments, if used to make instructional decisions.

Predictors of Reading Achievement

School-based data on early literacy skills can facilitate effective shaping and identification of research-driven theory and strategies for reading (Baker & Smith, 2001;Good, Simmons, & Kame'enui, 2001). Extensive research has identified the need for early identification of children at-risk for reading failure. Several studies have shown that first grade reading ability is indicative of long-term reading ability (Francis, Shaywitz, Stuebing, Shaywitz, & Fletcher, 1996; Torgeson & Burgess, 1998). Cunningham and Stanovich (1997) assert that first-grade reading achievement correlates with 11th-grade reading proficiency. The effects of poorly learned fundamental literacy skills create burgeoning and pervasive problems for weak readers (Torgesen, 1998). Poor readers have fewer opportunities for reading practice to improve reading (Allington, 1984). Over time, poor readers also tend to develop a negative mindset toward their weakness (Oka & Paris, 1986). Decreased vocabulary growth (Nagy, Herman, & Anderson, 1985) also leads to reduced reading comprehension strategies (Brown, Palinscar, & Purcell, 1986).

According to the Institute of the Development of Educational Achievement (2004), the bottom 25% of the reading field begin to significantly diverge from successful

peers by the end of first grade. Focus on crucial and basic early literacy skills can more closely monitor and reinforce achievement of fundamental skills. The National Reading Panel (2000) was given a congressional mandate to identify skills fundamental to reading achievement and review and evaluate research on reading instruction. The Panel identified five primary emphases in early reading literacy known as the "big ideas of beginning reading" (Institute of the Development of Educational Achievement, 2004). These are listed below.

- *Phonemic awareness* involves skills to identify, process, and manipulate sounds in spoken language.
- *Phonics instruction* requires students to understand the relationships between spoken and written language.
- *Vocabulary* involves identifying the meaning of spoken and written language in order to communicate effectively.
- *Fluency* involves skills needed to read text with accuracy and speed
- *Reading comprehension* requires students to understand written text and productively communicate meaning and application. Reading comprehension helps to extend general knowledge and academic achievement.

Reading readiness skills. Reading readiness skills and emergent literacy, such as learning how to hold a book and pencil, discriminating shapes, interpreting illustrations, letter identification, concepts and conventions of print, and phonemic awareness, are pre-requisites to reading and are linked to later reading achievement (Clay, 1966; Hammill & McNutt, 1980; Scarborough, 1998; Schumm, 1996; Snow, Burns, & Griffin, 1998;Sulzby & Teale, 1991). As children develop literacy skills, many contributing factors affect

overall achievement of basic reading skills including oral language abilities (expressive and receptive) and verbal memory (Scarborough, 1998; Snow, Burns, & Griffin, 1998). Reading proficiency must build upon fundamental skills (phonemic awareness, phonics, vocabulary) in order to achieve meaning from written text (fluency, comprehension) and meet higher academic demands (Alliance for Excellent Education, 2006).

Reading comprehension. In the second edition of the Partnership for Reading's *Putting Reading First* (Amrbruster, Lehr, and Osborn 2003) asserted, "Comprehension is the reason for reading. If readers can read the words but do not understand what they are reading, they are not really reading" (p. 47). Scott (2007) conceptualizes reading comprehension as constructing meaning through "decod[ing] words fluently, understand[ing] vocabulary, mak[ing] inferences, and relat[ing] the ideas in text to their prior knowledge and experiences" (p. 1).The National Center for Education Statistics (2005) defined reading comprehension as …"an active and complex process that involves understanding written text, developing and interpreting meaning, and using meaning as appropriate to type of text, purpose and situation" (p.2). Likewise, Torgeson (1998) asserted that "no matter what one's personal preferences for instructional method, the end goal is to help children comprehend written material at a level that is consistent with their general intellectual abilities" (p. 2).

Reading fluency. Oral reading fluency is both a theoretically and research-based indicator of reading comprehension and is comparable to direct measures of reading comprehension (Deno, Mirkin, & Chiang, 1982; Fuchs, Fuchs, & Maxwell, 1988; Hosp & Fuchs, 2005). Rasinski (n.d.) asserts that fluency extends beyond ability to read fast. He assesses reading fluency in three components: first, decoding skills; second,

automaticity; third, vocabulary skills. Fluency is a set of skills that allows readers to quickly and accurately decode and comprehend text simultaneously (National Reading Panel, 2001). Fluent readers do not have to think about reading (or decoding) words and can think about what the text means; therefore, fluency is an important indicator of reading competency, because it "frees students to understand what they read" (Amrbruster, Lehr, and Osborn, 2003, p. 31).

Curriculum-Based Assessment and High-Stakes Testing

Curriculum-based measurement (CBM) is increasingly utilized by local school districts to prepare for state tests of reading achievement (Fuchs & Fuchs, 2004). Important research has identified CBM as a practical and diagnostic measurement tool for long term objectives of regular and special education students (Shinn, 1998). Research has shown that brief, one-minute reading probes are accurate indicators of reading skill (Deno,Mirkin, & Chiang, 1982). Marston (1989) aggregated research showing that correlations between oral reading rates and multiple global reading skills, such as prediction, inferences, and comprehension, followed positive correlations for CBM. CBM research shows that oral reading fluency is a strong indicator of reading ability in elementary-aged children (Good & Jefferson, 1998; Marston, 1989; Shinn, 1989; 1998).

Predicting performance on state testing. Research has also provided evidence that CBM ameliorates prediction of state assessment performance outcomes (McGlinchey & Hixson, 2004; Stage & Jacobson, 2001). Continuous progress monitoring allows teachers to modify instruction according to students' needs and skill mastery, better preparing them for end-of-year state assessments (McGlinchey & Hixson). Stage and Jacobson (2001) compared student performance on CBM oral reading fluency probes for fall,

winter, and spring terms with spring administration of the Washington Assessment of Student Learning (WASL). Results indicated that CBM oral reading measures strengthened prediction of WASL performance as compared to base rates. Sibley, Biwer, and Hesch (2001) examined the relationship between CBM oral reading fluency measures and performance on state and local tests of reading achievement. According to their research, students meeting or exceeding established oral reading fluency benchmarks were likely to achieve proficiency on state standards testing.

Discussing reading fluency benchmarks. Likewise, students who did not meet oral reading fluency benchmarks were unlikely to achieve proficiency on state standards. In addition, similar correlations were shown between oral reading fluency benchmarks and local grade level reading tests, which were administered fall term. Other research supports the decision-making utility of oral reading fluency benchmarks as indicators of performance on global measures of reading and high-stakes, state-level assessments (Hintze & Silbertglitt, 2005); moreover, research supports the ongoing predictive value of performance on subsequent state assessments for the same students (Good, Simmons, & Kame'enui, 2001).

Concerns about High-Stakes Testing

High-stakes testing outcomes can produce heavy consequences for some local schools and students. A single test score indicating overall attainment of language arts content, including reading proficiency, attempts to establish accountability through an "end of a gun barrel approach, rather than building consensus" (Casbarro, 2005, p. 20). Casbarro implies that NCLB imposes accountability upon states through coercive regulations rather than through collaboration. Reeves (as cited in Olson, 2005) compared

high-stakes summative data to an autopsy, stating that results from end-of-year testing only indicates a deficit after the fact but does not indicate when, where, or to what extent the deficit occurred.

Decisions based upon high-stakes testing have created irrevocable problems for some students and schools. In 1999, New York City mistakenly sent thousands of students to summer school based upon incorrectly calibrated scores for the Citywide Tests (Steinberg & Henrique, 2001). Scoring and testing calculation error have occurred in numerous other states, including Washington, Ohio, Tennessee, Florida, and Wyoming (Kale, 2000). It appears that more problems than improvement have resulted from placing individual and local decisions upon high stakes.

Developmental Reading Assessment (DRA)

The Developmental Reading Assessment (DRA) is a diagnostic literature-based reading program that directs teacher instruction with baseline and benchmark data in grades K-8. The DRA was created in 1988 by the Upper Arlington City School District in Ohio. The aforementioned *A Nation at Risk* (National Commission on Excellence in Education, 1983) was a catalyst for development of the DRA, as Ohio required districts to identify students who were at risk of failing in reading. Most Ohio school districts chose standardized, norm-referenced testing to meet state requirements (Beaver, 2002). Upper Arlington adopted a competency-based framework which could more specifically link curriculum-based information and instructional utility.

Important features of DRA. The DRA has three specific features that broadened diagnostic strategies and instructional utility of previous reading programs used in Upper Arlington schools. First, the committee wanted to develop an assessment that teachers

could administer, as opposed to trained specialists, to more directly involve themselves in the evaluative process and inform them explicitly of students' strengths and weaknesses in reading. Second, oral reading fluency was the primary measurement for reading fluency; however, the reading assessment committee for the DRA asserted that comprehension is a critical element of reading and should be measured directly. Thirdly, the DRA was expanded to assess and monitor all students in kindergarten to third grade, rather than only at-risk first-grade students. The reading assessment committee for the DRA sought to employ comprehensive, research-based strategies and robust documentation to assess and monitor student reading over time.

The DRA matches students to an appropriate level of text difficulty. Independent reading levels identify the highest level book a child can read with 90% to 95% accuracy and with at least 70% comprehension. Students read a grade-appropriate book from which the teacher evaluates reading accuracy, fluency, and comprehension. In order to check fluency and comprehension, a teacher may ask a student to retell the story subsequent to reading, accounting for characters, thematic details, and predictions. After a student's reading level is determined, the teacher groups students by ability, building instruction and reading strategies upon already established skills and targeting skills for progression toward the next reading level. DRA administration varies by school district but is typically given at the beginning, middle, and end of the school year. Student performance is measured by benchmarks established locally or predetermined in the DRA curriculum. The frequent administration and curriculum-based framework for the DRA provides formative data to inform summative assessment.

Supporting Research of DRA. Beaver and Carter (2003) stated that the DRA was designed to measure and monitor student reading skills and strategies, support teachers in identifying student needs and increasing instructional effectiveness, and prepare students to meet local and state reading standards. Studies conducted by Williams (1999) and Weber (2000) presented effective utility of the DRA. Inter-rater reliability was 80-100% (Weber, 2000) and was also determined .74 across all teachers and students during a separate evaluation (Williams, 1999), indicating that teacher evaluation and scoring procedures were consistent across teachers. Internal consistency was high with a Cronbach's alpha of .98 for item separation reliability and .97 for text separation reliability (Williams, 1999).Correlation coefficients ranging from +.92 to +.99 indicate significant test-retest reliability (Weber, 2000). Construct and criterion validity were measured by correlating scores DRA reading level assessments with Iowa Test of Basic Skills (ITBS) Subscales of vocabulary, reading comprehension, and total reading. Construct validity was significant at the 0.01 level (2-tailed) for all 3 subscales with the most powerful Spearman's Rho rank-order correlation of +.71 for total reading (Williams, 1999). Criterion validity examined the extent to which the DRA independent reading level predicted performance on the reading comprehension subscale of the ITBS. Spearman rank-order correlation coefficients ranged from +.54 to +.83, suggesting a moderate level of criterion validity. Thus, studies reflect effective utility of DRA administration and inferential value in summative performance.

Dynamic Indicators of Basic Early Literacy Skills (DIBELS)

The Dynamic Measurement Group (DMG) is an educational research company which conducts extensive research on assessment and helps provide research-based

curricular tools for practical use in the classrooms. DMG was founded by Roland H. Good, III and Ruth Kaminski, authors of the DIBELS. The Reading First and Early Reading First initiatives stimulated extensive research to target effective skills and instructional techniques for reading achievement (Armbruster, Lehr, & Osborn, 2003). In response to the research outcomes of the National Reading Panel (2000) and NCLB, Good, Kaminski, and researchers at the University of Oregon, College of Education created the DIBELS to assess scientifically-based early literacy skills and aggregate data for the Reading First legislation. The DIBELS sought to address literacy skills formatively in order to target students at-risk of not achieving reading proficiency.

DIBELS is a standardized, individually administered K - 6 formative assessment which provides baseline and benchmark data to inform teacher literacy instruction and intervention. Good & Kaminski stated the following about the DIBELS.

> The measures were developed to assess student development of phonological awareness, alphabetic understanding, accuracy and fluency reading connected text, vocabulary and comprehension. Each measure has been thoroughly researched and demonstrated to be a reliable and valid indicator of early literacy development. When used as recommended, the results can be used to evaluate individual student development toward validated instructional objectives as well as provide feedback on effectiveness of intervention support. (www.dibels.org)

DIBELS subscales. The DIBELS subscales specifically examine initial sounds fluency, letter-naming fluency, phoneme segmentation fluency, nonsense word fluency, and oral reading fluency (ORF). Initial sounds, letter-naming, phoneme segmentation,

and nonsense word fluency are considered prerequisites to oral reading fluency and are targeted according to deficits. Initial sounds are administered to beginning kindergarteners. Additionally, kindergarteners and 1st grade students are administered letter naming, phoneme segmentation, and nonsense word fluency. First grade students are also administered oral reading fluency. The oral reading fluency and retell subscales are administered in 2nd through 6th grade. Student performance is measured by predetermined benchmarks established by scientifically-based criterion. Those who read below grade level are measured at their appropriate reading level. Data from the DIBELS is used to direct teacher instruction and individualize interventions for struggling students (www.dibels.uorgeon.edu).

Research supporting DIBELS. The formative value of the DIBELS has been linked to performance on high-stakes, state assessments. Criterion validity examined the extent to which the DIBELS predicted performance on the reading portion of various state assessments. Shaw and Shaw (2002) conducted research to determine the criterion validity of the DIBELS for the Colorado State Assessment Program (CSAP). For 3rd grade students, Shaw and Shaw (2002) concluded that the DIBELS had high criterion validity in relation to the CSAP with correlation coefficients ranging from .73 (fall and winter administrations) to .80 (spring administration). Barger (2003) examined criterion validity of the DIBELS in relation to the North Carolina End of Grade reading assessment.

Likewise, Barger (2003) found that oral reading fluency performance for the DIBELS was significant with a correlation coefficient of .73 for 3rd grade students. Buck and Torgeson (2002) conducted similar research with the criterion-referenced reading

Florida Comprehensive Assessment Test – Sunshine State Standards (FCAT-SSS) and norm referenced test (FCAT-NRT). Not surprisingly, the DIBELS oral reading fluency subscale was highly predictive of the reading FCAT-SSS ($r=.70$, $p<.001$) and FCAT-NRT ($r=.74$, $p<.001$). Additionally, the DIBELS was highly correlated with later performance on the Arizona Instrument to Measure Standards (Wilson, 2005) and the Ohio Proficiency Test in reading (Meer, Lentz, & Stollar, 2005). Extensive research has established oral reading fluency as a reliable and predictive measure of reading performance and outcomes on high-stakes, state assessment (Good, Simmons, & Kame'enui, 2001).

Utah Criterion-Referenced Tests (CRTs)

As part of the Utah Core Assessment Program (UCAP) and the Utah Performance Assessment System for Students (U-PASS), the Utah Core Assessment Criterion-Referenced Tests (CRTs) were originally designed to meet accountability provisions for state core curriculum. Utah first adopted a core curriculum as part of the implementation of state graduation requirements in 1984. Brett Moulding, a member of the Utah Educational Advisory Committee, defined the Core as

> …content knowledge and skills for all children. It is a set of minimum standards (the words "core" and "standards" are often interchanged) for each grade level. The Core consists of a set of standards and objectives that describe what students should know and be able to do. The Core describes the intended learning outcomes for instruction. (Educational Advisory Committee, 2007, p. 1)

Previously, local school and districts maintained significant flexibility in teaching standards and evaluating student performance. In 1990, the Utah State Office of

Education (USOE) revised the Utah Core Curriculum to regulate classroom instruction and standardize objectives and goals throughout Utah (Educational Advisory Committee, 2007). The curriculum established content expectations for grade K-12, aligned with National Association of Education Progress (NAEP).

Minutes of the Utah Educational Advisory Committee summarized Moulding's comments as follows:

> The process of core development first involves gathering input from the various stakeholders (teachers, administrators, parent groups, districts, state office of education, universities, professional organizations, informal education organizations, and experts in the field). The research base for the Core comes from national standards in the subject area. The Core must go through an open public hearings process to gather input from the public. The final step is consideration and approval by the State School Board. Once approved the core is implemented statewide. (Education Advisory Committee, 2007, p. 1)

Once a standardized state core curriculum was established, a standardized state assessment was developed to evaluate the core curriculum's effectiveness on student learning and performance (D. Smith, Personal Communication, January 16, 2008). The USOE initiated testing development and improvement in 1985 with research organizations and elementary, secondary, and post-secondary schools in Utah. The development and refining period was ongoing and is still considered ongoing as curriculum is improved. The NAEP also conducted state-by-state national assessment of both mathematics and reading nationwide in 1992. Participating states provided a

normative sample for performance, in which Utah was ranked 12th among 44 states for 4th grade (Nelson & Lawrence, 1994).

The Utah CRT was part of test development and modification, and in 1991the test was piloted in various districts across Utah. The primary purpose for the Utah CRT was to evaluate the core curriculum and meet state accountability provisions (D. Smith, personal communication, January 16, 2008). Nelson and Lawrence (1994) reported that "these tests are developed with great technical precision and are field tested at least three times. The end-of-level and end-of-course tests have two major purposes: first, they provide a final check on student attainment of core curriculum content; second, they help document program strengths and weaknesses" (p. 8).

Evaluation of Utah's CRT

Utah has conducted a vast selection of technical evaluations by state and local officers and psychometric contractors to ensure that state tests align with state content standards and instruction (D. Smith, personal communication, January 16, 2008). WestEd, a non-profit agency dedicated to educational assessment and accountability and program evaluation, is among the contractors evaluating Utah's curriculum and standards, providing feedback on alignment of assessment to the State Core, technical quality, and assessment utility. Utah utilizes the feedback to refine the curriculum, standards, and assessment to improve the state accountability system (WestEd, 2001). In addition, the USOE aims to align state content standards with national standards.

In 2000, the U-PASS legislation required all Utah school districts to provide annual report of assessments and state accountability plans. Likewise, NCLB holds schools accountable to "ensure that all public school students have access to a high-

quality and challenging education and become proficient in the core academic subjects of reading/language arts, mathematics, and science…" (U.S. Department of Education, 2003) Utah uses the CRT to meet achievement standards and accountability provisions for both U-PASS and NCLB.

Utah's CRT for Language Arts

The English Language Arts CRT (ELA-CRT) is the state assessment for receptive and expressive language, reading and spelling, vocabulary, comprehension, and writing in 2nd through 11th grades. The assessment evaluates core curriculum standards through text passages which require elementary-aged students to determine semantics and syntax, exhibit comprehension and problem solving skills, and demonstrate basic and persuasive writing skills. Students' scores are categorized in one of four levels of proficiency with a numerical value from 1 to 4 from 1 = minimal proficiency, 2 = partial proficiency, 3 = sufficient proficiency, and 4 = substantial proficiency (Utah State Office of Education, 2007). Students must earn a proficiency score of 3 or 4 in order to achieve a score of proficiency for U-PASS and NCLB.

Validity and Reliability of Utah's ELA-CRT

The Utah ELA-CRT undergoes a continually rigorous process to ensure evidence of validity and reliability and assessment utility (D. Smith, personal communication, January 16, 2008). The ELA-CRT was developed by a team of educators and administrators who utilized the Standards for Educational and Psychological Testing published by the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education (Utah Office of Education, 2007).

Functions of the ELA-CRT. In terms of U-PASS and NCLB, the two specific functions of the ELA-CRT are, first, "to provide evaluative information across public and academic domains (i.e., to the general public, the Utah Legislature, the State Board of Education, school districts, public schools, and school teachers) about academic [reading] proficiency so schools can design, assess, and evaluate the success of public school programs. Second, the ELA-CRT results are used to identify schools that are performing exceptionally and those needing assistance (additional resources) from the state for academic improvement" (Technical Report, p. 2).

Comparisons among the ELA-CRT and other assessments. The ELA-CRT undergoes continual modification, as additional items and tests are piloted for improved assessment. In 2000, additional psychometric data was collected for validity measures. Correlational analysis examined scaled scores on both the Utah ELA-CRT and State norm-referenced test (NRT), Stanford Achievement Test-SAT-9 for language arts in grades 3, 5, 8, and 11. Correlation coefficients ranging from .74 to.83 indicated that CRT scaled scores are measuring similar content. Strong convergent validity allowed for generalization to other non-sampled grades. Student performance was also compared to demographic characteristics such as social economic status, gender, migration, accommodations, English Language Learners (ELL), and ethnicity. The correlations between the ELA-CRT and the demographic characteristics were generally low (.10 - .20), indicating that student performance is generally independent of those student characteristics.

Internal consistency of ELA-CRT. In addition, internal consistency was examined across all grades and subgroups (including, desegregation by race, socioeconomic status,

gender, migration status, educational accommodation status, and language status). The Chronbach alpha coefficients ranged from +.79 to +.95 (.92. omnibus).The statistical analysis reports that the ELA-CRT is positively correlated to outside variables of academic achievement (i.e. NRT scaled scores) and poorly correlated to non-academic variables (i.e. demographics), indicating a reliable and robust measure of summative performance in language arts and reading skills (reference).

Purpose of the Study

Useful description of reading ability and achievement requires educators to monitor student progress in the multifaceted aspects of reading ability. Formative assessment for progress monitoring is especially important for students at risk for reading failure to prevent further deficits and delays in literacy achievement. Ultimately, the goal of formative assessment is to provide feedback that informs specific intervention to assist in mastery learning objectives and improving test scores. While research has validated its use to monitor student progress and predict performance high-stakes testing, no studies document the use of DRA and DIBELS for predicting student performance on the Utah ELA-CRT. The purpose of this study is to determine the correlation between DRA and DIBELS and student achievement on the Utah ELA-CRT. Based on previous research, the demographic variables were not identified as factors that strengthened prediction of performance; however, the interest of the study was to determine the degree to which these variables affected performance on the ELA-CRT. The study addresses the following questions:

1. To what degree do the scores on the DRA predict performance on the ELA-CRT for first and second grade students in an urban Utah school district,

controlling for ethnicity, income status, English language proficiency, and special education status?

2. To what degree do the scores on the DIBELS predict performance on the ELA-CRT for first and second grade students in an urban Utah school district, controlling for ethnicity, income status, English language proficiency, and special education status?

METHODS

Setting and Participants

The present study provides demographical and student performance data for both 1st and 2nd grade students in the Salt Lake City school district. A total of 2931 students in a 1st grade sample and 3018 students in the 2nd grade sample completed the DIBELS. A total of 1547 students in a 1st grade sample and 1497 students in a 2nd grade sample completed the DRA. The data used for this study were collected from Fall 2005 through Spring 2006. The participants in all groups included both general education and special education students. The samples were split 50% between male and female students in the DRA population and 48% to 52% for male and female students, respectively. The DIBELS samples were predominantly composed of ethnic minorities while the DRA samples were split evenly between Caucasian students and ethnic minority students. All student samples were approximately 50% free and reduced lunch, low income status. The Limited English Proficiency students ranged from 30-40% among the DIBELS and DRA samples. Table 1 illustrates specific samples broken down by gender, ethnicity, low-income status, English language proficiency, and special education status.

In describing the participants in the current study, it is important to understand how the composition of the sample populations compare to a larger context. By way of comparison, the National Center of Educational Statistics (NCES) reported that minorities accounted for 42% of the population nationwide and 17.3% for Utah in the 2003-2004 school year. In the 2005-2006 school year, in urban areas nationwide, 54.9% of students qualified for free and reduced lunch, and 45.8% in Utah. In addition, 25% of students spoke a language other than English at home and/or had limited English

Table 1

Demographic Information: Study Samples for DIBELS & DRA

Descriptive Information	Grade 1 DIBELS	Grade 2 DIBELS	Grade 1 DRA	Grade 2 DRA
Sample Population	n = 2931	n= 3018	n = 1547	n = 1497
Gender				
Male	49	50	48	49
Female	51	50	52	51
Ethnicity				
Caucasian	39	38	53	54
Hispanic	37	36	34	32
African American	5	4	4	3
Pacific Islander	5	5	4	4
Asian	4	4	4	5
American Indian	1.4	2	1	2
Unknown	< 1	< 1	< 1	< 1
LEP	38	40	29	30
Special Education	10	15	9	14
Low Income	56	56	49	47

proficiency nationwide. In the 2003-2004 school-year, the percentage of students with disabilities enrolled in U.S. public schools was reportedly 13.7% with Utah at 11.6% (NCES, 2007). The Utah samples in the current study included a moderately higher percentage of ethnic minorities, 50-70%, than the nationwide sample and a significantly higher percentage than the Utah population. Limited English proficiency was slightly higher in the current sample than the national average of 25% compared to approximately 30-40% limited English proficient. In addition, students taking the DIBELS who were from low income backgrounds (approximately 55% of all students, those qualifying for free and reduced lunch) were comparable to the national sample, while students taking the DRA (approximately 48%) were comparable to the Utah sample. In sum, the sample in the present study contained a higher volume of students from minority and low-income backgrounds, which provide valuable information pertaining to utility of the study with historically disadvantaged populations.

Measures and Procedures

Three measures of student reading performance were used in the present study: (a) oral reading fluency performance for the DIBELS (b) oral reading fluency performance for the DRA, 1st edition (c) proficiency on the ELA-CRT. DRA oral reading fluency data was reported for the fall term of 2005. DIBELS oral reading fluency data was reported for the fall term of 2005 and the winter and spring terms of 2006.ELA-CRT proficiency levels were reported for the spring of 2006.

The DIBELS oral reading fluency subtest was administered at the beginning, middle, and end of the 2005-2006 school year for second grade students and middle and end of the school year for first grade students. Students were administered a 1-minute

grade-level reading probe. The teacher noted specific mistakes, such as syntax, pronunciation, and omitted words, and recorded the number of words read correctly. The schools used benchmarks directed by DIBELS research.

The DRA oral reading fluency subtest was administered at the beginning of the 2005-2006 school year for first and second grade students. Teachers chose a grade-level text in which students were asked to read. Teachers recorded the number of words per minute read correctly by the student and noted specific observations, such as semantics, syntactical, graphophonic errors. The schools used benchmarks directed by DRA research.

The ELA-CRT was administered in the spring of 2006, as an end-of-year testing of the core curriculum for the 2005-2006 school year. Students received the test through oral and paper/pencil administration. Some portions of the test were read orally, and students were directed to mark the correct answer on the paper. Other portions of the test required students to define vocabulary and read passages with comprehension questions. Administration was standardized, and testing for each portion was timed.

Statistical Analysis

The current interest of the study is to extend research for formative and predictive value of curriculum-based assessment for high-stakes testing used as federal and statewide accountability measures. Descriptive statistics outline the data used in the statistical analysis. A correlation matrix depicts the associations between the independent variables (i.e. demographics, DRA and DIBELS performance) and the dependent variable (ELA-CRT). A multiple linear regression was conducted to evaluate the predictive statistics of the DRA and the DIBELS with student performance on the ELA-CRT.In

addition, other independent variables, such as gender, ethnicity, English language proficiency, and economic status are analyzed for impact on student performance on the ELA-CRT. The 1st model for the regression analysis includes the demographical student data in terms of their predictive value and impact on ELA-CRT performance. The 2nd model for the regression analysis includes DRA or DIBELS performance for 1st and 2nd grades in conjunction with demographical student data. The results and implications of these analyses will be discussed subsequently.

Due to the large number of participants involved in this study, it was highly probable that the traditional level of statistical significance ($p < .05$) would greatly underestimate the practical significance of the result. Moreover, multiple analyses were conducted, such that the likelihood of obtaining statistically significant results was artificially inflated based on the increased likelihood of obtaining values below $p < .05$. Therefore, in the present study, the level of statistical significance was set at $p < .001$. In interpreting the results, it is also important to attend to the magnitude of the associations observed in the data. Correlation coefficients and beta weights below absolute value 0.1 generally indicate a very weak relationship, even if the analysis proves to reach statistical significance. Hence, the analyses conducted in this thesis will attend to the magnitude of the association more than to the level of statistical significance.

RESULTS

Descriptive Statistics for Predictors and Criterion Variables

Descriptive statistics for Grades 1 and 2 of the DIBELS and DRA and end-of-year administration of the ELA-CRT are shown in Table 2. Statistical normality was reflected in the DRA student populations, indicating the data approached the expected normal distribution. DIBELS student populations were less likely to concentrate around the mean, indicating a flatter and more widely distributed performance range. Mean values for each assessment indicate the general level of performance for that student population. For the ELA-CRT population, students' scores were categorized in one of six levels of proficiency with a numerical value from 1 to 6 from: 1 = minimal proficiency/bottom half; 2 = minimal proficiency/top half; 3 = partial proficiency/bottom half; 4 = partial proficiency/top half; 5 = sufficient proficiency; and 6 = substantial proficiency. In order to obtain proficiency, students must earn a score of 5 or 6. ELA-CRT students scored generally between partial proficiency/top half (4) and sufficient proficiency (5), indicating that many students obtained proficiency. The median value (5) indicates that the most frequently occurring score was sufficient proficiency.

DIBELS ORF scores were placed in (a) at risk, (b) some risk, and (c) low risk areas, based on each student's words read per minute. Mean values indicate that Grade 1 students were generally between the some to low risk ranges; whereas, more Grade 2 students performed in the at-risk to some risk range.

DRA scores ranged from levels 1 - 44, with specific grades assigned to a range of levels. According to DRA, levels 14 - 16 are considered Grade 1 range, and levels 18 - 28 are considered Grade 2 range; therefore, both Grades 1 and 2 mean values show that

students were generally within expected benchmarks for their respective grades. Although the DIBELS data indicated a platykurtic distribution, the following analyses were justified through the proposed regression models without additional statistical adjustments.

Table 2

Descriptive Statistics for Predictors and Criterion Variables

Variables	Mean	SD	Skewness	*Skew* SD	Kurtosis	*Kurtosis* Mean
ELA-CRT	4.5	1.6	- 1.02	.02	- .21	.05
Grade 1 DIBELS ORF, mid.	2.25	.08	- 4.8	.05	- 1.30	.09
Grade 2 DIBELS ORF, beg.	2.06	.89	- .12	.05	- 1.71	.10
Grade 1 DRA	15.97	7.67	.16	.06	.20	.12
Grade 2 DRA	26.03	8.7	- .85	.06	1.02	.13

Correlations Between and Among Predictors and Criterion Variables

Tables 3 and 4 report correlational analysis between performance on the Utah ELA-CRT and curriculum-based assessments, DRA and DIBELS, revealing significant correlations between CBM and high-stakes testing in Utah. In Table 3, correlations of control variables for the Grade 1 mid-year and Grade 2 beginning administrations of the DIBELS are presented. Pearson correlations between Hispanic, Black, and Asian students and performance on the ELA-CRT were significant for both Grade 1 and Grade 2. Gender had a significant correlation for Grade 1 (r= .086) and Grade 2 (r=.119), indicating that female students were slightly more likely to score proficiently on the

ELA-CRT. Students who had low English language proficiency (LEP) (Grade 1, *r*=-.369; Grade 2, *r*= -.306) and/or were classified in special education (Grade 1, *r*=-.214; Grade 2, *r*= -.273) were less likely to perform in the proficient range. In addition, students with low-income status were less likely to score proficiently (Grade 1, *r*=-.379; Grade 2, *r*= -.371). Grade 1 DIBELS ORF mid-year (*r*= .687) and Grade 2 DIBELS ORF (*r*=.644) beginning year resulted in a significant correlation, indicating that students who met expected benchmarks for the DIBELS also attained proficiency on the ELA-CRT.

Table 4 presents fall administration of the DRA for Grades 1 and 2 and correlations of control variables with end-of-year performance on the ELA-CRT. Pearson correlations between Hispanic, Black, and Asian students and performance on the ELA-CRT were significant for both Grade 1 and Grade 2. Gender was not a significant correlation for Grade 1 or Grade 2. Students who had low LEP (Grade 1, *r*=-.386; Grade 2, *r*= -.327) and/or were classified in special education (Grade 1, *r*=-.145; Grade 2, *r*= -.213) were less likely to perform in the proficient range. In addition, students with low-income status were less likely to score proficiently (Grade 1, *r*=-.357; Grade 2, *r*= -.425). Grade 1 DRA (*r*= .699) and Grade 2 DRA (*r*=.785) fall administrations resulted in a strong and statistically significant correlation, indicating that students who met expected benchmarks for the DRA also attained proficiency on the ELA-CRT.

Table 3

Pearson Correlations with Proficiency on the ELA-CRT: Control Variables for Grade I DIBELS ORF(Mid-year) and Grade 2 DIBELS ORF(Beginning of year)

Variables	DIBELS Grade 1 (n = 2,482)	DIBELS Grade 2 (n = 2,073)
Ethnicity		
Hispanic	-.331***	-.275***
Black	-.102***	-.157***
American Indian	-.037	-.021
Pacific Islander	.001	.006
Asian	.063**	.112***
Gender	.086***	.119***
LEP	-.369***	-.306***
Special Education Status	-.214***	-.273***
Low Income Status	-.379***	-.371***
ELA-CRT	.687***	.644***

Note. * = $p < .05$; ** = $p < .01$; *** = $p < .001$

Table 4

Pearson Correlations:
Control Variables for DRA Grade 1 and Control Variables for DRA Grade 2

Variables	DRA Grade 1 (n = 1,539)	DRA Grade 2 (n = 1,495)
Ethnicity		
Hispanic	-.383***	-.338***
Black	-.073**	-.134***
American Indian	-.021	-.059*
Pacific Islander	-.016	-.001
Asian	.077**	.104***
Gender	.043	.048
LEP	-.386***	-.327***
Special Education Status	-.145***	-.213***
Low Income Status	-.357***	-.425***
ELA-CRT	.699***	.785***

Note. * = $p < .05$; ** = $p < .01$; *** = $p < .001$

Regression of Predictors for Performance on the ELA-CRT

Tables 5 through 8 contain the results of multiple linear regression models measuring the extent to which each of the five control variables and the DRA and DIBELS for Grades 1 and 2 impacted performance on the ELA-CRT and the amount of variance on the ELA-CRT which were accounted for by these variables. Table 5 provides data for mid-year Grade 1 DIBELS ORF and performance on the ELA-CRT. Model 1 had an adjusted R^2 of .256 (p= <.001), demonstrating that the demographic variables significantly affect and predict performance on the ELA-CRT. Beta weights and their corresponding t-values were statistically significant for Hispanic and Black students, as well as LEP, special education, and low income students, indicating that these students are less likely to attain proficiency on the ELA-CRT. Model 2 indicates that when adding the Grade 1 DIBELS ORF to the statistical model, the association between Hispanic and Black students, and special education and low income status no longer reach statistical significance, showing that the Grade 1 DIBELS ORF accounts for variations in subsequent slopes even more than do student characteristics.

Table 6 provides data for beginning Grade 2 DIBELS ORF and performance on the ELA-CRT. Model 1 had an adjusted R^2 of .274 (ΔP= <.001), demonstrating that the demographic variables significantly affect and predict performance on the ELA-CRT. Beta weights and their corresponding t-values were statistically significant for Hispanic and Black students, as well as LEP, special education, and low income students, indicating that these students are less likely to attain proficiency on the ELA-CRT. Model 2 indicates that when adding the Grade 2 DIBELS ORF in the statistical model, the association between Hispanic and Black students, and special education and low income

Table 5

Regression Models Predicting End of Year ELA-CRT Proficiency from Demographic Variables and Mid-year Grade 1 DIBELS ORF (n = 2,581)

Variables	Adjusted R^2	ΔP	B	*t*	*p*
MODEL 1	.256	<.001			
Ethnicity					
Hispanic			-.163	-6.3	<.001
Black			-.121	-6.4	<.001
American Indian			-.054	-3.0	.002
Pacific Islander			-.025	-.1	.173
Asian			.030	1.7	.098
Gender			.058	3.3	.001
LEP			-.187	-8.0	<.001
Special Education Status			-.199	-11.3	<.001
Low Income Status			-.189	-8.9	<.001
MODEL 2	.512	<.001			
Ethnicity					
Hispanic			-.079	-3.7	<.001
Black			-.075	-4.9	<.001
American Indian			-.018	-1.2	.221
Pacific Islander			-.042	-2.8	.005
Asian			-.003	-.2	.853
Gender			.028	2.0	.046
LEP			-.108	-5.6	<.001
Special Education Status			-.079	-5.4	<.001
Low Income Status			-.056	-3.2	.002
Grade 1 DIBELS ORF			.582	36.1	<.001

Table 6

Regression Models Predicting End of Year ELA-CRT Proficiency from Demographic Variables and Beginning Grade 2 DIBELS ORF (n = 2,072)

Variables	Adjusted R^2	ΔP	B	*t*	*P*
MODEL 1	.274	<.001			
Ethnicity					
Hispanic			-.107	-3.7	<.001
Black			-.148	-7.2	<.001
American Indian			-.036	-1.9	.060
Pacific Islander			-.020	-1.0	.338
Asian			.070	3.5	.001
Gender			.087	4.6	<.001
LEP			-.163	-6.4	<.001
Special Education Status			-.270	-14.2	<.001
Low Income Status			-.223	-9.7	<.001
MODEL 2	.481	<.001			
Ethnicity					
Hispanic			-.083	-3.4	.001
Black			-.114	6.5	<.001
American Indian			-.031	-1.9	.057
Pacific Islander			-.044	-2.5	.011
Asian			.022	1.3	.193
Gender			.056	3.5	<.001
LEP			-.090	-4.2	<.001
Special Education Status			-.152	-9.2	<.001
Low Income Status			-.088	-4.4	<.001
Grade 2 DIBELS ORF			.514	28.7	<.001

status no longer reach statistical significance, showing that the Grade 2 DIBELS ORF accounts for some variations in subsequent slopes even more than do student characteristics; however, in this sample, special education status remained a factor for ELA-CRT, in the presence of the Grade 2 DIBELS ORF.

Table 7 provides data for beginning Grade 1DRA and performance on the ELA-CRT. Model 1 had an adjusted R^2 of .241 (Δ P= <.001), demonstrating that the demographic variables significantly affect and predict performance on the ELA-CRT. Beta weights and their corresponding t-values were statistically significant for Hispanic and Black students, as well as low LEP, special education, and low income students, indicating that these students are less likely to attain proficiency on the ELA-CRT. Model 2 indicates that when adding the Grade 1 DRA to the statistical model, the association between Hispanic and Black students, and special education and low income status no longer reach statistical significance, showing that the Grade 1 DRA accounts for variations in subsequent slopes even more than do student characteristics. Students with low LEP remain less likely to attain proficiency on the ELA-CRT.

Table 8 provides data for beginning Grade 2 DIBELS ORF and performance on the ELA-CRT. Model 1 had an adjusted R^2 of .290 (Δ P= <.001), demonstrating that the demographic variables significantly affect and predict performance on the ELA-CRT. Beta weights and their corresponding t-values were statistically significant for Hispanic and Black students, as well as LEP, special education, and low income students, indicating that these students are less likely to attain proficiency on the ELA-CRT.

Table 7

Regression Models Predicting End of Year ELA-CRT Proficiency from Demographic Variables and Beginning Grade 1 DRA (n =1,550)

Variable	Adjusted R^2	ΔP	B	*t*	*p*
MODEL 1	.241	<.001			
Ethnicity					
Hispanic			-.208	-6.3	<.001
Black			-.100	-4.4	<.001
American Indian			-.033	-1.5	.148
Pacific Islander			-.030	-1.3	.190
Asian			.036	1.5	.123
Gender			.019	.9	.397
LEP			-.189	-6.2	<.001
Special Education Status			-.144	-6.4	<.001
Low Income Status			-.163	-6.1	<.001
MODEL 2	.530	<.001			
Ethnicity					
Hispanic			-.103	-3.9	<.001
Black			-.058	-3.2	.001
American Indian			.005	.3	.762
Pacific Islander			-.039	-2.1	.033
Asian			-.002	-.1	.903
Gender			.007	.4	.684
English Lang. Proficiency			-.097	-4.0	<.001
Special Education Status			-.044	-2.5	.014
Low Income Status			.005	.2	.807
Grade 1 DRA			.628	30.8	<.001

Table 8

Regression Models Predicting End of Year ELA-CRT Proficiency from Demographic Variables and beginning Grade 2 DRA (n = 1,497)

Variables	Adjusted R^2	ΔP	β	*t*	*P*
MODEL 1	.290	<.001			
Ethnicity					
Hispanic			-.127	-3.7	<.001
Black			-.144	-6.3	<.001
American Indian			-.061	-2.7	.007
Pacific Islander			-.002	-.1	.928
Asian			.076	3.2	.001
Gender			.038	1.7	.086
LEP			-.159	-5.1	<.001
Special Education Status			-.225	-10.1	<.001
Low Income Status			-.273	-10.3	<.001
MODEL 2	.638	<.001			
Ethnicity					
Hispanic			-.044	-1.8	.075
Black			-.038	-2.3	.023
American Indian			-.028	-1.7	.084
Pacific Islander			-.001	-.1	.948
Asian			.018	1.1	.294
Gender			.015	1.0	.330
LEP			-.031	-1.4	.175
Special Education Status			-.072	-4.4	<.001
Low Income Status			-.095	-4.8	<.001
Grade 1 DRA			.698	37.8	<.001

Note. p<.001

Model 2 indicates that when adding the Grade 2 DRA in the statistical model, the association between Hispanic and Black students, and special education and low income status no longer reach statistical significance, showing that the Grade 2 DRA accounts for variations in subsequent slopes even more than do student characteristics. Students with low LEP and special education status remain less likely to attain proficiency on the ELA-CRT.

Supplementary Analyses

Descriptive Statistics for Supplementary Predictor Variables

While primary analyses focused on the most formative assessment of reading ability (Grade 1 mid-year and Grade 2 beginning administrations) and its predictive relationship to latter ELA-CRT performance, supplementary analyses help to validate utility of the DIBELS as formative and cross-sectional assessment for performance on summative, high-stakes testing. Descriptive statistics for supplementary and cross-sectional analyses including additional benchmarks for Grades 1 and 2 of the DIBELS are shown in Table 5. Mean values for each assessment indicate the general level of performance for that student population.

DIBELS ORF was placed in (a) at risk, (b) some risk, and (c) low risk areas, based on each student's words read per minute. Mean values indicate that Grade 1 end-year students were generally between the some to low risk ranges; and Grade 2 mid-year students were in some to low risk ranges while the end-year students were generally in the some risk range. DIBELS data indicated a platykurtic distribution; however, proposed regression models were considered appropriate without additional statistical adjustments.

Table 9

Descriptive Statistics for Supplementary Predictor Variables

Variables	Mean	SD	Skewness	Skew SD	Kurtosis	Kurtosis Mean
Grade 1 DIBELS ORF, End	2.20	.86	-.39	.05	-1.54	.09
Grade 2 DIBELS ORF, Mid.	2.17	.91	-.35	.05	-1.71	.10
Grade 2 DIBELS ORF, End	2.05	.91	-.10	.05	-1.8	.09

Correlations Between and Among Supplementary and Criterion Variables

Table 10 reports correlational analysis between performance on the Utah ELA-CRT and the Grade 1end-year and Grade 2 mid and end-year administrations of the DIBELS, revealing significant correlations between supplementary and cross-sectional analyses of CBM and high-stakes testing in Utah. Pearson correlations between Hispanic, Black, and Asian students and performance on the ELA-CRT were significant for both Grade 1 end-year and Grade 2 mid and end-year. Gender had a correlation for benchmarks in both Grades 1 and 2, indicating that female students were slightly more likely to score proficiently on the ELA-CRT. Students who had low English language proficiency and/or were classified in special education were less likely to perform in the proficient range. In addition, students with low-income status were less likely to score proficiently. Grade 1 DIBELS ORF end-year and Grade 2 DIBELS ORF mid and end-year performances resulted in a significant correlation, indicating that students who met expected targets for the DIBELS in both Grades 1 and 2 at mid and end-year benchmarks also attained proficiency on the ELA-CRT.

Table 10

Pearson Correlations with Proficiency on the ELA-CRT
Control Variables for Grade I DIBELS ORF, End-year; Grade 2 DIBELS ORF, Middle; and Grade 2 DIBELS ORF, End-year

Variables	DIBELS Grade 1 (n = 2,482)	DIBELS Grade 2 (n = 2,073)	DIBELS Grade 2 (n = 2610)
Ethnicity			
Hispanic	-.334***	-.290***	-.304***
Black	-.107***	-.159***	-.158***
American Indian	-.027	-.031	-.037
Pacific Islander	.008	<.001	.004
Asian	.070***	.101***	.101***
Gender	.069***	0.91***	.078***
LEP	-.361***	-.308***	-.319***
Special Education Status	-.200***	-.265***	-.244***
Low Income Status	-.363***	-.396***	-.400***
ELA-CRT	.702***	.670***	.654***

Note. * = $p < .05$; ** = $p < .01$; *** = $p < .001$

Regression of Predictors of Performance on the ELA-CRT, Supplementary Analyses

Tables 11-13 contain the results of multiple linear regression models measuring the extent to which each of the five control variables for Grades 1 end-year and 2 mid and end-year DIBELS impacted performance on the ELA-CRT and the amount of variance on the ELA-CRT which were accounted for by these variables. Table 11 provides data for end-year Grade 1 DIBELS ORF and performance on the ELA-CRT. Model 1 had an adjusted R^2 of .246 (ΔP= <.001), demonstrating that the demographic variables significantly affect and predict performance on the ELA-CRT. Beta weights and their corresponding t-values were statistically significant for Hispanic and Black students, as well as LEP, special education, and low income students, indicating that these students are less likely to attain proficiency on the ELA-CRT. Model 2 indicates that when adding the Grade 1 DIBELS ORF end-year to the statistical model, the association between Hispanic and Black students, and low income status no longer reach statistical significance, showing that the Grade 1 DIBELS ORF accounts for variations in subsequent slopes even more than do student characteristics. Students with low LEP and special education status remain at risk in their ability to attain proficiency for the ELA-CRT.

Table 12 provides data for beginning Grade 2 DIBELS ORF and performance on the ELA-CRT. Model 1 had an adjusted R^2 of .283 (ΔP= <.001), demonstrating that the demographic variables significantly affect and predict performance on the ELA-CRT. Beta weights and their corresponding t-values were statistically significant for Hispanic and Black students, as well as LEP, special education, and low income students, indicating that these students are less likely to attain proficiency on the ELA-CRT.

Table 11

Regression Models Predicting End of Year ELA-CRT Proficiency from Demographic Variables and End-year Grade 1 DIBELS ORF

Variables	Adjusted R^2	Δ P	β	*t*	*P*
MODEL 1	.246	<.001			
Ethnicity					
Hispanic			-.193	-7.5	<.001
Black			-.138	-7.4	<.001
American Indian			-.050	-2.9	.004
Pacific Islander			-.028	-1.5	.126
Asian			.030	1.6	.103
Gender			.046	2.7	.008
LEP			-.177	-7.7	<.001
Special Education Status			-.194	-11.2	<.001
Low Income Status			-.166	-7.9	<.001
MODEL 2	.535	<.001			
Ethnicity					
Hispanic			-.107	-5.2	<.001
Black			-.084	-5.8	<.001
American Indian			-.013	-.9	.350
Pacific Islander			-.047	-3.3	.001
Asian			-.011	-.7	.457
Gender			.017	1.3	.211
LEP			-.112	-6.2	<.001
Special Education Status			-.064	-4.5	<.001
Low Income Status			-.034	-2.0	.044
Grade 1 DIBELS ORF			.608	40.0	<.001

Note. n = 2,581

Table 12

Regression Models Predicting End of Year ELA-CRT Proficiency from Demographic Variables and mid-year Grade 2 DIBELS ORF (n = 2,467)

Variables	Adjusted R^2	Δ P	β	*t*	*p*
MODEL 1	.283	<.001			
Ethnicity					
Hispanic			-.121	-4.6	<.001
Black			-.154	-8.3	<.001
American Indian			-.045	-2.6	.009
Pacific Islander			-.020	-1.1	.285
Asian			.059	3.2	.001
Gender			.071	4.1	<.001
LEP			-.148	-6.4	<.001
Special Education Status			-.264	-15.3	<.001
Low Income Status			-.244	-11.5	<.001
MODEL 2	.506	<.001			
Ethnicity					
Hispanic			-.081	-3.7	<.001
Black			-.106	-6.8	<.001
American Indian			-.029	-1.9	.047
Pacific Islander			-.025	-1.6	.102
Asian			.007	.44	.661
Gender			.043	2.9	.003
LEP			-.068	-3.5	<.001
Special Education Status			-.136	-9.1	<.001
Low Income Status			-.110	-6.1	<.001
Grade 2 DIBELS ORF			.540	33.3	<.001

Note. p <.001

Model 2 indicates that when adding the Grade 2 DIBELS ORF to the statistical model, the association between Hispanic and Black students, and low income status no longer reach statistical significance, showing that the Grade 2 DIBELS ORF accounts for some variations in subsequent slopes even more than do student characteristics; however, once again, LEP and special education status remained an impact for ELA-CRT, in the presence of the Grade 2 DIBELS ORF.

Table 13 provides data for beginning Grade 1DRA and performance on the ELA-CRT. Model 1 had an adjusted R^2 of .281 ($\Delta P = <.001$), demonstrating that the demographic variables significantly affect and predict performance on the ELA-CRT. Beta weights and their corresponding t-values were statistically significant for Hispanic and Black students, as well as low LEP, special education, and low income students, indicating that these students are less likely to attain proficiency on the ELA-CRT. Model 2 indicates that when adding the Grade 1 DRA to the statistical model, the performance of Hispanic and Black ethnicity, and low income status improve meaningfully, showing that the Grade 1 DRA accounts for variations in subsequent slopes even more than do student characteristics. However, in this sample, LEP and special education students remain less likely to attain proficiency on the ELA-CRT.

Table 13

Regression Models: Cross-sectional data for End of Year ELA-CRT Proficiency from Demographic Variables and End-year Grade 2 DIBELS (n = 2,609)

Variables	Adjusted R^2	Δ P	β	*t*	*p*
MODEL 1	.281	<.001			
Ethnicity					
Hispanic			-.139	-5.4	<.001
Black			-.156	-8.6	<.001
American Indian			-.059	-3.5	<.001
Pacific Islander			-.018	-1.0	.314
Asian			.056	3.1	.002
Gender			.058	3.5	.001
LEP			-.150	-6.6	<.001
Special Education Status			-.247	-14.7	<.001
Low Income Status			-.241	-11.7	<.001
MODEL 2	.498	<.001			
Ethnicity					
Hispanic			-.092	-4.2	<.001
Black			-.115	-7.5	<.001
American Indian			.033	-2.3	.021
Pacific Islander			-.035	-2.3	.022
Asian			.007	.50	.622
Gender			.035	2.5	.013
LEP			-.097	-5.1	<.001
Special Education Status			-.139	-9.6	<.001
Low Income Status			-.106	-5.7	<.001
Grade 1 DIBELS ORF			.523	33.5	<.001

Note. p<.001

DISCUSSION

The purpose of this study was to examine the predictive nature of two commonly used reading CBMs. Statistical analyses examined the extent to which the DIBELS 1st grade mid-year and 2nd grade beginning year scores predict performance on the Utah ELA-CRT. Analyses also examined the extent to which the DRA 1st and 2nd grade beginning year scores predict performance on the Utah ELA-CRT. In addition, the study investigated the degree to which the variables of student gender, ethnicity, English language proficiency, special education status, and low-income status affected performance on the Utah ELA-CRT. Supplementary analyses examined cross-sectional end-year DIBELS data for Grades 1 and 2 in comparison to the Utah ELA-CRT. Overall, the findings indicate that the scores for students of minority ethnicity, low LEP, special education, and low-income status were related to lower academic performance, which result is similar to previous research regarding students at risk of academic failure (Ding & Davison, 2004; Lewelling, 1991; Wang & Kovach, 1995). These results therefore provide continuing support for established efforts to decrease the reading achievement gap for minority students and students from low economic backgrounds, low LEP, and special education status. Thus, the current study confirms environmental and individual factors that contribute to the performance of disadvantaged students on high-stakes testing.

This study contributes to literature for formative assessment and high-stakes testing in several ways. It demonstrates an ability to predict a significant amount of variance in performance on the ELA-CRT based on brief oral reading fluency assessments administered during the Fall semester of the school year. Results from the

current study indicate that CBM proactively identified students at risk for poor performance on high-stakes testing. Results also demonstrate that CBM may help predict performance on end-of-year high-stakes testing as early as the beginning of the school year.

The statistically significant correlations between formative CBM and summative high-stakes testing in the current research align with previous studies. According to Sibley, Biwer, and Hesch (2001), students meeting established oral reading fluency benchmarks were likely to achieve proficiency on state standards testing. Moreover, oral reading fluency benchmarks have decision-making utility for students' eventual performance on global reading measures and high-stakes testing (Hintze & Silbertglitt, 2005). Correlations between DRA and DIBELS and the Utah ELA-CRT ranged from .64 to .79. These correlations were comparable to those between the DRA and ITBS, which ranged from .53 to .83 (Williams, 1999). Findings were also similar to correlations between the DIBELS and the CSAP (r = .73), North Carolina End of Grade (r = .73), and the FCAT-SSS (r = .70).

Few previous researchers have accounted for the influence of variables such as gender, ethnicity, English language proficiency, and low-income status (Buck & Torgeson, 2002; Wilson, 2005), despite the fact that these variables are widely known to be relevant to student performance. A literature search conducted by the author revealed no other research studies that explained variance on high-stakes testing after accounting for these relevant demographic variables. Hence, the current study makes a substantive contribution to research literature with additional analyses for these variables.

Even within a single urban school district, the demographics varied considerably across schools and student populations. A higher proportion of ethnic minority students were administered the DIBELS (61-62%) than were administered the DRA (46-47%). This may have impacted the correlation between DIBELS results and the ELA-CRT as compared to DRA results and the ELA CRT, especially if ethnicity implies within-child differences that are not controlled in the testing. For example, the number of students designated as LEP in the DIBELS sample was moderately higher than the DRA sample, indicating possible language differential factors affecting DIBELS and ELA-CRT performance. Moreover, even when controlling for demographics, half of the variance is still uncertain, indicating that several unknown factors affect performance on high-stakes testing. Other research indicates that these factors may include teacher characteristics (Hanushek, Kain, & Rivkin, 1998), parental support (Akimoff, 1996), and community resources (Kegler et al., 2005).

While correlations and regression statistics were statistically significant, it is necessary to emphasize that correlation does not reflect causation. More specifically, while the predictive variables, performance on the DIBELS and DRA, and demographic variables were highly correlated with performance on the ELA-CRT, correlations between these variables are associations; the predictive and demographic variables cannot be interpreted as direct causes of performance on the ELA-CRT. As previously mentioned, the regression analysis reflects only half the variance in end-of-year scores, indicating that half of the variability in performance on the ELA-CRT cannot be explained by the variables included in the current regression analyses.

Overall, the DIBELS and the DRA appear to be highly predictive of ELA-CRT scores, even when known moderating variables associated with demographic variables are included in the regression model. Moreover, in the presence of the DIBELS or DRA test scores, the association of these demographic variables with ELA-CRT scores becomes minimal, although the results were still statistically significant. This finding confirms that a child's reading ability tends to be stable over the course of an academic year, even when controlling for known factors associated with initial reading ability.

Limitations

Results of this study provide many useful insights regarding reading achievement; however, limitations of the study itself must be addressed. First, possible threats to the internal validity of the study involve the accuracy of the database, integrity of the test administration, and possible confounds related to differential participant selection and attrition. Accuracy of the data gathering process for the database is unknown, as well as attrition rate, data entry methods, and consideration for unknown factors previously mentioned. In addition, human error in data entry is likely minimal, but may affect data results. Integrity for test administration of the DIBELS, DRA, and the ELA-CRT is also unknown. It is assumed that trained professionals administered each respective test; however, this feature of the research is indiscernible, as the data was gathered prior to the current analysis. In addition, the previously mentioned unknown factors, specifically teacher characteristics, may have affected testing administration.

Second, two cohorts were used in the current study, indicating that formative assessment is useful to predict performance on high-stakes testing for early elementary

school grades; however, the data do not provide evidence for students in higher elementary grades.

Third, the results are limited in terms of their external validity. Caution is suggested in generalizing specific outcomes and implications to other urban populations. The overrepresentation of minority and low-income students, while providing useful information for this district, may skew the utility of the study outcomes for other populations. Factors such as ethnicity, English language proficiency, and low income status may have influenced many of the unknown variables which affect performance on summative high-stakes testing. Because of varying demographics for urban areas, cities and schools are encouraged to conduct their own research to better understand the influences of demographics specific to their location on assessment and testing.

In another way, this overrepresentation may actually be a strength of the current study. Students from minority, LEP, and low income backgrounds are more likely to be at risk for academic problems, so demonstrating that CBM can be used for predicting performance on high-stakes tests for these students gives districts and schools a useful tool for addressing these students' needs prior to end-of-level testing.

Implications

Implications for Practice

This study demonstrated the utility of both the DIBELS and the DRA to predict academic achievement on the Utah ELA-CRT while controlling for potential factors known to be associated with reading achievement, such as ethnicity and low-income status. The results of the study have particular import for schools and districts in that it provides educators ways to identify students at risk for inadequate or partial proficiency

on high-stakes tests. Educators should use the results of this study as an impetus for collecting data to guide instructional decisions and provide timely intervention for low-performing students. Educators should also examine what strategies are currently in place to address student needs and evaluate their current effectiveness, according to student assessment.

The outcomes of the current research also provide direct ways to save both time and money in the testing milieu. Schools do not need to juggle two or more types of formative assessment to monitor progress and predict high-stakes performance. Neither the DIBELS and nor the DRA appear biased by demographic factors in this population. With comparable evidence of reliability and validity, schools and districts may choose a single test (DIBELS or DRA, in this case), rather than several, to identify student needs and provide useful interventions for improved learning and performance. The use of several different assessments results in multiple data sources and interpretations, requiring more time for acquiring and analyzing data, as opposed to quicker interpretation and earlier intervention.

Implications for Future Research

Future research may extend the current research findings in several important directions. First, prospective research should focus on an equal number of benchmark assessments for both the DRA and DIBELS to provide a better data match between the two tests. The current study provided only one assessment of the DRA and three administrations of the DIBELS. While both tests show comparability to predict performance on the ELA-CRT, an equal number of assessments would facilitate more accurate analyses.

Second, this study did not control for previous or subsequent years' performance on the ELA-CRT and/or benchmark assessments. Consideration of previous scores on the ELA-CRT and benchmark assessments may indicate predictive power for performance in following years and specify even earlier indications of student deficits. Longitudinal data could decrease the number of benchmarks needed to provide the same results currently obtained with three or more benchmark administrations. If this were the case, time and money spent on the preparation and administration of benchmark assessments could be spent on effective interventions driven by such data. In addition, examination of the latter elementary grades (3rd-6th) would provide further predictive data.

A fourth suggestion and caution for further research involves the risk of biases caused by formative assessment. Neither teacher characteristics nor responses to testing results were analyzed in the current study. Once educators obtain early assessments, in what ways do those scores affect how and what teachers teach their students? While formative assessment is intended to inform instruction and intervention, early scores may also pigeon-hole students or establish teacher expectations for performance, thereby labeling students who are at risk of failing and inadvertently decreasing attention to student needs. Research shows the impact that teacher attitude and characteristics have on student performance. Factors such as self-fulfilling prophecy (Rosenthal & Jacobsen, 1968) and stereotype threats (Steele, 1997) depress student performance, especially if conveyed in strongly negative and consistent conditions. In addition, the degree to which teacher characteristics affect performance for this particular school district would inform administrators of needs for teacher development and ways to improve the learning environment.

This research emphasizes the need to inform intervention using data for specific schools, specific classrooms and, specific students. While formative assessment is utilized through both the DIBELS and the DRA, data must then be used to shape instruction. Schools have evidenced consistency in data collection; however, little evidence is shown for how educators are using the data subsequent to collection. While much of the emphasis on formative assessment is informational, that information is not useful unless the acquired knowledge catalyzes change toward more appropriate instructional strategies and techniques geared toward improving student learning and performance. Greater emphasis should be placed on the utility of data collected from formative assessment for meeting student needs.

Conclusion

Benchmark assessments can be useful for monitoring progress toward performance on end-of-level tests. Progress monitoring data can be used by schools to identify students who may need added intervention during the school year to prepare for acceptable performance on high-stakes tests. Districts and schools that use benchmark assessments for this purpose should determine which measures correlate most highly with the test the students will take. This study indicates that both the DRA and the DIBELS fill this role for the Utah ELA CRT.

REFERENCES

The Access Center: Improving Outcomes for All Students K-8. (2005). *Early reading assessment: A guiding tool for instruction.* Washington, DC: The Access Center.

Adams, M. J. (l990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

Afflerbach, P. (2002). The road to folly and redemption: Perspectives on the legitimacy of high-stakes testing. *Reading Research Quarterly, 37*(3), 348-360.

Afflerbach, P. (2005). National reading conference policy brief: High-stakes testing and reading assessment. *Journal of Literacy Research, 37*(2), 151-162.

Akimoff, K. G. (1996). *Parental involvement: An essential ingredient for a successful school*. Unpublished master's thesis, Dominican College, San Rafael, CA. (ERIC Document Reproduction Service No. ED400930)

Alliance for Excellent Education. (2005). *Policy brief: Why the crisis in adolescent literacy.* Retrieved April 18, 2008, from http://www.adlit.org/article/19968

Allington, R. L. (l984). Content coverage and contextual reading in reading groups. *Journal of Reading Behavior, 16*, 85-96.

Ananda, S.,& Rabinowitz, S. (2000). *The high stakes of high-stakes testing* [policy brief]. San Francisco, CA: WestEd.

Armbruster, B. B., Lehr, F., & Osborn, J. (2003). *Put reading first: The research building blocks for teaching children to read.* Washington, DC: Partnership for Reading.

Baker, S., & Smith, S. (2001). Linking school assessments to research-based practices in beginning reading: Improving programs and outcomes for students with and without disabilities. *Teacher Education and Special Education, 24*(4), 315-332.

Barger, J. (2003). *Comparing the DIBELS oral reading fluency indicator and the North Carolina end-of-grade reading assessment*. Asheville, NC: North Carolina Teacher Academy.

Beaver, J. (2001). *Developmental reading assessment.* Upper Arlington, OH: Celebration Press.

Beaver, J. M.,& Carter, M. A. (2003). *Teacher guide: Developmental reading assessment, grades 4-8.* Parsippany, NJ: Pearson Education, Inc.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment*, *1*(2), 1-12.

Black, P., & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2): 139-149.

Bluebello, L. (2000). *High-stakes testing*. Retrieved March 3, 2008, from http://www.muse.widener.edu/~egrozyck/EDControversy/Bluebello.html

Brown, A. L., Palincsar, A. S., & Purcell, L. (1986). Poor readers: Teach, don't label. In U. Neisser (Ed.), *The school achievement of minority children: New perspectives* (pp. 105-143). Hillsdale, NJ: Lawrence Erlbaum Associates.

Buck, J., & Torgesen, J. (2003). The relationship between performance on a measure of oral reading fluency and performance on the Florida Comprehensive Assessment Test. Tallahassee, FL: Florida Center for Reading Research.

Casbarro, J. (2005). The politics of high-stakes testing. *Principal, 84*, 16-20.

The Center for Public Education. (2006). *A guide to the No Child Left Behind Act*. Retrieved January 30, 2008, from: *http://www.centerforpubliceducation.org/site/c.kjJXJ5MPIwE/b.1505669/k.D349/A_guide_to_the_No_Child_Left_Behind_Act.htm*

Clay, M. M. (1966). *Emergent reading behavior*. Unpublished doctoral dissertation, University of Auckland, New Zealand.

Cohen, P.A., Kulik, J.A., & Kulik, C.L.C.(1982). Educational outcomes of tutoring: A meta-analysis of findings. *American Educational Research Journal, 19*, 237-248.

Committee on Labor and Public Welfare. (1965). Elementary and Secondary Education Act of 1965. *Hearings before Subcommittee on Education of the Committee on Labor and Public Welfare, United States Senate*.: Washington, DC: U.S. Government Printing Office.

Coordinated Campaign for Learning Disabilities (1997). *Early warning signs of learning disabilities*. Retrieved December 28, 2007, from http://www.readingrockets.org/article/226

Cowie, B., & Bell, B. (1999). A model of formative assessment in science education, *Assessment in Education, 6*, 101-116.

Crooks, T. (2001, September). The validity of formative assessments. Paper presented to the British Educational Research Association Annual Conference, University of Leeds, September 13-15, 2001.

Cunningham, A.E., & Stanovich, K.E. (1997). Early reading acquisition and its relation to reading experience and ability 10 years later. *Developmental Psychology, 33*, 934-945.

Davis, A. (1998). *The limits of educational assessment*. Oxford, England: Blackwell.

DeBruin-Parecki, A. (2004). Evaluating early literacy skills and providing instruction in a meaningful context. *High/Scope Resource: A Magazine for Educators, 23*(3), 510.

Deno, S. L. (1985). Curriculum-based measurement: The emerging alternative. *Exceptional Children, 49,* 36-45.

Deno, S., Mirkin, P., & Chiang, H. (1982). Identifying valid measures of reading. *Exceptional Children*, *49*(1), 36-45.

Dickson, S. V., & Bursuck, W. D. (1999). Implementing a model for preventing reading failure: A report from the field. *Learning Disabilities Research & Practice, 14,* 191-202.

Ding, C. & Davison, M. (2004). A longitudinal study of math achievement gains for initially low achieving students. *Contemporary Educational Psychology, 30*(1), 81-95

Duffy, C. (2001). *The education standards movement spells trouble for private and home schools.* Retrieved January 25, 2008, from http://www.home-school.com/exclusive/standards.html

Durrell, D. D. (1955). *Durrell analysis of reading difficulty*. New York: Harcourt, Brace, and World.

The Education Trust. (2004). *The ABCs of AYP*. Retrieved February 20, 2008, from http://www2.edtrust.org/NR/rdonlyres/37B8652D-84F4-4FA1-AA8D-319EAD5A6D89/0/ABCAYP.PDF

Ehri, L. C. (1998). Grapheme-phoneme knowledge is essential for learning to read words in English. In J. Metsala & L. Ehri (Eds.), *Word recognition in beginning reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Formative Assessment. (2005, September). *Technology and Learning*, *26*(2), 49.

Francis, D. J., Shaywitz, S. E., Stuebing, K. K., Shaywitz, B. A., & Fletcher, J. M. (1996). Developmental lag versus deficit models of reading disability: A longitudinal, individual growth curves analysis. *Journal of Educational Psychology, 88*, 317.

Fuchs, L.S., & Fuchs, D. (2004). Determining adequate yearly progress from kindergarten through grade six with curriculum-based measurement. *Assessment for Effective Instruction, 29*(4), 25-38.

Fuchs, L. S., Fuchs, D., & Compton, D. L. (2004) Monitoring early reading development in first grade: Word identification fluency versus nonsense word fluency. *Exceptional Children, 71*(1), 7-21.

Fuchs, L. S., Fuchs, D., & Deno, S. L. (1982). Reliability and validity of curriculum-based informal reading inventories. *Reading Research Quarterly, 18*, 6–26.

Fuchs, L. S., Fuchs, D., & Maxwell, L. (1988). The validity of informal measures of reading comprehension. *Remedial and Special Education, 9*(2), 20-28.

Gagne, R. M. (1965).*The conditions of learning.* New York: Holt, Rinehart and Winston.

Gagne, R. M., & Medsker, K. L. (1996). *"The conditions of learning: Training applications."* Fort Worth, TX: Harcourt Brace College Publishers.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist, 18*, 510-522.

Good, R. H.,& Jefferson, G. (1998). Contemporary perspectives on Curriculum-Based Measurement validity. In M. R. Shinn (Ed.), *Advanced applications of Curriculum-Based Measurement* (pp. 61-88). New York: Guilford.

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.

Good, R. H., Simmons, D. C., & Kame'enui, E. J. (2001). The importance and decision-making utility of a continuum of fluency-based indicators of foundational reading skills for third-grade high-stakes outcomes. *Scientific Studies of Reading, 5*(3), 257-288.

Gough, P. B. (1996). How children learn to read and why they fail. *Annals of Dyslexia, 46,* 320.

Griffeth, P., & Olson, M. (1992). Phonemic awareness helps beginning readers break the code. *The Reading Teacher, 45*(7), 516-523.

Hammill, D. D.,& McNutt, G. (1980). Language abilities and reading: A review of the literature on their relationship. *Elementary School Journal, 80*(5), 269-277.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (1998). Teachers, schools, and academic achievement. Retrieved on February 9, 2008 from http://www.nber.org/papers/w6691.pdf

Hare, M. G. (2002). 'Reading Czar' has talk with educators. *The Baltimore Sun*. Retrieved February 2, 2008, from http://www.sunspot.net

Hintze, J. & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review*, *34*(3), 372-386.

Hosp, M. K., & Fuchs, L. S. (2005). Using CBM as an indicator of decoding, word reading, and comprehension: Do the relations change with grade? *School Psychology Review, 34*(1), 9-26.

Huey, E. B. (1908). *The psychology and pedagogy of reading*. New York: The Macmillan Company.

International Reading Assocation. (1999). High-stakes assessment in reading: A position statement of the International Reading Association. *Journal of Adolescent an Adult Literacy, 43*(3).

Jeffrey, J. R. (1978). *Education for the children of the poor: A study of the origins and implementation of the Elementary and Secondary Education Act of 1965.* Columbus, OH: Ohio State University Press.

Juel, C. (1996). What makes literacy tutoring effective? *Reading Research Quarterly, 31*, 268-289.

Kahl, S. (2000, Winter). Stakes, mistakes, & statewide testing. *The State Education Standard*, Winter, 18-21. Retrieved on March 21, 2008 from http://www.measuredprogress.org/resources/assessment/stakesmistakes.html

Kantor, H. (1991). Education, social reform, and the state: ESEA and federal education policy in the 1960s. *American Journal of Education, 100*(1), 47-83.

Kegler, M. C., Oman, R. F., Vesely, S. K., McLeroy, K. R., Aspy, C. B., Rodine, S., et al. (2005). Relationships among youth assets and neighborhood and community resources. *Health Education and Behavior, 32*(3), 380-397.

Langenfeld, K. L., Thurlow, M. L., & Scott, D. L. (1997). *High-stakes testing for students: Unanswered questions and implications for students with disabilities*

(Synthesis Report 26). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Learning First Alliance. (2000). *Every child reading: A professional development guide*. Retrieved January 22, 2008, from http://www.learningfirst.org/publications/reading/guide/content.html

Lewelling, V. (1991). *Academic achievement in a second language*. Washington, DC: ERIC Clearinghouse on Languages and Linguistics. Retrieved March 29, 2008, from http://www.ericdigests.org/pre-9219/second.htm

Lewis, A. (2000). *High-stakes testing: Trends and issues* [policy brief]. Aurora, CO: Mid-Continent Research for Education and Learning.

Marston, D. (1989). A curriculum-based measurement approach to assessing academic performance: What it is and why do it. In M. R. Shinn (Ed.), *Curriculum-based measurement: Assessing special children* (pp. 18–78). New York: Guilford.

Mastropieri, M. A., & Scruggs, T. E. (1997). Best practices in promoting reading comprehension in students with learning disabilities: 1976-1996. *Remedial and Special Education 18*, 197-213.

Mathes, P. G., Denton, C.A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics on the skills of struggling readers. *Reading Research Quarterly, 40*(2), 148–182.

McColl, A. (2005). Has “No Child” left behind the constitution? *Phi Delta Kappan, 86*, 604-10.

McGlinchey, M. T., & Hixon, M. D. (2004). Using curriculum-based measurement to predict performance on state assessments in reading. *School Psychology Review, 33*(2), 193-203.

McNeil, M., Coppola, E., Radigan, J., & Vasquez Heilig, J. (2008). Avoidable losses: High-stakes accountability and the dropout crisis. *Education Policy Analysis Archives, 16*(3), 1-48. Retrieved on March 21, 2008 from http://epaa.asu.edu/epaa/v16n3/

Meyer, B. J. F.,& Rice, E. (1984). The structure of text. In P. D. Pearson, R. Barr, M. L. Kamil, & P. Mosenthal (Eds.), *The handbook of reading research* (pp. 319-52). New York: Longmans.

Nagy, W. E., Herman, P. A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly, 20*, 233-253.

National Center for Education Statistics (2003-2004). "*State nonfiscal survey of public elementary/secondary education.*" The NCES common core of data (CCD). Retrieved on March 30, 2008, from http://nces.ed.gov/programs/digest/d05/tables/dt05_052.asp

National Council of Teachers of English (2005). *Status on the implementation of the No Child Left Behind Act (NCLB).* Retrieved January 13, 2008, from http://www.ncte.org/about/issues/national/views/116435.htm

National Center for Educational Statistics. (2005*). 2009 NAEP reading framework.* Washington, DC: Author.

National Center for Educational Statistics. (2006). *States that use criterion-referenced tests (CRTs) aligned to state standards, by subject area and level*. Retrieved

February 2, 2008, from http://nces.ed.gov/programs/digest/d07/tables/dt07_159.asp

National Commission of Excellence in Education, U.S. Department of Education. (1983). *A nation at risk: The imperative for educational reform*. Retrieved February 2, 2008, from http://www.ed.gov.pubs.NatAtRisk

National Institute of Child Health and Human Development. (2000). 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). Washington, DC: U.S. Government Printing Office.

Nelson, D. E., & Lawrence, B. J. (1994). *Utah's major student assessment programs*. Salt Lake City, UT: Utah State Office of Education.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), pp.199-218.

Nichols, S., Glass, G., & Berliner, D. (2005). *High-stakes testing and student achievement: Problems for the No Child Left Behind Act.* Tempe, AZ: Educational Policy Research Unit.

Nichols, S., Glass, G., & Berliner, D. (2007). *Collateral damage: How high-stakes testing corrupts American's schools*. Cambridge, MA: Harvard Education Press.

Oka, E., & Paris, S. (1986). Patterns of motivation and reading skills in underachieving children. In S. Ceci (Ed.*), Handbook of cognitive, social, and neuropsychological aspects of learning disabilities* (Vol. 2, pp. 115-146). Hillsdale, NJ: Erlbaum.

Olson, L. (2005,November 30).Benchmark assessments offer regular checkups on student achievement. *Education Week, 25*(13), 13-14. Retrieved January 30, 2008, from http://www.edweek.com

O'Shea, M. (2006). Beyond compliance: Steps to achieving standards. *Principal Leadership,6*(8), 28-31.

Paige, R. (2006). No Child Left Behind: The ongoing movement for public education reform. *Harvard Educational Review,* 76(4), 461-473.

Paris, S. G. (2007). Assessment of reading comprehension. In *Encyclopedia of language and literacy development* (pp. 1-8). London, ON: Canadian Language and Literacy Research Network. Retrieved January 23, 2008, from http://www.literacyencyclopedia.ca/pdfs/topic.php?topId=226

Rashotte, C. A., Torgesen, J. K., & Wagner, R. K. (1997, September). *Growth in reading accuracy and fluency as a result of intensive intervention*. Paper presented at the annual meetings of the Florida Branch of the International Dyslexia Association, Miami, FL.

Rasinski, T. V. (n.d.). *Assessing reading fluency*. *Pacific resources for education and learning.* Retrieved January 21, 2008, from http://www.prel.org/products/re_/assessing-fluency.htm

Ravitch, D. (1995). 50 ways to teach them grammar. In D. Ravitch, *National standards in American education*. Washington, D.C.: Brookings Institution Press. Retrieved on January 15, 2008, from http://www.brookings.edu/opinions/1996/0411education_ravitch.aspx

Rothkope, A. J. (2007). *Elementary and secondary education act reauthorization: Improving NCLB to close the achievement.* Presentation to the Senate Committee on Health, Education, Labor and Pensions, and the House Committee on Education and Labor, March 13, 2007. Retrieved March 21, 2008, from http://www.publiceducation.org/nclb_main/Reauth_Positions_Organizations.asp

Samuels, S. J.,& Edwall, G. E. (1975). Measuring reading achievement: A case for criterion-referenced testing and accountability. *NCME Measurement in Education, 6*(2), 1-8.

Sarroub, L., & Pearson, P. D. (1998). Two steps forward, three steps back: The stormy history of reading comprehension assessment. *Clearing House, 72*(2), 97-105.

Scarborough, H. S. (1998). Early identification of children at risk for reading disabilities: Phonological awareness and some other promising predictors. In B. K. Shapiro, P. J. Accardo, & A. J. Capute (Eds.), *Specific reading disability: A view of the spectrum* (pp. 75-120). Timonium, MD: York Press.

Schemo, D. J., & Fessenden, F. (2003, December 3). Gains in Houston schools: How real are they? *The New York Times*, p. A1.

Schumm, J. S. (2006). *Reading assessment and instruction for all learners*. New York: Guilford Press.

Schugurensky, D. (2002). History of education: Selected moments of the 20th century. Retrieved February 18, 2008, from Website of Daniel Schugurensky: http://fcis.oise.utoronto.ca/~daniel_schugurensky/assignment1/1965elemsec.html

Siegel, L. S. (1989). IQ is irrelevant to the definition of learning disabilities. *Journal of Learning Disabilities, 22*, 469-479.

Shaw, R., & Shaw D. (2002). DIBELS oral reading fluency-based indicators of third grade reading skills for Colorado state assessment program (CSAP) (Technical Report). Eugene, OR: University of Oregon. Retrieved February 1, 2006, from http://dibels.uoregon.edu/techreports/index.php

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

Shinn, M. R., & Bamonto, S. (1998). Advanced applications of curriculum-based measurement: "Big ideas" and avoiding confusion. In M. R. Shinn (Ed.), *Advanced Applications of Curriculum-Based Measurement* (pp.1-31). New York, NY: Guilford Press.

Shinn, M. R., Shinn, M. M., Hamilton, C., & Clarke, B. (2002).Using curriculum-based measurement in general education classrooms to promote reading success. In M. R. Shinn (Ed.), *Interventions for academic and behavior problems II: Preventive and remedial approaches (*pp. 113-142). Bethesda, MD: NASP.

Sibly, D., Biwer, D., & Hesch, A. (2001). *Establishing curriculum-based measurement oral reading fluency performance standards to predict success on local and state tests of reading achievement.* Paper presented at the Annual Meeting of the National Association of School Psychologists, Washington, DC.

Singer, H. (1983). A century of landmarks in reading and learning from text at the high school level: Research, theories, and instructional strategies. *Journal of Reading, 26,* 332-342.

Smith, N. B. (2002). *American reading instruction.* (Special Ed.). Newark, DE: International Reading Association.

Snow, C. E., Burns, M. S., & Griffin, P. (Eds.). (1998). *Preventing reading difficulties in young children*. Washington, DC: National Academy Press.

Spring, J. (1993). *Conflicts of interests: The politics of American education*. New York: Longman.

Stage, S. A., & Jacobsen, M. D. (2001). Predicting student success on a state-mandated performance-based assessment using oral reading fluency. *School Psychology Review 30*(3), 407-319.

Stecker, P. M. (n.d.). Monitoring student progress in individualized educational programs using curriculum-based measurement. National Center on Student Progress Monitoring, Washington, D.C. Retrieved February 5, 2008, from http://www.osepideasthatwork.org/parentkit/14%20-%20Monitoring%20Student%20Progress%20in%20IEPs%20using%20CBM.pdf

Steinberg, J. (2000, December 30). Roderick Raynor Paige. *New York Times*, p. A10.

Steinberg, J.,& Henrique, D. B. (2001, May 21). When a test fails the schools, careers and reputations suffer. *The New York Times*. Retrieved March 31, 2008, from www.nytimes.com/learning/testing

Sulzby, E., & Teale, W. H. (1991). Emergent literacy. In R. Barr, M. L. Kamil, P. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research, Vol. II* (pp. 727-757). New York: Longman.

Thorndike, E. L. (1971). Reading as reasoning: A study of mistakes in paragraph reading. *Research Quarterly, 6*(4), 425-434. (Reprinted from *The Journal of Educational Psychology*, 1917, June)

Torgesen, J. K. (1998). Catch them if before they fall: Identification and assessment to prevent reading failure in young children. *American Educator, 22*(1), 32-39

Torgeson, J. K.,& Burgess, S. R. (1998). Consistency of reading-related phonological processes throughout early childhood: Evidence from longitudinal-correlational and instructional studies. In J. Metsala & L. Ehri (Eds.), *Word recognition in beginning reading*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Tyler, R. (1972, February). Why evaluate education? *Compact, 6*(1) p.3-4.

United States Department of Education. (n.d.). *No Child Left Behind: Foreword by President George W. Bush.* Retrieved February 12, 2008, from http://www.ed.gov/nclb/overview/intro/presidentplan/proposal.pdf

United States Department of Education. (2002). *The No Child Left Behind Act of 2001: Executive summary*. Retrieved December 11, 2008 from http://www.ed.gov/nclb/overview/intro/execsumm.html

United States Department of Education. (2003). *Standards and assessment guidance*. Retrieved February 19, 2008, from www.ed.gov/policy/elsec/guid/saaguidance03.doc

United States Department of Education. (2004). *Four pillars of NCLB.* (2004). Retrieved January 18, 2008, from http://www.ed.gov/nclb/overview/intro/4pillars.html

United States Department of Education. (2005). *The facts about state standards*. Retrieved January 18, 2008, from http://www.ed.gov/nclb/accountability/standards/standards.html

United States Department of Education, National Center for Education Statistics. (2004-2005). "*State nonfiscal survey of public elementary/secondary education*" (NCES

Common Core of Data [CCD]). Retrieved March 30, 2008, from http://nces.ed.gov/pubs 2007/minoritytrends/tables/table_7_2.asp?referrer=report

United States Department of Education, National Center for Education Statistics. (2004-2005). "*Public elementary/secondary school universe survey,"* version 1A. (NCES Common Core of Data [CCD]). Retrieved March 30, 2008, from http://nces.ed.gov/pubs2007/pesschools06/tables/table_7.asp?referrer=report

Utah Educational Advisory Committee. (2007). *Minutes from meeting: April 5, 2007*. Retrieved February 20, 2008, from http://www.solveednow.org/Documents/04-05-2007_EAC%20Minutes.pdf

Utah State Office of Education. (2004). *Technical report*. Salt Lake City, UT.

Vander Meer, C. D., Lentz, F. E., & Stoller, S. (2005). *The relationship between oral reading fluency and Ohio proficiency testing in reading* (Technical Report). Eugene, OR: University of Oregon. Retrieved November 30, 2005, from http://dibels.uoregon.edu/techreports/ohio.pdf

Vaughn, S.,& Schumm, J. S. (1996). Classroom ecologies: Classroom interactions and implications for inclusion of students with learning disabilities. In D. L. Speece & B. K. Keogh (Eds.), *Research on classroom ecologies* (pp.107-124). Mahwah, NJ: Lawrence Erlbaum Associates.

Vygotsky, L. (1986). *Thought and language*. Cambridge, MA: MIT Press.

Wang, M. & Kovach, J. (1995). Bridging the achievement gap in urban schools: Reducing the educational segregation and advancing resilience – Promoting strategies. Paper presented at a Conference of the Urban Education National Network of the Regional Education Laboratories (Washington, DC, May 5, 1995).

Weber, W. A. (2000). *Developmental reading assessment and evaluacion del desarrollo del la lectura: A validation study* (Research Report). Houston, TX: University of Texas. Retrieved March 11, 2006, from www.pearsonlearning.com

WestEd Assessment and Standards Development Services. (2001). Standards to support standards-based assessment. Retrieved on February 15, 2008, from http://www.wested.org/asds/standards.shtml#utah

Wiliam, D. (2005). *The formative purpose: assessment must first promote learning*. CCSSO Presentation, 35th Annual National Conference on Large-Scale Assessment, San Antonio, TX.

Williams, E. J. (1999). *Developmental reading assessment reliability study* (Research Report). Lebanon, IN: Pearson Learning Group. Retrieved March 11, 2006, from www.pearsonlearning.com

Wilson, J. (2005). *The relationship of dynamic indicators of basic early literacy skills (DIBELS) oral reading fluency to performance on Arizona instrument to measure standards (AIMS).* Tempe, AZ: Tempe School District No. 3. Retrieved February 1, 2006, from http://dibels.uoregon.edu/techreports/index.php

Wren, S. (2004). *Descriptions of early reading assessments*. Southwest Educational Developmental Laboratory. Retrieved February 2, 2008, from http://www.balancedreading.com/assessment/assessment/pdf.