2017-03-01

Comparing Two Individually Administered Reading Assessments for Predicting Outcomes on SAGE Reading

Meighan Noelle Stevens
Brigham Young University

Comparing Two Individually Administered Reading Assessments for Predicting Outcomes on SAGE Reading

Meighan Noelle Stevens

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Educational Specialist

Gordon Stanley Gibb, Chair
Richard R. Sudweeks
Ellie L.Young

Department of Counseling Psychology and Special Education

Brigham Young University

ABSTRACT

Comparing Two Individually Administered Reading Assessments
for Predicting Outcomes on SAGE Reading

Meighan Noelle Stevens
Department of Counseling Psychology and Special Education, BYU
Educational Specialist

Accountability for student learning outcomes is of importance to parents and school and district administrators, especially since the passage of The No Child Left Behind Act in 2001. The requirement for high-stakes testing to measure progress has fostered interest in ways to monitor student preparedness during the school year. This study used 2014 and 2015 test data from of 154 students from one elementary school to measure the correlation between individually administered Kaufman Test of Educational Achievement Brief Reading and DIBELS Next reading assessments and outcomes on the high-stakes Utah SAGE test. This correlational study used Pearson correlation coefficients to determine redundancy across the tests, and used multiple regression to assess how well scores on the KTEA and DIBELS Next tests predict students' subsequent scores on the SAGE test. Results indicate that DIBELS Next was a strong predictor of SAGE outcomes while KTEA Brief results were moderate predictors.

Keywords: high stakes testing, Student Assessment of Growth and Excellence, Dynamic Indicators of Basic Early Literacy Skills Next, Kaufman Test of Educational Achievement, curriculum-based assessment, reading

ACKNOWLEDGMENTS

I would like to acknowledge my chair Mr. Gordon Gibb for all of his help and dedication to the project. His encouragement and guidance were invaluable throughout the process. I would also like to express appreciation to my family for always supporting my ambitions and goals in life. They help me strive for success in all avenues of life. Finally, I would like to acknowledge the faculty and staff at Brigham Young University for providing a tremendous academic environment to learn and grow in.

TABLE OF CONTENTS

LIST OF TABLES

DESCRIPTION OF THESIS STRUCTURE AND CONTENT

This thesis, *Comparing Two Individually Administered Reading Assessments for Predicting Outcomes on SAGE Reading*, is written in a hybrid format. This format combines the elements of traditional thesis requirements and journal publication formats. The initial pages of this thesis represent requirements for submission to the university. The thesis report is presented as a manuscript consistent with length and style requirements for education journals. It includes the following sections: introduction, method, results, discussion, and references. Appendix A, included after the main body of the thesis, includes an extended literature review as presented at the prospectus defense. Appendix A has a separate reference list that includes only those references that are cited in Appendix A.

Introduction

The multiplicity of assessments in modern American schools reflects the demand for accountability to the nation, to states, and to district and school patrons. Assessment in schools can be formative or summative. Formative assessment is "concerned with how judgments about the quality of student responses (performances, pieces, or works) can be used to shape and improve the students' competence by short-circuiting the randomness and inefficiency of trial-and-error learning" (Sadler, 1989, p. 120). Educators use formative assessment to determine where students are functioning after a lesson or unit in the classroom. Formative data can be used by teachers to alter or continue instructional strategies depending on student progress. Therefore, the purpose of formative assessment is to drive instructional decisions. Summative assessment "contrasts with formative assessment in that it is concerned with summing up or summarizing the achievement status of a student, and is geared towards reporting the end of a course of study" (Sadler, 1989, p. 120). Educators use summative assessment to show achievement and growth at the end of instructional units, courses, or from one school year to the next. Formative assessments are usually informal and often created by teachers, whereas summative assessment can be informal or standardized.

As educational practices continue to evolve so does the multitude of assessments used in schools and districts. Testing that impacts school, district, and state standing within the requirements of the No Child Left Behind Act (NCLB) is often called high-stakes testing (Lewis & Hardy, 2015). Haladyna (2006) defines high-stakes testing as "the use of test scores that have a significant consequence for students, teachers, schools, and school districts," including "graduation, promotions, school accountability, federal funding, No Child Left Behind, and merit pay" (p. 23). Performance on high-stakes testing can grade districts, schools, and even teachers

in ways that inform students, teachers, and administrators. Good performance on standardized tests can indicate successful teaching and leadership, while poor performance can label schools as ineffective.

Curriculum-based testing differs from standardized testing in that it is used by educators to track and improve student performance. These data show the effectiveness of the curriculum and instruction on the student achievement (Deno, 1992). As teachers track the effectiveness of instruction in the classroom, instructional methods can be altered to find the best positive projective. This also provides teachers with data that can be used as evidence of where the students have been and where they are going (Marchand & Furrer, 2014).

There are two general types of curriculum-based measures: curriculum-based assessment and curriculum-based measurement. Curriculum-based assessment is commonly defined as one-time testing performed at the end of a unit, semester or year (Deno, 1992). In contrast, curriculum-based measurement is used repeatedly to assess the growth in student learning and can be compared to formative assessment for demonstrating progress.

Progress monitoring is an application of curriculum-based measurement that allows teachers to see and create change in a student's performance over a period of time (Deno, 1992). Progress monitoring is used in schools to track the progress of students in between benchmark assessments and through the year, to project whether or not individual students will achieve benchmarks or be on grade level by the end of the year. This approach is not just a record of the students' achievement. It also measures the effects of teachers' practices in the classroom (Deno, 1992). Progress monitoring indicates the effectiveness of classroom instruction by tracking students' performance to demonstrate what areas need improvement or need to be

taught in more depth for students to reach proficiency. Instruction is then targeted at the specific point in the academic learning that is weak (Catts, Nielsen, Bridges, Liu, & Bontempo, 2015).

Progress monitoring results are usually graphed to depict the slope of student achievement over time (Deno, 1992). When the target is a specific benchmark or aim point, then the slope indicates whether or not students are on track to meet achievement objectives. Progress monitoring can also be quite indicative in a short period of time. Thornblad and Christ (2014) found that six weeks of daily monitoring produced valid and reliable indicators of student growth for reading. Jenkins, Graff, and Miglioretе (2009) noted that less frequent monitoring will also produce accurate predictions provided that a sufficient number of scores are used.

Researchers have used progress monitoring to predict scores on standardized summative tests. Richardson, Hawken, and Kircher (2012) used maze comprehension testing to compare students who speak predominantly Spanish at home with native English speakers, and to compare Hispanic student outcomes to those of Caucasian students. The authors found that maze measures under predicted scores on end-of-year high stakes tests for Hispanic students and non-native English speakers, but accurately predicted outcomes for Caucasian students and native English speakers. Curriculum-based measures of comprehension used with Tennessee 3rd grade readers demonstrated strong predictive power for end-of-level tests (Miller, Bell, & McCallum, 2015), as did oral reading and maze measures for Nebraska students in grades two through five (Merino & Beckman, 2010).

Dynamic Indicators of Basic Early Literacy Skills

Dynamic Indicators of Basic Early Literacy Skills (DIBELS) is a progress-monitoring assessment used to measure reading skills from kindergarten through 6th grade. The National Reading Panel (Armbruster, Lehr, & Osborn, 2001) identified five critical components for

successful reading, including phonemic awareness, phonics, fluency, vocabulary, and text comprehension. DIBELS assesses each of these components and can be administered in a few minutes, providing a quick view of student ability that informs early intervention (Good & Kaminski, 2011).

DIBELS was designed in the late 1980's based on curriculum-based measurement, with initial research at the University of Oregon. Curriculum-based measurement and DIBELS both aim for an economical and efficient way to demonstrate student growth towards achievement. DIBELS Next, the current version, measures six areas in reading that can be monitored for progress over time, including first sound fluency, letter naming fluency, phoneme segmentation fluency, nonsense word fluency, oral reading fluency, and text comprehension (Good & Kaminski, 2011).

Kaufman Test of Educational Achievement

The Kaufman Test of Educational Achievement (KTEA) is a standardized test developed by clinical psychologists Alan S. Kaufman and Nadeen L. Kaufman. Currently in its third edition, the KTEA is a norm-referenced test designed to compare student achievement to that of other students in the same age and grade (Cole, 2012). Each test is designed with an increase of testing time based on grade level: Pre-Kindergarten and Kindergarten are given a 30-minute testing session, grades 1-2 a 50-minute testing session, and grade 3 and above an 80-minute session.

KTEA subtests include comprehensive achievement, reading, math, decoding, written language, sound symbol, oral language, and oral language fluency. Subtest scores are combined to form a composite score. Separate types of questions within each section measure students' knowledge. For example, the math composite score is a combination of math concepts and

application, math computation, and math fluency, all of which encompass multiple topic areas (Kaufman & Kaufman, 2004).

Student Assessment for Growth and Excellence (SAGE)

The Student Assessment for Growth and Excellence (SAGE) is a computer-adaptive summative test specific to the Utah core curriculum and used as the end-of-year high-stakes assessment in Utah schools. SAGE tests reading, language arts, math, science, and writing. In reading, students are tested on listening comprehension as well as the reading standards that align with the student's grade level. SAGE is an adaptive test, meaning that the testing software analyzes a student's answer to one question to determine the question that will follow (Utah State Board of Education, 2015a). Therefore, "the difficulty of the test will adjust to each student's skills, providing a better measure of what each student know and can do" (Utah State Board of Education, 2015a, p. 2).

SAGE also offers optional fall and winter versions that schools can use to inform instruction and monitor student progress. A formative teacher instruction manual with guidance on classroom instruction is available, but not required for use. The computerized SAGE requires students to "create graphs, interact with science simulation, and write and respond in multiple different ways. These question types will assess higher order thinking skills" (Utah State Board of Education, 2015b, p. 1). With the Utah Core relying more upon higher order analysis and problem solving skills, these questions provide a deeper questioning assessment. By having the students think in this manner, Utah is planning to guide students into the careers of the 21st century (Utah State Board of Education, 2015b).

The Utah State Office of Education reports that no norming procedures were conducted prior to implementing the assessment. Instead, the state collects test data as it is implemented in the schools. Therefore, the reliability and validity of SAGE was yet to be reported in 2015.

Purpose of Study and Research Questions

The purpose of this study was to determine to what degree two standardized reading assessments predict reading outcomes on the SAGE test, and to determine to what extent the assessments provide redundant information. Searches of research literature using the terms KTEA, DIBELS Next, SAGE, standardized testing, reliability, validity, and combinations thereof on the ERIC, Google Scholar, and PsycInfo databases produced no studies of the three tests, most likely due to the recent implementation of SAGE. The study addressed two questions:

1. To what extent do reading scores obtained from the KTEA test and DIBELS Next test predict scores obtained from each of the two SAGE reading subtests?
2. To what extent do DIBELS Next, KTEA, and SAGE reading scores provide redundant information?

Method

Research Participants

The study used existing data for students in grades 3 through 6 enrolled in a public elementary school in an urban Utah school district during the 2014-2015 school year. The school provided the data with anonymous codes for student names so the investigator could not identify the participants other than by grade level. This group was selected because these students participated in the DIBELS Next, KTEA and SAGE testing. The overall sample included 154 male and female students ranging in age from 8 to 13 years old.

Setting

The Utah State Office of Education website (2015) reported that during the 2014–2015 school year there were 592 students enrolled in the school, including 345 males and 247 females. Three hundred and three students came from low-income homes and 20% were classified as English as a Second Language. Student ethnicity is depicted in Table 1. Additionally, grades 3-6 had a 25:1 students-to-teacher ratio; median class size is depicted in Table 2 (Utah State Office of Education, 2015c).

Table 1

Ethnicity Populations in the Study School

Ethnicity	Number	Percent
American Indian	5	1%
Asian	7	1%
Black	7	1%
Hispanic	247	42%
Pacific Islander	11	2%
White	306	53%
Total	593	100%

Table 2

Average Number of Students in Each Classroom by Grade Level

Grade level	Students per classroom
3	25
4	27
5	23
6	27

Statistics that highlight the number of students qualifying for free and reduced lunches are also a portion of the demographics of the school. During 2014-2015, 64.92% of students were on a free or reduced lunch plan (Utah State Office of Education, 2015d).

Predictor Variables

Predictor variables enable prediction of a criterion variable, or the variable of interest in the study (Marchand & Furrer, 2014). Predictor variables for the study were the KTEA Reading Comprehension scores and DIBELS Next Oral Reading Fluency scores. These scores were used to predict the criterion variables.

Criterion Variables

The criterion variables are the variables of interest in this study. The criterion variables in this study are the SAGE reading literature and reading for information scaled scores for 2015.

Instruments

The instruments used in the study were the KTEA-II Brief, the DIBELS Next, and the SAGE standardized reading assessment. The KTEA-II Brief measures reading, decoding, oral language, and oral language fluency. The KTEA-II Reading Composite score was a predictor variable in the calculations. DIBELS Next measures reading skills including phonemic awareness, phonics, fluency, vocabulary, and text comprehension. The DIBELS Oral Reading Fluency was the other predictor variable used in calculations. SAGE measured Informational Text as well as Reading Literature.

Design

This correlational study involved "collecting two sets of data and determining the extent to which they covary (or vary together)" (Martella et al., 2013, p. 208). Data sets include reading scores on the KTEA andreading scores on DIBELS Next predictor variables, and reading scores

on SAGE criterion variable. The study investigated predictive correlations for reading scores between KTEA and SAGE and DIBELS Next and SAGE.

Data Collection

Data included the KTEA 2014, DIBELS Next 2015, and SAGE 2015 assessment scores obtained from the target school. Assessment scores were analyzed for the 154 students who completed all three assessments or 75% of students in in grades 3–6.

Data Analysis

Two procedures were used to analyze the data. First, the Pearson correlation coefficient was used to calculate zero-order correlations between SAGE and KTEA, SAGE and DIBELS Next, and KTEA and DIBELS Next to determine the level of redundancy provided by each test. Second, multiple regression was used to regress SAGE scores on KTEA scores and on DIBELS Next scores to measure the predictive power of each benchmark assessment on SAGE outcomes, as in $SAGE = b_0 + b_1(KTEA) + b_2(DIBELS\ Next) + e$.

Redundancy of the information was calculated using the coefficient of determination between each set of scores. The correlation between KTEA and DIBELS Next scores provided the redundancy through r^2.

Results

Separate multiple linear regression analyses were calculated to determine relationships among the two SAGE subtests: 2015 reading literature and 2015 informational text. KTEA and DIBELS Next were then correlated using the Pearson Product Moment correlation coefficient.

The first research question asked the extent to which reading scores on DIBELS Next and KTEA predict reading scores on SAGE. The SAGE scores for Reading Literature and Informational Text were analyzed with the KTEA and DIBELS Next data in a predictive

manner. KTEA was significantly correlated with both SAGE tests and DIBELS Next was significantly correlated with both SAGE tests.

The second research question asked the extent to which reading literature scores provide redundant information. Analysis shows that KTEA, DIBELS Next, and SAGE reading literature provide redundant information. The zero-order correlations between KTEA, SAGE and DIBELS Next are 0.478 and 0.409 which are significant. Individually, each test is a predictor. However, the partial correlations are 0.272 and -0.007, which are not significant. In the presence of each other, they do not have significant unique predictive ability, indicating redundancy.

The zero-order correlations for KTEA and DIBELS Next with SAGE Informational Text were 0.241 and 0.141, which were not significant. Individually, the tests are not significant predictors of SAGE 2015 Informational Text. The partial correlation coefficients were 0.238 and -0.136, which were not significant. The predictors do not provide significant unique prediction in the presence of each other.

The zero-order correlations between KTEA and DIBELS Next with SAGE Reading Literature were 0.713 and 0.703, which are significant (Table 3). Each test is a significant predictor of Reading Literature by itself. However, the partial correlation coefficients were 0.306 and 0.261, which were not significant.

The zero order correlations between KTEA and DIBELS Next) with SAGE Informational Text were 0.425 and 0.571 which were significant. The partial correlations were -.142 and 0.440. Only the DIBELS Next 15 provided unique information in the presence of KTEA 14. KTEA 14 in the presence of DIBELS Next 14 does not provide new information.

Table 3

Results of the Regression of SAGE 2015 Reading Literature Scores on 2014 KTEA Scores and DIBELS Next 2015 Scores

	Unstandardized Coefficients		Standardized Coefficients		Correlations	
Model	B	Std. Error	Beta	T	Zero-Order	Partial
KTEA	1.583	0.947	0.415	1.671	0.713	0.306
DIBELS Next	0.215	0.153	0.350	1.407	0.703	0.261

Table 4

Results of the Regression for SAGE 2015 Reading Informational Text Scores on 2014 KTEA and 2015 DIBELS Scores

	Unstandardized coefficients		Standardized coefficients			Correlations	
Model	B	Std. Error	Beta	T	Sig.	Zero-order	Partial
Constant	212.202	80.163		2.647	.013		
KTEA 15	1.745	1.322	.464	1.320	.197	.241	.238
DIBELS 16	-.154	.209	-.259	-.737	.467	.141	-.136

The zero order correlations between KTEA and DIBELS Next with SAGE Informational Text were 0.425 and 0.571 which were significant. The partial correlations were -.142 and 0.440. Only the DIBELS Next provided unique information in the presence of KTEA. KTEA in the presence of DIBELS Next does not provide new information.

Discussion

Zero-order correlations are significant for both KTEA and DIBELS Next in predicting SAGE scores, indicating that each test predicts SAGE scores to some degree. Partial correlations for KTEA, DIBELS Next, and SAGE are not significant, meaning that combining KTEA and

DIBELS Next produces enough redundancy that the combination does little to improve predictive power. The zero-order significance indicates that the test with the largest correlation would be most useful for predicting SAGE scores at the school.

Pearson correlation coefficients indicate the degree to which scores on KTEA and DIBELS Next predicted SAGE subtest scores. The KTEA predicted SAGE Reading Literature scores with a coefficient of .488 and the SAGE Informational Text scores with a coefficient of .464, both considered moderate (Cronk, 2010). In 2015, DIBELS Next predicted both the SAGE-15 Reading Literature and Informational Text scores with a coefficient of .760, a strong correlation. Therefore, the DIBELS Next was a better predictor of SAGE scores than the KTEA.

These results add to those reported by Espin et al. (2010), Marchand and Furrer (2014), Merino and Beckman (2010), Miller et al. (2015), and Nese et al. (2011); all of which reported that curriculum-based measurement benchmarks reliably predicted performance on high stakes tests. Data indicate that DIBELS Next predicts comprehension scores as measured by SAGE Reading Literature and Informational Text. The authors did not find peer-reviewed research regarding the predictive power of KTEA for end-of-level test scores, despite searching multiple data bases. This is likely due to the specific nature of the KTEA. The KTEA is an individually administered achievement battery commonly used for identifying specific skill strengths and deficits in children with significant academic deficits or suspected learning disabilities. It would be expensive in time and other resources to assess every child individually on a test that is not designed for school-wide benchmarking and takes significantly longer to administer than DIBELS Next.

The problems addressed by this study were twofold: teachers currently administer more reading assessments than may be necessary to monitor student progress during the school year,

and teachers lack information about the effectiveness of these assessments for predicting outcomes on the end-of-year SAGE assessment. Results of the study indicate that school-wide DIBELS Next testing at benchmark points during the year is useful in predicting performance on the SAGE test. Therefore, it is unnecessary for the school to administer more than one predictive test during the year.

Limitations

The study was limited in that participant scores represented 75% of the students in grades 3-6, leaving one-fourth of student scores unaccounted for in the analysis. There is no way to know the possible effects of the additional scores in calculating the impact on the two predictor variables. The study was also limited in that the SAGE scores were from the second year of statewide testing with no validity or reliability studies conducted to establish these key characteristics of the test. In addition, the results only apply to the years represented by the predictor and criterion data used and cannot be generalized to other assessments in subsequent years.

Implications for Practice

Results of the study indicate that DIBELS Next benchmark scores are strongly correlated with results on the SAGE test. Schools should feel confident in using the DIBELS Next results to identify and intervene with students who may be at risk for low performance on the end-of-level test.

Implications for Further Research

Richardson et al. 2012 compared students for whom home languages were either Spanish or English and found that curriculum-based measurement under predicted end of year test scores for non-native English speakers. The current study did not differentiate between native and non-

native English speakers, but these associations are worthy of future study. Further, it would be useful to compare results across multiple schools and to look at Social Economic Status or other variables.

This study compared KTEA data from 2014 and DIBELS Next and SAGE data from 2015. Further studies should compare tests given in the same year or years with the corresponding end-of-level tests to avoid possible differences in the SAGE itself or in the study participants across years.

Conclusion

Periodic benchmark assessments can help schools predict student outcomes on high-stakes tests. The utility of these predictions lies in using benchmark data to modify instruction, student groupings, or other elements of the school program to enhance student learning and preparedness for end-of-level assessments. This study shows that DIBELS Next strongly predicts outcomes on the Utah SAGE assessment, and can be used by schools and districts to track student preparation for high-stakes tests.

References

Armbruster, B. B., Lehr, F., & Osborn, J. (2001). *Put reading first: The research building blocks for teaching children to read*. Washington, DC: U.S. Department of Education.

Catts, H. W., Nielsen, D. C, Bridges, M. S., Liu, Y. S., & Bontempo, D. E. (2015). Early identification of reading disabilities within an RTI framework. *Journal of Learning Disabilities, 48*(3), 281–297. doi: 10.1177/0022219413498115

Cole, A. (2012). *Kaufman Test of Educational Achievement*. Retrieved from http://amandacoleassessment.weebly.com/ktea.html

Deno, S. (1992). The nature and development of curriculum-based measurement. *Preventing School Failure, 36*(2), 5–10. doi: 10.1080/1045988X.1992.9944262

Good, R. H., & Kaminski, R. A. (2011). *DIBELS Next assessment manual*. Eugene, OR: Dynamic Measurement Group.

Haladyna, T. M. (2006). Perils of standardized achievement testing. *Educational Horizons, 85*(1), 30–43.

Jenkins, J., Graff, J., & Migliorette, D. (2009). Estimating reading growth using intermittent CBM progress monitoring. *Exceptional Children, 75*(2), 151–163. doi: 10.1177/001440290907500202

Kaufman, A. S, & Kaufman, N. L. (2004). *K-TEA II: Kaufman Test of Educational Achievement: Comprehensive Form* (2nd ed.). Circle Pines, MN: AGS Pub.

Kaufman, A. S., & Kaufman, N. L. (2014). *K-TEA III: Kaufman Test of Educational Achievement: Form A* (3rd ed.). Bloomington, MN: PsychCorp.

Lewis, S., & Hardy, I. (2014). Funding, reputations, and targets: The discursive logics of high-stakes testing. *Cambridge Journal of Education, 45*(2), 245–264. doi:

10.1080/0305764X.2014.936826

Marchand, G. C., & Furrer, C. J. (2014). Formative, informative, and summative assessment: The relationship among curriculum-based measurement of reading, classroom engagement, and reading performance. *Psychology in the Schools, 51*(7), 659–676. doi: 10.1002/pits.21779

Merino, K., & Beckman, T. O. (2010). Using reading curriculum bases measurements as predictors for the measure academic progress standardized test in Nebraska. *International Journal of Psychology, 85*(8), 85–98.

Miller, K. C., Bell, S. M., & McCallum, R. S. (2015). Using reading rate and comprehension CBM to predict high-stakes achievement. *Journal of Psychoeducational Assessment, 33*(8), 707–718. doi: 10.1177/0734282915574028

Richardson, R. D., Hawken, L. S., & Kircher, J. (2012). Bias using maze to predict high-stakes test performance among Hispanic and Spanish-speaking students. *Assessment for Effective Intervention, 37*(3), 159–170. doi: 10.1177/1534508411430320

Sadler, D. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144. doi: 10.1007/BF00117714

Thornblad, S., & Christ, T. (2014). Curriculum based measurement of reading: Is 6 weeks of daily progress monitoring enough? *School Psychology Review, 43*(1), 19–29.

Utah State Board of Education. (2015a). *Information for educators on Utah's SAGE assessment system.* Retrieved from http://sageportal.org/wp-content/uploads/2014/02/UTEducatorbrochure-Final.pdf

Utah State Office of Education. (2015b). *State of Utah October 1st enrollment counts for school year 2015-2016*. Retrieved from http://www.schools.utah.gov/data/Reports/Enrollment-Demographics.aspx

APPENDIX A: EXTENDED LITERATURE REVIEW

Pretest, posttest, grade-level test, pass, fail, above grade level and below grade level are terms of importance to administrators, educators, parents, and students. Assessment is a way of measuring not only a student's academic knowledge, but also a teacher's effectiveness. The multiplicity of assessments in modern American schools represents the demand for accountability to the nation, to states, and to district and school patrons.

Assessment in schools can be formative or summative. Formative assessment is "concerned with how judgements about the quality of student responses (performances, pieces, or works) can be used to shape and improve the students' competence by short-circuiting the randomness and inefficiency of trial-and-error learning" (Sadler, 1989, p. 120). Educators use formative assessment to determine where students are functioning after a lesson or unit in the classroom. Formative data can be used by teachers to alter or continue instructional strategies depending on student progress. Therefore, the purpose of formative assessment is to drive instructional decisions. Summative assessment "contrasts with formative assessment in that it is concerned with summing up or summarizing the achievement status of a student, and is geared towards reporting the end of a course of study" (Sadler, 1989, p. 120). Educators use summative assessment to show achievement and growth at the end of instructional units, courses, or from one school year to the Next. Formative assessments are usually informal and often created by teachers, whereas summative assessment can be informal or standardized.

Standardized Testing

Standardized tests are those that insure the same conditions for all participants by using specific test items common to all test takers and that use directions common to each administration (Glossary of Educational Reform, 2015). Test results compare individuals or

groups to establish normative standards and are the commonly used tools of accountability in public education. As educational practices continue to evolve, so does the multitude of assessments used in schools and districts. Testing that impacts school, district, and state standing within the requirements of the No Child Left Behind Act (NCLB) is often called high-stakes testing (Lewis & Hardy, 2014). Haladyna (2006) defines high-stakes testing as "the use of test scores that have a significant consequence for students, teachers, schools, and school districts," including decisions about "graduation, promotion, school accountability, federal funding, No Child Left Behind, and merit pay" (p. 23). Performance on high-stakes testing are used to grade districts, schools, and even teachers in ways that inform students, teachers, and administrators. Good performance on valid and reliable tests can indicate successful teaching and leadership, while poor performance can label schools as ineffective. These data interest teachers and administrators who recognize that "reputational capital" is important to build and protect in test-based accountability systems (Lewis & Hardy, 2014 p. 259).

Unfortunately, high-stakes tests can cause concern for students, teachers, administrators and families who lack access to, or understanding of, factors beyond percentile ranks. Because standardized tests do not take all human and environmental variables into consideration, results may be difficult to interpret. Haladyna (2006) highlighted key influences on student performance that are not indicated by test scores including student motivation, test-taking skills, differences in classroom instruction, student strategies for preparation, differences in test construction, and variation in the results of hand-scored tests. Studies of specific performance variables on high-stakes testing identified test anxiety (Segool, Carlson, Goforth, von der Embse, & Barterian, 2013; von der Embse & Witmer, 2014) and student boredom with test preparation (Mora, 2011). These variables may affect the complexity of test results and the inferences based

on them, but the effects are difficult to measure (Brennan, 2015). Still, standardized assessments generally provide the reliability of test scores demanded by stakeholders even though the validity may be questioned (Brennan, 2015).

History of Standardized Testing in the US

Until the mid-1800s there was no uniform system of public schooling in the US; therefore, there was no systematic assessment of student progress. Most of the population lived in small rural communities in which schools were ungraded, sporadically attended, and led by one adult who was usually not trained for teaching (Reese, 2013). This began to change when Horace Mann became the secretary of the newly formed Massachusetts Board of Education in 1837. Mann's investigations of existing schools in America and in Europe convinced him that the only democratic way to educate America's youth was to establish a school system with common standards. In 1843 he recommended that the Board control textbooks used in schools to insure that students had uniform access to information (Mann, 1843). He then applied the same reasoning to testing. At the time, students were assessed by memorizing book passages and reciting them when called upon (Mann, 1845). Mann realized that recitation depended more on mechanical recall than on thinking skills, and he believed that students attended to content only as needed to recite it to the teacher. In 1845, he recommended written questions common to students by grade and subject that could be compared impartially (Mann, 1845). This early move toward standardization eventually led to modern testing.

Early in the 20th century, psychologists began using scientific methods for assessing intelligence (Giordano, 2005). Binet, Galton, Terman and others designed tests that produced scores and could be compared across individuals. Soon academicians began to apply the same science to academic studies (Cubberly, 1934; Giordano, 2005). Standardized testing became a

way to determine if schools were fair and efficient (U.S. Congress, 1992). Parents needed to know that their children's education was equivalent to others' education, and government and administrative stake holders needed assurance that schools were orderly and efficient in educating all children. The 1965 Elementary and Secondary Education Act was the U.S. government's first mandate for using standardized testing to measure and compare students, requiring it for Title 1 schools (Scott, 2004). Since that time, standardized testing has become the accepted and primary means for measuring and reporting accountability in the nation's schools (Gallagher, 2003; Wiliam, 2010).

Curriculum-Based Assessment and Measurement

Curriculum-based assessment differs from standardized testing in that it is used by educators to track and improve student performance. These data show the teacher the progress and effectiveness that the curriculum and instruction are having on the student (Deno, 1992). By teachers tracking the effectiveness of instruction in the classroom, instructional methods can be altered to find the best positive projective. This also provides the teachers with data that can be used as evidence of where the students have been and where they are going (Marchand & Furrer, 2014).

There are two general types of curriculum-based measures: curriculum-based assessment and curriculum-based measurement. Curriculum-based assessment is commonly defined as one-time testing performed at the end of a unit, semester or year (Deno, 1992). In contrast, curriculum-based measurement is used repeatedly to assess the growth in student learning and can be compared to formative assessment for demonstrating progress.

Nese, Park, Alonzo, and Tindal (2011) applied curriculum-based measurement as a predictor of high stakes test scores. By examining the relationship between curriculum-based

benchmarks for passage reading fluency, vocabulary, and reading comprehension; and state end-of-year assessments reading scores, Nese et al. found that teachers could reliably use the benchmark scores to identify students at risk for failing the state test.

Another study demonstrated the effects of curriculum-based measurement with end-of-year testing. Marchand and Furrer (2014) analyzed data from 750 students in grades 3, 4 and 5 using teacher and student reports of student classroom engagement and Nevada state criterion-referenced reading test scores. Student engagement was reported by both teachers and students using a four-point scale with options ranging from *not true* to *true*. Results showed that teacher reports and oral reading fluency scores from CBM-Reading strongly predicted students' scores for informational, literary, and functional text on the state end-of-year test. Findings also showed that student engagement added to the importance of increasing performance of students with reading difficulties.

Additional studies have investigated the amount of data needed from curriculum-based measurement to establish a valid growth estimation in reading scores (Jenkins, Graff, & Migloretti, 2009). The authors found that the amount of progress monitoring in classrooms could be greatly reduced and still produce adequate information to produce a valid projected score. The authors looked at recording scores every third through ninth week to see if they could still create the optimal projective rate. They found that teachers who record data points one or two times a week can decrease to every other week for all students. Using less time to measure progress gave teachers more time for instruction.

Curriculum-based measurement research has also been applied with maze measures of reading comprehension. To assess the reliability of curriculum-based measurement with reading comprehension, Espin et al. (2010) had students read passages aloud for one, two or three

minutes. After completing the oral reading, the students completed a two- to four-minute maze task in which they selected one of three word choices to fill in each blank on the sheet. The authors then looked to see if these scores were compatible with the state testing scores. Results indicated that the reading comprehension scores were predictors of the state end-of-level test.

Progress Monitoring

Progress monitoring is an application of curriculum-based measurement that allows teachers to see and create change in a student's performance over a period of time (Deno, 1992). Progress monitoring is used in schools to track the progress of students in between benchmark assessments, as well as through the year, to project whether or not individual students will achieve benchmarks or be on grade level by the end of the year. This approach is not just a record of the students' achievement. It also measures the effects of teachers' practices in the classroom through their teaching strategies (Deno, 1992). Progress monitoring indicates the effectiveness of classroom instruction by tracking students' performance to demonstrate what areas need improvement or need to be taught in more depth for students to reach proficiency. Instruction is then targeted at the specific point in the academic learning that is weak (Catts, Nielsen, Bridges, Liu, & Bontempo, 2015).

Progress monitoring results are usually graphed to depict the slope of student achievement over time (Deno, 1992). When the target is a specific benchmark or aim point, then the slope indicates whether or not students are on track to meet achievement objectives. Progress monitoring can also be quite indicative in a short period of time. Thornblad and Christ (2014) found that six weeks of daily monitoring produced valid and reliable indicators of student growth for reading. Jenkins et al. (2009) noted that less frequent monitoring will also produce accurate predictions provided that a sufficient number of scores are used.

Repeated measures over time require alternate forms of progress monitoring assessments to avoid test-retest gain (Catron & Thompson, 1979) and writers have emphasized the importance of verifying the reliability of alternate forms (Espin, Wallace, Lembke, Campbell, & Long, 2010). This is especially important when using progress monitoring to predict outcomes on high-stakes tests (Espin et al., 2010; Nese et al., 2011). Researchers have found that while forms may be comparable, it is more difficult to achieve equivalence. Betts, Pickart, and Heistad (2009) found that alternate forms of curriculum-based measurement passages for reading were not equivalent when compared using identical raw scores, but the results across forms were comparable within grade levels.

Researchers have used progress monitoring to predict scores on summative standardized tests. Richardson, Hawken, and Kircher (2012) used maze comprehension testing to compare students who speak predominantly Spanish at home with native English speakers, and to compare Hispanic student outcomes to those of Caucasian students. The authors found that maze measures under predicted scores on end-of-year high stakes tests for Hispanic students and non-native English speakers, but accurately predicted outcomes for Caucasian students and native English speakers. Curriculum-based measures of comprehension used with Tennessee 3rd grade readers demonstrated strong predictive power for end-of-level tests (Miller, Bell, & McCallum, 2015), as did oral reading and maze measures for Nebraska students in grades two through five (Merino & Beckman, 2010).

Dynamic Indicators of Basic Early Literacy Skills

The Dynamic Indicators of Basic Early Literacy Skills (DIBELS Next) measure reading skills from kindergarten through 6th grade. The National Reading Panel (Armbrustur, Lehr, & Osborn, 2001) identified five critical components for successful reading including phonemic

awareness, phonics, fluency, vocabulary, and text comprehension. DIBELS NEXT assesses each of these components and can be administered in a few minutes, providing a quick view of student ability that informs early intervention (Good & Kaminski, 2011).

DIBELS NEXT was designed in the late 1980's based on curriculum-based measurement as a result of initial research at the University of Oregon. Curriculum-based measurement and DIBELS NEXT both aim for an economical and efficient way to demonstrate student growth towards achievement. DIBELS NEXT, the current version, measures six areas in reading that can be monitored for progress over time, including first sound fluency, letter naming fluency, phoneme segmentation fluency, nonsense word fluency, oral reading fluency, and text comprehension (Good & Kaminski, 2011). Elliot, Lee and Tollefson (2001) described the skill sets in a study of the validity and reliability of DIBELS NEXT:

> The DIBELS NEXT evaluate a set of early literacy skills identified in the literature as directly related to and facilitative of later reading competence. Student knowledge of the letter names, sound-symbol relationships, and phonemic awareness in kindergarten have all been identified as important predicators of later literacy. (p. 34)

DIBELS NEXT Reliability and Validity

Elliot et al. (2001) tested the reliability and validity of DIBELS NEXT using a population of 75 kindergartners from four classrooms in three different elementary schools. The authors used repeated administrations to calculate reliability and validity of the tests as a whole as well and for individual subtests. Table 1 shows that the subtests have relatively strong reliability between interraters and from one test episode to the next, but less so for equivalent forms.

Table 1

Single-Passage and Three-Passage (Triad) Alternate-Form Reliability for DIBELS NEXT ORF Benchmark Passage from Study D

Grade	Number of Students	Number of Passages	Median Passage Means	Median Passage SD's	Median Triad Means	Median Triad SDs	Median Reliability	
							Single-passage	Triad
Third	22	32	109.89	39.13	110.44	38.01	.93	.97
Fourth	23	32	131.87	31.99	132.47	31.01	.90	.94
Fifth	23	32	136.24	36.07	137.33	34.62	.92	.96
Sixth	24	32	150.99	28.63	148.02	27.63	.84	.90

Note. Based on Good, Kaminsky, Dewey, Wallin, Powell-Smith, & Latimer, 2013, p. 89.

Reliability is important because progress monitoring requires the comparison of scores from one assessment to the Next. Reliability is the statement of consistency of results across multiple measures. If assessments do not contain equivalent content for each alternate form, then students can be proficient in a test one day and not do as well the Next. The Elliot et al. (2001) data indicate a strong reliability for the DIBELS NEXT subtests.

Kaufman Test of Educational Achievement

The Kaufman Test of Educational Achievement (KTEA) was developed by clinical psychologists Alan S. Kaufman, PhD and Nadeen L. Kaufman, PhD. with the original name of K-ABC (Pearson, 2016). Currently in its third edition, the KTEA is a norm-referenced test designed to compare student achievement to that of other students in the same age and grade (Cole, 2012). Each test is designed with an increase of testing time based on grade level: preK-K are given a 30 minute testing session, grades 1–2 a 50 minute testing session, and grade 3 and above an 80 minute session.

KTEA subtests include comprehensive achievement, reading, math, decoding, written language, sound symbol, oral language, and oral language fluency. Subtest scores are combined to form a composite score. Separate types of questions within each section measure students' knowledge. For example, the math composite score is a combination of math concepts and application, math computation, and math fluency, all of which encompass multiple topic areas (Kaufman & Kaufman, 2004).

KTEA-3 Brief Reliability

The KTEA-3 Brief form includes letter and word recognition along with reading comprehension. The reliability of each test score is a statement of its ability to consistently reproduce similar results. In order for test results to be reliable, valid and interpretable, they must be relatively resistant to errors of measurement that could cause the score to change substantially if the child were retested on a different occasion or with an alternative set of items (Kaufman, 2004). Reliability is based on four key factors: (a) the items measure the same skill (b) the extent that performance on tasks is determined by the skill being measured, (c) the number of items being measured (the larger the quantity of items the higher the reliability), and (d) the consistency with which a person responds to the same or parallel items in different situations (Kaufman, 2004). Strong reliability allows educators to use data to develop instruction to help students achieve desired goals.

All but two KTEA subtests have reliability coefficients. Word recognition fluency and decoding fluency provide only one opportunity for students to complete the task, so reliability has not been determined. The reliability of the KTEA with all the subtests was calculated to find the overall stability of the test over time using the four criteria listed above (Kaufman, 2004) and is reflected in terms of internal consistency. The internal consistency of the KTEA was

calculated using Cronbach's Alpha applied to the reliabilities of each subtest, which are .90 and above for the KTEA. The breakdown of consistencies include comprehensive achievement composite (.97), reading composite (.96), math composite (.96), decoding composite (.97), and written language composite and sound symbol composite (.93). The oral language composite and oral fluency composite are the only two that fall below .90, but are still a respectable .87 for the oral language composite and .85 the oral fluency composite (Kaufman, 2004). Overall, the KTEA has a high probability of reliable measurement over multiple occasions and settings.

Student Assessment for Growth and Excellence (SAGE)

The Student Assessment for Growth and Excellence (SAGE) is a computer-adaptive summative test aligned with the Utah core curriculum and used as the end-of-year high-stakes assessment in Utah schools. SAGE tests reading, language arts, math, science, and writing in grades 3 through 6. In reading, students are tested on listening comprehension as well as the reading standards that align with the student's grade level. SAGE is an adaptive test, meaning that the testing software analyzes a student's answer to one question to determine the question that will follow (Utah State Board of Education, 2015a). Therefore, "the difficulty of the test will adjust to each student's skills, providing a better measure of what each student knows and can do" (Utah State Board of Education, 2015a, p. 2).

SAGE also offers optional fall and winter versions that schools can use to inform instruction and monitor student progress. A formative teacher instruction manual with guidance on classroom instruction is available, but not required for use. The computerized SAGE requires students to create "graphs, interact with science simulation, and write and respond in multiple different ways. These question types will assess higher-order thinking skills" (Utah State Board of Education, 2015b, p. 1). With the Utah State Core relying more upon higher order analysis

and problem solving skills, these questions provide a more in depth questioning assessment. By having the students think in this manner, Utah is planning to prepare students for careers of the 21st century (Utah State Board of Education, 2015b).

During the first year of SAGE testing in Spring 2014, processing the results took a long time. Results are now calculated electronically and available soon after students have completed the assessment (Utah State Board of Education, 2015b). The first year of testing also returned lower scores due to more rigorous standards and student unfamiliarity with the testing format (Utah State Board of Education, 2015b). The State office maintains that "students' proficiency will increase as students, parents, and educator's work together to implement the new standards and assessments" (Utah State Board of Education, 2015a, p. 1).

The Utah State Office of Education reports that no norming procedures were conducted prior to implementing the assessment. Instead, the State is collecting norming data as the tests are implemented in the schools. Therefore, the reliability and validity of SAGE was yet to be assessed or verified in 2015.

References

Betts, J., Pickart, M., & Heistad, D. (2009). An investigation of the psychometric evidence of CBM-R passage equivalence: Utility of readability statistics and equating for alternate forms. *Journal of School Psychology, 47*, 1–17. doi:10.1016/j.jsp.2008.09.001

Brennan, R. L. (2015). Testing for accountability: A balancing act that challenges current testing practices and theories. *Measurement: Interdisciplinary Research and Perspectives, 13*(1), 35–38.

Catron, D. W., & Thompson, C. C. (1979). Test-retest in WAISS scores after four intervals, *Journal of Clinical Psychology, 35*(2), 352–357.

Catts, H. W., Nielsen, D. C, Bridges, M. S., Liu, Y. S., & Bontempo, D. E. (2015). Early identification of reading disabilities within and RTI framework. *Journal of Learning Disabilities, 48*(3), 281–297.

Cole, A. (n.d.). *KTEA*. Retrieved from http://amandacoleassessment.weebly.com/ktea.html

Cubberly, E. P. (1934). *Public education in the United States; A study and interpretation of American educational history* (rev. ed.). Boston, MA: Houghton Mifflin.

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

Espin, C., Wallace, T., Lembke, E., Campbell, H., & Long, J. D. (2010). Creating a progress-monitoring system in reading for middle-school students: Tracking progress toward meeting high-stakes standards. *Learning Disabilities Research & Practice, 25*(2), 60–75.

Gallagher, C. J. (2003). Reconciling a tradition of testing with a new learning paradigm. *Educational Psychology Review, 15*(1). 83–99.

Giordano, G. (2005). *How testing came to dominate American schools*. New York, NY: Peter Lang.

Glossary of Educational Reform. (2015). *Standardized Testing*. Retrieved from http://edglossary.org/standardized-test/

Good, R. H., & Kaminski, R. A. (2011). *DIBELS NEXT assessment manual*. Eugene, OR: Dynamic Measurement Group.

Good, R. H., Kaminski R. A., Dewey, E. N., Wallin, J., Powell-Smith, K. A., & Latimer, R. J. (2013). *DIBELS NEXT technical manual.* Eugene, OR: Dynamic Measurement Group.

Haladyna, T. M. (2006). Perils of standardized achievement testing. *Educational Horizons, 85*(1), 30–43.

Jenkins, J., Graff, J., & Migliorette, D. (2009). Estimating reading growth using intermittent CBM progress monitoring. *Exceptional Children, 75*(2), 151–163.

Kaufman, A. S (2016). *Clinical psychology*. Retrieved from http://www.pearsonclinical.com/psychology/authors/kaufman-alan.html

Kaufman, A. S, & Kaufman, N. L. (2004). *K-TEA II: Kaufman Test of Educational Achievement: Comprehensive form* (2nd ed.). Circle Pines, MN: AGS Pub.

Kaufman, A. S, & Kaufman, N. L. (2014). *K-TEA III: Kaufman Test of Educational Achievement: Form A* (3nd ed.). Bloomington, MN: PsychCorp.

Lewis, S., & Hardy, I. (2014). Funding, reputations, and targets: The discursive logics of high-stakes testing. *Cambridge Journal of Education, 45*(2), 245–264.

Macmillan/McGraw-Hill. (2007). *Treasures.* New York, NY: Author.

Mann, H. (1843). *Seventh annual report of the Board of Education together with the seventh annual report of the secretary of the board.* Boston, MA: Dutton and Wentworth. Retrieved from http://hdl.handle.net/2027/uma.ark:/13960/t2g740j6p

Mann, H. (1845). *Ninth annual report of the Board of Education together with the seventh annual report of the secretary of the board.* Boston, MA: Dutton and Wentworth.

Marchand, G. C., & Furrer, C. J. (2014). Formative, informative, and summative assessment: The relationship among curriculum-based measurement of reading. Classroom engagement, and reading performance. *Psychology in the Schools, 51*(7), 659–676.

Martella, R. C., Nelson, J. R., Morgan, R. L, & Marchand-Martella, N. E. (2013). *Understanding and interpreting educational research.* New York, NY: Guilford.

Merino, K., & Beckman, T. O. (2010). Using reading curriculum based measurements as predictors for the measure academic progress standardized test in Nebraska. *International Journal of Psychology, 85*(8), 85–98.

Miller, K. C., Bell, S. M., & McCallum, R. S. (2015). Using reading rate and comprehension CBM to predict high-stakes achievement. *Journal of Psychoeducational Assessment, 33*(8), 707–718.

Mora, R. (2011). "School is so boring." High-stakes testing and boredom at an urban middle school. *Penn GSE Perspectives on Urban Education, 9*(1), 35–42.

Nese, J. F. T., Bitnara, J. P, Alonzo, J., & Tindal, G. (2011). Applied curriculum-based measurement and predictor of high stakes assessment. *The Elementary School Journal, 111*(4), 609–624.

Reese, W. J. (2013). *Testing wars in public schools.* Boston, MA: Harvard University.

Richardson, R. D., Hawken, L. S., & Kircher, J. (2012). Bias using maze to predict high-stakes test performance among Hispanic and Spanish-speaking students. *Assessment for Effective Intervention, 37*(3), 159–170.

Sadler, D. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*, 119–144.

Scott, T. (2004) Teaching the ideology of assessment. *Radical Teacher, 71*(4), 30–37.

Segool, N. K, Carlson, J. S., Goforth, A. N., von der Embse, N., & Baterian, J. A. (2013). Heightened test anxiety among young children: Elementary school students' anxious responses to high-stakes testing. *Psychology in the Schools, 50*(5), 489–499.

Thornblad, S., & Christ, T. (2014). Curriculum based measurement of reading: Is 6 weeks of daily progress monitoring enough? *School Psychology Review, 43*(1), 19–29.

U.S. Congress. (1992). *Testing in American schools: Asking the right questions, OTA-SET-519*. Washington, DC: Office of Technology Assessment.

Utah State Board of Education, (2015a). *Information for educators on Utah's SAGE assessment system.* Retrieved from http://sageportal.org/wp-content/uploads/2014/02/UTEducatorbrochure-Final.pdf

Utah State Office of Education. (2015b). *State of Utah October 1st enrollment counts for school year 2015–2016*. Retrieved from http://www.schools.utah.gov/data/Reports/Enrollment-Demographics.aspx

Utah State Office of Education. (2015c). *National school lunch program October survey*. Retrieved from http://schools.utah.gov/data/Reports/Child-Nutrition.aspx

Von der Embse, N., & Witmer S. (2014). High stakes accountability: Student anxiety and large scale testing. *Journal of Applied School Psychology, 30*(2), 132–156.

Wiliam, D. (2010). Standardized testing and school accountability. *Educational Psychologist, 45*(2), 107–12.