INVESTIGATING THE SELECTED VALIDITY OF AUTHENTIC ASSESSMENT IN WRITTEN LANGUAGE FOR STUDENTS WITH AND WITHOUT LEARNING DISABILITIES

DISSERTATION

Presented to the Graduate Council of the

University of North Texas in Partial

Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Pamela K. Peak, B.S., M.Ed.

Denton, Texas

August, 1994

INVESTIGATING THE SELECTED VALIDITY OF AUTHENTIC ASSESSMENT IN WRITTEN LANGUAGE FOR STUDENTS WITH AND WITHOUT LEARNING DISABILITIES

DISSERTATION

Presented to the Graduate Council of the

University of North Texas in Partial

Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY

By

Pamela K. Peak, B.S., M.Ed.

Denton, Texas

August, 1994

Peak, Pamela K., Investigating the Selected Validity of Authentic Assessment in Written Language for Students With and Without Learning Disabilities. Doctor of Philosophy (Special Education), August, 1994, 99 pp., 6 tables, bibliography, 154 titles.

Methods of assessing educational growth have not kept pace with the changing curriculum. The vision that is directing many of the current efforts to transform assessment is provided by authentic assessment practices. These assessments seek to display student performance on meaningful and challenging tasks as close as possible to real world responsibilities.

This research study was designed to investigate whether authentic assessment in written language is a valid assessment tool for students with and without learning disabilities. Teacher judgements were used to evaluate students' authentic writing assessments gathered from the classroom. Students' report card grades, authentic writing assessments, and two standardized writing assessments, the Test of Written Language-Revised and Written Language Assessment, were correlated to provide evidence of the validity of authentic assessment practices in written language.

The subjects for this study (N = 84) were drawn from the population of students in a large urban school district in North Central Texas. Subjects who were in fourth or fifth grade, had never been retained, and were receiving no special education services (N = 46) made up the sample of students without learning disabilities. Subjects who were in the fourth and fifth grade and had a diagnosed learning disability in language arts according

to Texas state guidelines (N = 38) made up the sample of students with learning disabilities.

Correlation coefficients indicated that the various types of validity related to authentic assessment in written language are only minimally supported. Teachers have the expertise to produce an accurate evaluation of students' progress. Furthermore, the evidence pertaining to the assumption that because authentic assessment encompasses a basic school subject, students who do well in writing should do well in other areas of school supported the construct validity of authentic assessment. Nevertheless, the findings fail to present evidence that the test statistics for criterion-related concurrent validity differ for students with and without learning disabilities.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION TO THE STUDY

The present focus on authentic assessment and the measurement of student outcomes grew out of a number of national reports such as A Nation At Risk: The Imperative for School Reform (National Commission on Excellence in Education, 1983) and A Time for Results (National Governors' Association, 1986) and national test data that focused public attention on what was considered the mediocre education of the nation's students. A growing body of research addresses the issue of authentic assessment (Adams & Hamm, 1992; Barrett, 1992; Bracey, 1993; Hambleton & Murphy, 1992; Herman, Aschbacher, & Winters, 1992; Jongsma, 1989; Levi, 1990; Linn, 1993; Linn, Baker, & Dunbar, 1991; Maeroff, 1991; O'Neil, 1992; Paulson, Paulson, & Meyer, 1991; Quinta & McKenna, 1991; Simmons, 1990; Valencia, 1990; Vavrus, 1990; Wiggins, 1989; Worthen, 1993; Worthen & Spandel, 1991).

"Most countries we compete with in Europe and Asia that out achieve us use essays, oral exams and exhibits of students' work" ("Not as Easy as ABC," 1990). In response to increasing concern over American students' poor international standing, the National Commission on Testing and Policy (1990) has recommended that alternative forms of assessment be adopted in American schools. More and more, American educators are insisting that assessment practices become more authentic (i.e., assess

meaningful skills and abilities in a realistic and integral way that enable students to become successful, productive adults) (Hacker & Hathaway, 1991).

Critics and Proponents of Traditional Assessment

Traditional assessment in American education has led to much controversy. Traditional testing practices embrace a variety of approaches to assessment; however, in the following discussion, the phrase traditional testing refers specifically to standardized, norm-referenced examinations (Moody, 1991). While some researchers are in favor of more standardized testing (Killoran, 1992; Miller, 1991; Wilson, 1991), many researchers are opposed to the idea (Adams & Hamm, 1992; Farr, 1992; Linn, 1993; Worthen, 1993).

Standardized testing is a pervasive part of American education (Hartle & Battaglia, 1993). The National Commission on Testing and Public Policy (1990) has estimated that each year elementary and secondary school students take 127 million separate tests. Some students may take as many as 12 tests a year. Researchers have found that essays written under controlled settings are more valid predictors of students' success than essays written by students in a more naturalistic setting (Anastasi, 1988). Some of the advantages of traditional tests include the ability to (a) demonstrate a student's progress over time (Mercer & Corbett, 1991), (b) depict a student's strengths and weaknesses (Wallace & Larsen, 1978), and (c) compare a student's performance to a measured ability to determine over/under-achievement (Weiderholt, Hammill, & Brown, 1983).

The tendency for assessment to shape instruction and learning is not necessarily negative. However, within American education, the types of competencies measured with traditional assessment tools do not match well with the competencies mandated by parents, legislators, and the nation at large. On the one hand, the nation is demanding that public educators teach and reinforce a broad array of academic and nonacademic competencies; on the other, traditional forms of assessment focus on a fairly narrow range of academic abilities. For instance, the national goals established by the National Governor's Association at the First Educational Summit meeting in 1990 cited competencies such as learning to utilize one's mind and complex reasoning as two areas that must be improved by the year 2000. Similarly, the report from the Department of Labor entitled What Work Requires of School: A SCANS Report of America 2000 (Secretary's Commission on Achieving Necessary Skills, 1991) lists a broad array of both academic and nonacademic competencies, such as (a) creative thinking, (b) decision making, (c) problem solving, (d) seeing in the mind's eye, and (e) self management. Other national and regional reports have identified similar sets of diverse competencies (Eisner, 1991; Resnick, 1987; Stiggins, 1991).

Controversy and skepticism have followed the utilization of traditional testing. Critics argue that the information measured by traditional assessment tools is not indicative of the learning within the classroom (Rogers, 1989; Worthen, 1993). Traditional tests assume a theory of learning that is incompatible with current understanding; one in which knowledge is viewed as discrete and hierarchically arranged, and one in which increased expertise is represented along a single dimension (Shepard,

1991; Wolf, Bixby, Glenn, & Gardner, 1991). Current standardized testing practices have also (a) been unfair to some students; (b) given false information; (c) corrupted the processes of teaching and learning; and (d) focused time, energy, and attention on simpler skills (Haney & Madaus, 1989).

Another related set of concerns focuses on the effects of traditional forms of assessment on the educational context. When high-stakes decisions are attached to test scores, as has been the case through the 1980s, assessments often determine educational goals. The indicators of achievement then become confused with goals; as a result, they lose their value as indicators while distorting movement toward more fundamental goals (Linn, Baker, & Dunbar, 1991). The effect is most damaging if performance on a test bears little resemblance to the performances seen as essential to education. Thus, educators who believe that learning involves complex performances in which students integrate a variety of skills and call on different kinds of knowledge and mental representations are likely to see the emphasis on multiple-choice test scores as particularly detrimental to genuine learning.

In a series of studies at the Mid-Continent Regional Educational Laboratory, 6,942 items from the Stanford Achievement batteries and the California Test of Basic Skills were analyzed to determine the extent to which they assess general cognitive competencies (Marzano, 1990; Marzano & Costa, 1988; Marzano & Jesse, 1987). Of the 22 general cognitive competencies considered for the study, items in the 2 test batteries assessed only 9 of them. Of the 9 competencies covered, retrieval or recall of information was the mental ability most commonly assessed by a factor of 5-to-1 relative to the next

most commonly assessed cognitive competency. In brief, a growing body of researchers support the assertion that the current systems of standardized tests measure a fairly narrow range of competencies which are limited to the academic domain.

The implementation of standardized testing has increased dramatically and so have the complaints. Even though an increasing number of educators are annoyed with the utilization of traditional assessment practices, American students still remain the most tested, least examined students in the world (Resnick, 1989). Dissatisfaction has led researchers and practitioners to turn toward a more authentic assessment (Farr & Carey, 1986; Haney & Madaus, 1989; McClennon, 1988; Valencia & Pearson, 1987).

Critics and Proponents of Authentic Assessment

Methods of assessing educational growth have not kept pace with the changing curriculum (Quinta & McKenna, 1991). The vision that is guiding many of the current efforts to transform assessment is provided by authentic assessment practices. The phrase authentic assessment refers to the gathering and evaluation of evidence of student performance which is produced in an integrated manner and in a naturalistic time frame and context (Archibald & Newmann, 1988). These assessments attempt to reveal student performance on meaningful and challenging tasks that are as close as possible to real world responsibilities.

A number of advantages support the use of authentic assessment. One of the greatest advantages claimed for authentic assessment is that it can test directly what educators want children to know. Assessing academic skills in context is beneficial to

students' overall education (Archibald & Newmann, 1988; Mitchell, 1989; Snow, 1989; Valencia, 1990). Performances should be based in meaningful tasks---tasks that are complex and challenging, that are consistent with goals for learning and inherently valuable to learning, that are closely related to real-world skills and challenges, and that allow students to use the processes and strategies that are relevant to genuine performance. Authentic assessment recognizes higher order thinking skills, personal judgement, and collaboration techniques. Evaluating students' ability to speak, write, analyze, and do experiments provides a clearer picture of the student as a learner. Authentic assessment is designed to create an environment in which students can "show" what they know, leaving the power in their hands and allowing them to utilize higher thinking skills (Levi, 1990; Valencia, 1990).

Finally, authentic assessment methods, which examine actual student work samples, have received considerable attention because they are considered to have high face and content validity (Charney, 1984; Moran, 1987). These desirable characteristics can be more feasibly incorporated in the assessment performance if they occur within the immediate educational context (i.e., the classroom or school). In addition, when the assessment of performance becomes part of the life of the educational institution, the procedures associated with it can inform and be informed by the educational community that is likely to feel its effects most acutely (Camp, 1993). Although the costs of standardized testing are astonishing, critics argue that authentic assessment is much more costly. When compared to computer scored, multiple choice items, considerably more

time is necessary on the teacher's and student's part when analyzing authentic assessment (Valencia, 1990).

The new views of educational assessment are based on theories of learning that are unlike those assumed by conventional test theory. It is not surprising, therefore, that they have created some controversy about measurement issues. Authentic assessment lacks quantitative analysis. Policy makers want percentiles and quotients to use in modifying programs and budgets (Hacker & Hathaway, 1991; Levi, 1990). Without these exact measurements, policy makers find it difficult, if not impossible, to implement authentic assessment procedures.

Validity of Assessment Instruments

Validity, which means a test measures what it purports to measure, is the single most important aspect of a test (Anastasi, 1988; Mehrens & Lehmann, 1987). Validity reveals whether a test measures what it claims to measure, how well the test measures what it claims, and what can be inferred from that measurement. Additionally, the concept that validity is a matter of inferences, not direct measures, is widely recognized. The determination of validity hinges on the uses to which a measure is put, as well as the care that has gone into its development. Thus, validity can only be determined in the context of the relationship between the specific uses to which the test results will be put and the construct that is being measured. Information and conclusions regarding the validity of a given test in one context may not be relevant and applicable in other contexts (Gronlund & Linn, 1990; Messick, 1989; Neill & Medina, 1989).

Because the validity of a test's results is relevant and contingent upon the purpose for which the test will be used, an assortment of validity evidence should be accumulated. Authors of educational and psychological measurement textbooks have suggested that at least three types of validity (e.g., content validity, criterion-related validity, and construct validity) should be present in current assessment measures (Anastasi, 1988; Gronlund & Linn, 1990; Salvia & Ysseldyke, 1991).

Validity and other statistical measures of traditional, norm-referenced tests are usually established on groups of normal children. The response limitations of many children with disabilities seriously interfere with performance on such measures and make it difficult to obtain accurate estimates of abilities and precise measures of reliability and validity (Keogh & Sheehan, 1981). Assessment specialists must find ways to circumvent these difficulties in order to secure measures of functional status. While a number of options are available, a valid option is to include norm-referenced, criterion-referenced, and judgement-based assessment measures from multiple sources across multiple settings (Bagnato, Neisworth, & Munson, 1989).

Purpose of the Study

The purpose of this research study was to investigate the selected validity of authentic assessment in written language to determine whether types of validity vary between individuals with and without learning disabilities (LD).

Significance of the Study

The new views of educational assessment have created some controversy about measurement issues. Authentic assessment lacks quantitative analysis, and policy makers want percentiles and quotients to use in modifying programs and budgets (Hacker & Hathaway, 1991; Levi, 1990). Without these exact measurements, policy makers find it difficult, if not impossible, to implement authentic assessment practices. As states begin to adopt authentic assessment practices, there is a need to inspect the selected validity of authentic assessment in written language.

The use of discrepancy formulas and timed, multiple-choice tests in the referral and identification processes of students with LD have peaked concerns (Council for Learning Disabilities, 1986; Fuchs & Fuchs, 1990; Shepard, 1989). Also, as various districts move toward an inclusive environment and develop partnerships with regular education in meeting the needs of students' with disabilities, it is likely that students with LD will be included in authentic assessment practices. Therefore, determining whether types of validity vary between individuals with and without LD is essential. In this study, the selected validity of authentic assessment in written language is investigated and determination of whether types of validity vary between individuals with and without LD is made.

Problem Statement

The problem statement for this study was "Does the selected validity of authentic assessment practices vary between individuals with and without LD? If they do, what areas of validity are similar, and what areas of validity differ?"

Limitations

Identifying students with and without LD who were willing to participate and who met the specific criteria for subject selection was a limitation of this study. The scoring procedures of the standardized and authentic writing assessments were also a limitation of the study due to the possible bias of the scorer.

Definition of Terms

The following definitions are used in this study:

Traditional assessment refers specifically to standardized, norm-referenced examinations (Moody, 1991).

Authentic assessment refers to the gathering and evaluation of evidence of student performance which is produced in an integrated manner and in a naturalistic time frame and context (Archibald & Newmann, 1988).

Students with learning disabilities

> exhibit a disorder in one or more of the basic psychological processes involved in understanding or in using language, spoken or written, which may manifest itself in an imperfect ability to listen, think, speak, read, write, spell or to do mathematical calculations. The term includes such conditions as perceptual handicaps, brain injury, minimal brain dysfunction, dyslexia, and developmental aphasia. The term does not include children who have learning problems which are primarily the result of visual, hearing, or motor handicaps, of mental retardation, of emotional disturbance, or of environmental, cultural, or economic disadvantage. (U. S. Department of Education, 1991, pp. 11-12)

Validity is the degree to which a test measures what it claims to measure (Anastasi, 1988; Mehrens & Lehmann, 1987).

Face validity refers to whether a test appears to be a measure of the proposed constructs (Anastasi, 1988) or resembles the construct or skill of interest (Mehrens & Lehmann, 1987).

Content validity is determined by the systematic examination of content to determine if a representative sample depicts the domain to be measured (Anastasi, 1988).

Criterion-related validity is determined by the systematic examination of the relationship of the measurement criteria to some outside criterion. The two types of criterion-related validity are predictive and concurrent validity (McLoughlin & Lewis, 1990).

Predictive validity, a type of criterion-related validity, is a test's ability to predict future performance (Anastasi, 1988).

Concurrent validity, also a type of criterion-related validity, is a test's ability to correlate with some current criterion (Salvia & Ysseldyke, 1991).

Construct validity is a test's ability to measure a particular construct or trait underlying the test (Salvia & Ysseldyke, 1991).

CHAPTER 2

LITERATURE REVIEW

Assessment policies and practices at the local, state, and national levels are in transition. The direct assessment of complex student performances provides the vision that is steering many current efforts to transform assessment. Examples include a strong emphasis on the use of more open-ended problems, essays, hands-on activities, computer simulations of real-world problems, and portfolios of student work. Collectively, such measures are frequently referred to as authentic assessments because they involve the performance of tasks that are appreciated in their own right (Archibald & Newmann, 1988). In contrast, paper-and-pencil, multiple-choice tests derive their value primarily as indicators of other valued performances (Linn, Baker, & Dunbar, 1991). This review of literature is focused on (a) authentic assessment practices, (b) measuring validity, (c) research on the validity of authentic assessment in written language, (d) the importance of teacher judgements, and (e) the importance of investigating the validity of authentic assessment for students with learning disabilities (LD).

Authentic Assessment Practices

The American education system is under excessive scrutiny and pressure from essentially every educational, political, business, or interest group in our society for its alleged faults. This pressure centers researchers attention on an alternative approach to

educational assessment, authentic assessment. The focus on authentic assessment and measurement of student outcomes grew out of a number of national reports and national test data that concentrated public attention on what was considered the mediocre education of the nation's students. The challenge to public education is to produce students who can compete intellectually in the global society (National Commission on Excellence in Education, 1983; National Governors' Association, 1986).

As the pressure mounts for significant educational reforms, the public is demanding more and stronger evidence that such reforms are working to produce students who can think, communicate, and solve problems. Proponents of authentic assessment are willing to accept this challenge. The term authentic assessment means the gathering and evaluation of evidence of student performance which is produced in an integrated manner and in a naturalistic time frame and context (Archibald & Newmann, 1988). This type of assessment of educational achievement directly measures students' academic performance and awards students with the opportunity to be viewed as a whole learner (Levi, 1990; Valencia, 1990).

Authentic assessment includes many writing tasks. Most students who receive services for LD have severe writing difficulties that persist over time (Graham & Harris, 1989). Warner, Alley, Deshler, and Schumaker (1980) found that the majority of the students with LD in their study scored at or below the 10th percentile in written language. This low academic performance differentiates students with LD from other students.

Educational assessment should be an ongoing process rather than an after-the-fact procedure (Wiggins, 1989). Traditional, standardized tests demand hours of instructional

time preparing students for the test. If, in fact, teachers and schools are teaching to the test, then assessments should be developed in which teaching to the test is a valid use of instructional time. Authentic assessment allows teachers to teach to the test without destroying validity. Frederiksen and Collins (1989) argued that systemically valid assessment instruments are those that foster the type of learning and performance that is deemed critical to the educational mission. Assessment activities that are systemically valid are worthwhile learning tasks in and of themselves. Teachers who teach and test directly what children need to know in order to be successful community members can only enhance student learning and benefit society (Wiggins, 1989).

Measuring Validity

As mentioned previously, a growing body of researchers have addressed the issue of authentic assessment; however, much of their research lacks quantitative analysis (Dunbar, Koretz, & Hoover, 1991; Hacker & Hathaway, 1991; Wilson, 1991). Difficulties in authentic assessment stem from the problems encountered in attempts to make assessment results valid, reliable, and comparable. This is the type of information that policy makers need and want to have answers to so that they can modify policies, programs, and resources in productive ways (Hacker & Hathaway, 1991).

Validity, which basically means that a test measures what it claims to measure, is the single most important aspect of a test (Anastasi, 1988; Mehrens & Lehmann, 1987). Measurement experts know that although reliability is necessary, it is not a sufficient condition for validity. Determining the validity of an assessment instrument depends

upon the specific use of the instrument. Assessments can be valid for some purposes, but inappropriate for others (Herman, Aschbacher, & Winters, 1992).

Because the validity of a test's results is relevant and contingent on the purpose for which the test is to be used, an assortment of validity evidence should be accumulated. Authors of educational and psychological measurement textbooks have suggested that at least three types of validity (e.g., content validity, criterion-related validity, and construct validity) should be present in current assessment measures (Anastasi, 1988; Gronlund & Linn, 1990; Salvia & Ysseldyke, 1991). A fourth type of validity, face validity, is typically not considered in establishing a test's validity, even though it is most beneficial in the early stages of constructing a test.

Measuring Face Validity

Face validity refers to whether a test appears to be a measure of the proposed constructs (Anastasi, 1988) or resembles the construct or skill of interest (Mehrens & Lehmann, 1987). Face validity can never take the place of any other form of test validity; however, it is still vitally important because most people react more favorably to assessments that have high face validity (Borg, Worthen, & Valcarce, 1988). Nevo (1985) noted that assessments with high face validity are more apt to (a) bring about higher levels of cooperation and motivation on the part of students; (b) reduce students' feelings of dissatisfaction; (c) help convince potential users (e.g., teachers and school administrators) to implement the test; and (d) improve public relations, because laypersons can more easily see a relationship between the test and the performance.

Measuring Content Validity

Content validity involves the systematic examination of content to determine if a representative sample depicts the domain to be measured (Anastasi, 1988). Obviously, this type of validity has to be incorporated into a test at the time of item development and selection. The determination of content validity is a matter of judgement and is closely tied to the procedures used to construct the assessment tool. By determining the rationale underlying the selection of the testing formats and items and of the statistical procedures used to choose good items, a test has relevance for content validity as well as for item and format selection. Certain questions must be taken into consideration when working with content validity. For instance, (a) What area is the content trying to assess? (b) Is the content assessing the entire universe of content, or a specific portion? (c) If the content is assessing a sample, is the sample representative? and (d) What tasks are being used to assess the content (McLoughlin & Lewis, 1990)?

Evidence of content validity is highly subjective. Only if the selection of items is consonant with the theoretical orientation of a test will examiners agree that the test measures the domain in question. In other words, the authenticity of the task must be closely connected to the task's objective. Unlike reliability, where there is partial external validity, content validity is based on personal judgement (Hresko, 1988).

Measuring Criterion-Related Validity

Criterion-related validity is determined by the systematic examination of the relationship of the measurement criteria to some outside criterion. It is assumed, of course, that the outside criterion is a valid measurement (McLoughlin & Lewis, 1990).

There are two types of criterion-related validity. The first, predictive validity, is a test's ability to predict future performance (Anastasi, 1988). A measure is given to a particular group of students and then, some time in the future, a criterion is administered to the same group of students. For instance, the predictive validity of a readiness test could be established by administering the test to a group of kindergartners. At the end of first grade, the test scores could be correlated to the teacher's judgement of the students' performance. If the teacher's judgement of the students' performance correlates well with the readiness test, the predictive validity of the readiness test is supported.

The second type of criterion-related validity, concurrent validity, is a test's ability to correlate with some current criterion. Here, the measure in question is correlated with a specific criterion measure, to the same group of individuals and at the same time (Salvia & Ysseldyke, 1991). For example, new reading achievement tests could be correlated with the students' grades in reading. If the correlation was high, the concurrent validity of the reading achievement test would be supported.

Measuring Construct Validity

Construct validity is a test's ability to measure a particular construct or trait underlying the test (Salvia & Ysseldyke, 1991). For example, many tests attempt to measure the constructs of intelligence and visual perception; however, constructs cannot be measured directly and must be inferred from observed behaviors (McLoughlin & Lewis, 1990). The definition of the construct and the theory from which the construct is

originated allows certain predictions to be confirmed or disconfirmed (Salvia & Ysseldyke, 1991).

Gronlund (1985) introduced a three-step procedure for demonstrating construct validity. First, several constructs that are presumed to account for test performance are identified. Second, hypotheses are generated that are based on the identified constructs. Third, the hypotheses are verified using logical or empirical methods.

Research on the Validity of Authentic Assessment in Written Language

Validity, which basically means that a test measures what it claims to measure, is regarded as the single most important aspect of a test (Anastasi, 1988; Mehrens & Lehmann, 1987). In reviewing the research on authentic assessment in written language, there appear to be no studies which have resulted in validity coefficients.

Modern views of assessment are based on theories of learning that are unlike those assumed by conventional test theory. It is not surprising that they have generated some disputes about measurement issues. At the very least, the new approaches require expansion of concepts that are already in transformation in the measurement community. Thus, the unified and extended notions of validity developed by Cronbach (1988), Cole and Moss (1989), and Messick (1989) are challenged to yet further growth as comprehension of learning encompasses additional aspects of cognition and social interaction, thereby extending the construct or constructs to be accounted for. More recent views of assessment call for understandings such as those considered in Frederiksen and Collins' (1989) notion of systematic validity, which centered attention on

the effect of assessment on the whole of the educational system it is designed to serve---clearly an extension of the earlier emphasis on social consequences as a consideration for validity.

In fact, as Linn et al. (1991) suggested, "modern views of validity already provide the theoretical rationale for expanding the range of criteria" to be applied to new as well as familiar approaches to assessment, even though "in practice, . . . validity is usually viewed too narrowly and given short shrift," whereas "reliability has too often been overemphasized at the expense of validity" (p. 23). A number of writers have identified criteria and concerns to be contemplated in evaluating the merit of new forms of assessment such as authentic assessment practices (e.g., Frederiksen & Collins, 1989; Linn et al., 1991; Snow, 1989; Wiggins, 1989). Other writers have identified criteria and concerns specific to written language assessment (e.g., Keech-Lucas, 1988; Valencia, McGinley, & Pearson, 1990). Although none of the sets of criteria or issues is thought to be complete or conclusive, they do suggest a common effort to expand the criteria for judging the value of an assessment beyond narrow concerns with efficiency, reliability, and validity (Camp, 1993).

In the years that multiple-choice tests and writing samples have constituted the dominant approaches to writing assessment, teachers, researchers, and writing assessment practitioners have worked to improve the measurement properties of the writing sample. A considerable body of research has been gathered on writing assessment, as evidenced by the sizable and rapidly expanding number of publications on the subject (Breland, Camp, Jones, Morris, & Rock, 1987; Cooper, 1981; Greenberg, Weiner, & Donovan,

1986; Keech-Lucas, 1988; Lucas & Carlson, 1989; Ruth & Murphy, 1988). As a result of this research and repeated refinements in practice, procedures have been developed for analyzing, revising, pretesting, and further revising the prompts used to elicit writing (Ruth & Murphy, 1988).

Because of the need for authentic assessment of students' writing skills (Archibald & Newmann, 1988), many of the authentic assessment practices employed have centered around writing or written language (McKendy, 1992; Swartz, Patience, & Whitney, 1985). Baker (1991) found that "the most useful studies" on authentic assessment in the ERIC educational database are "those conducted in the writing assessment area" (p. 3).

Much emphasis has been placed on the reliability of holistic scoring procedures of authentic assessment practices in written language. The constant significance placed on reliability has caused the educational profession to assume, confuse, and otherwise ignore the validity of holistic scoring procedures of authentic assessment (Huot, 1990). A good illustration of this is found in C. R. Cooper's chapter, "Holistic Evaluation of Writing" (1977), where Cooper indicated that "holistic evaluation can be as reliable as multiple-choice testing . . . is always more valid" (p. 15). Cooper offers no empirical evidence or theoretical support, however, for this claim concerning the validity of holistic scoring procedures used in authentic assessment.

In The Evaluation of Composition Instruction, Davis, Scriven & Thomas (1981) noted that reliability is as critical as validity; in fact, it is a prerequisite for validity. On the other hand, test reliability is a necessary, but not sufficient condition for test validity. Furthermore, a test's ability to be consistent (reliable) means nothing unless the

assessment instrument measures what it claims to measure (validity) (Popham, 1981). Validity is the single most important aspect of a test (Anastasi, 1988; Mehrens & Lehmann, 1987). These contrasting accounts of validity and reliability are a clear depiction of the exaggerated position of reliability and the subsequent neglect of validity in literature on writing evaluation.

Many of the authentic assessment tasks employed have centered around writing or written language. These writing samples have become the routine method of testing and placement at many colleges and universities (Swartz et al., 1985; McKendy, 1992). The evaluation of writing samples is subjective and relies heavily on teachers' judgement.

Importance of Teachers' Judgement

The accurate evaluation of students' progress through school is basic to responsible education. Without reliable and valid assessment of students' performance in areas such as reading and written expression, the provision of appropriate teaching and learning procedures becomes haphazard. Both high- and low-achieving students are ignored because teaching is focused on the large middle-of-the-range group of students that teachers presume makes up the majority of their classes (Sharpley & Edgar, 1986).

Perhaps because of the large amount of time spent with students and the very important role that they play in the development of students' academic skills, teachers are expected to accurately assess the achievement levels of their students. In fact, teachers are very often required, by parents, school administrations, and other teachers, to make such assessments. Day-to-day decisions are made by teachers regarding the selection of

materials, teaching strategies, curriculum content and objectives, formation of teaching groups, and the placement of children in accelerated or remedial classes. These decisions must be made accurately if students' progress is to be ensured (Sharpley & Edgar, 1986).

Student progress in writing was long considered a problematic area for those who teach written language and conduct educational research. For years, researchers struggled with the development of methods that enabled educators to provide reliable and valid means of directly assessing students' writing ability (Huot, 1990). The conflict in scoring rests in the value placed on human judgement. In authentic assessment, teachers' judgement is highly respected. Wiggins (1989) pointed out that, "In the contest of testing, equity requires us to insure that human judgement is not overrun or made obsolete by an efficient, mechanical scoring system" (p. 708). Consequently, performance standards and rubrics are produced through a consensual process. Usually, this process is local and is expected to (a) give prominence to the judgements of those who will be affected by the assessment; (b) serve as a learning experience by improving knowledge of valued performance and how to assess it; and (c) acquire a sense of ownership in the process. Differences in judgements, or between judgements and the results of standardized measures, are socially moderated, consideration is often given to the perceptions of students' classroom teacher, who knows the examinee the very best (Department of Education and Science, 1987; Wiggins, 1989). Judgement-based assessment plays an important role in the assessment of writing (Neisworth & Bagnato, 1988).

Studies in which teachers' judgements have been employed (e.g., Kellaghan, Madaus, & Airasian, 1982; Pedulla, Airasian, & Madaus, 1980) have revealed substantial correlations between teachers' judgements and scores on standardized tests. Since the 1920s, literally dozens of researchers have reported correlations in the order of .5 to .6 between teachers' judgements and various standardized tests (Cronbach, 1961).

Researchers have examined the correlation between students' performance on standardized tests of academic achievement and teachers' a priori judgements of their performance. Gerber and Semmel (1984) presented an argument for using classroom teachers as useful and defendable tests of academic achievement.

Educators use ratings of developmental abilities for making important educational and instructional decisions (Clark & Peterson, 1986; Peterson, 1988; Shavelson & Stern, 1981). Teachers use ratings to determine groupings of students for instructional purposes, to determine whether students are comprehending the lesson, and in making decisions regarding whether instructional groups or strategies should be changed. Teacher judgements are used daily in regular and special education classrooms (Hoge, 1983).

Teacher ratings are used constantly in instructional settings (Gerber & Semmel, 1984). In the academic area of written language, teachers have the capability of producing an accurate evaluation of student progress. A review of studies shows a direct relationship between teachers' judgements and academic achievement. The results of the studies suggest that teachers' judgements share common dimensions with standardized achievement test data. For example, the correlations between teachers' judgements and

language arts and English are as follows: (a) .61 (DuPaul, Rapport, & Perriello, 1991); (b) .65 (Egan & Archer, 1985); (c) .74 (Hopkins, George, & Williams, 1985); (d) .76 (Wright & Wiese, 1988) and (e) .89 (Hammill & Hresko, 1994) (see Appendix A). The correlations are all positive, and fairly substantial, which validates the use of teachers' judgements of achievement. These research studies provide evidence of the validity of teacher judgements in evaluating students' educational achievement in written language.

Teachers' judgements and evaluations are essential in determining a student's eligibility for special education services. Teachers' judgements, by way of teacher rating scales, are often preferred by teachers (Ysseldyke, Algozzine, & Richey, 1982).

Authentic assessment practices clearly rely on teachers' judgements (Gresham, Reschly, & Carey, 1987). Teachers' judgements make it possible to augment, enhance, and corroborate data obtained from educational assessment measures in a sound and valid manner (Gerber & Semmel, 1984; Simeonsson, Huntington, Short, & Ware, 1982). Teachers' judgements provide a comprehensive and quick way to evaluate academic curriculum content such as individual students' writing samples (DuPaul et al., 1991). Writing, which is evaluated using teachers' judgements, is incorporated into many authentic assessment tasks. Teachers' judgements provide a comprehensive and quick way to evaluate individual students' writing samples. Just as written language relies heavily on teachers' judgements for assessment purposes, so do authentic assessment practices (Gresham et al., 1987).

Importance of Investigating the Validity of Authentic Assessment for Students With Learning Disabilities

The history of writing assessment has, in recent decades, frequently involved controversy over formats and methodology (Camp, 1993). A majority of the research in this area has involved typical learners; however, a small but growing body of research in special education has provided educators with information about the written language abilities of students with LD (Graves, Montague, & Wong, 1990; Isaacson & Mattoon, 1990; Laughton & Morris, 1989; Tindell & Parker, 1989; Vallecorsa & Garriss, 1990). The importance of investigating the validity of authentic assessment in written language with students with LD is discussed within the context of (a) writing problems, (b) referral procedures, and (c) movement toward integration of students with LD.

Writing Problems of Students With LD

Dramatic changes in the emphasis on writing in school curricula within the past decade have given rise to substantial research investigating children's ability to compose. Although most research examining children's writing skills has involved typical learners, an increasing number of studies have been conducted, primarily in the last decade, which address the composition skills of students with LD (Newcomer & Barenbaum, 1991).

Most students who receive services for LD have severe writing difficulties that persist over time (Graham & Harris, 1989). Research has revealed evidence of depressed performance on the part of the students with LD that did not diminish with age or years in school (Anderson, 1982; Poplin, Gray, Larsen, Banikowski, & Mehring, 1980). Warner, Alley, Deshler, and Schumaker (1980) reported that of the students with LD in their

study, the majority scored at or below the 10th percentile in written language. This low academic performance differentiates students with LD from other students.

Students with LD have difficulty composing stories. Moreover, when the stories of students' with LD were analyzed for various mechanical, vocabulary, and syntactic/fluency components, considerable challenges and frustrations were encountered on the part of the students (Barenbaum, Newcomer, & Nodine, 1987; MacArthur & Graham, 1987; Nodine, Barenbaum, & Newcomer, 1985; Tindal & Parker, 1989).

Students with LD have difficulties with the mechanics of written language. Mechanics refers to capitalization, punctuation, and spelling. Punctuation has been identified as a specific problem. Poplin et al. (1980) reported comprehensive and persistent deficits in students' performance in all areas of the Test of Written Language (Hammill & Larsen, 1978), including spelling and style (capitalization and punctuation). When comparing students with LD with typical learners across a variety of educational levels, fourth grade through college, researchers have consistently found significantly more mechanical errors in the work of learning disabled students, particularly in spelling (Houck & Billingsley, 1989; Moran, 1981; Vogel & Moran, 1982).

Tindal and Parker (1989) demonstrated a significant predictive relationship between both the percentage of words correctly spelled and the percentage of words correctly sequenced and the overall holistic rating of stories written by learning disabled students. Regardless of the purposes for these significant positive correlations, the implications for students with LD, who have extreme difficulty with spelling and sequencing, are evident. Researchers have established a high correlation between an

assortment of mechanical writing skills, including words spelled correctly and use of novel words, and standardized writing test scores. Research has also revealed that thought-units or T-units, a frequent measure of syntax, are not predictive of universal performance in writing (Deno, Marston, & Mirkin, 1982).

The next category of research focuses more directly on components of stories produced by students with LD. Most researchers have examined only one mode of story composition, handwritten essays. As a group, students with LD write fewer stories than their regular-education peers. For example, students with LD generally fail to meet the most basic criteria for story generation, by failing to include a conflict and a resolution of that conflict. Students' fluency, the total number of words per composition, correlates with story production, and students with LD are less fluent in all composition types (Gajar, 1989; Houck & Billingsley, 1989; Morris & Crump, 1982).

Moving away from a central concern with syntax, mechanics, and fluency, the research examining knowledge of story schema reveals that while some school-age children with LD have gained a knowledge of the basic components of a story and can utilize the knowledge to write stories, the majority have difficulty with some important aspect of the task. Data frequently show subjects with LD performing below their normally achieving peers on some creative or organizational dimension. The compositions written by most students with LD lack certain critical components of stories (e.g., setting, characters, conflict, and resolution) and are often classified as less sophisticated compositions (Englert & Thomas, 1987; Englert et al., 1988; Newcomer & Barenbaum, 1991).

Another persistent problem in the writing abilities of students with LD is related to some aspect of cohesion. Compositions by writers with LD often contain extraneous ideas, confusing words, or unclear referents that interrupt the meaningful flow of the stories and make them difficult to read (Gregg, 1983; Gregg & Hoy, 1989; Nodine et al., 1985). However, Laughton and Morris (1989) found that, unlike younger subjects with LD, sixth-grade students with LD paralleled their regular education peers in story production, which suggests improvement with maturation and experience.

Other researchers have measured students' story-composing ability when more than one mode of production was utilized. When comparing dictated compositions with compositions written by hand, a key variable to consider is the effect of mechanical skills on story production. Researchers have confirmed that students with LD know considerably more about story production than they are able to convey on paper because they are constrained by mechanical problems. Furthermore, students with LD organize their dictated stories according to acceptable story grammar structure (Ripich & Griffith, 1988; Roth & Speakman, 1986; Stein & Glenn, 1979).

Students with LD have difficulty with most facets of mechanics, syntax, and fluency and are less masterful than other students in writing stories. Students with LD also have difficulty generalizing writing skills. Schmidt, Deshler, Shumaker, and Alley (1988) provided the most important conclusion regarding generalizability--the involvement of each student in taking control of the process of writing is essential for improvement in substantive skills. Authentic assessment practices allow students to have greater control over their learning (Valencia, 1990).

Referral Procedures of Students With LD

When teachers refer students for special education, they set in motion a series of decision-making activities that can significantly influence whether students receive special services. Referral and identification procedures that typically are used to determine student eligibility often have a number of problems, including negligible relationships between assessment data and eligibility decisions and questionable decision making during the placement process (Ysseldyke et al., 1982). There is little evidence that any assessment intended to diagnose the disability of a student provides data relevant to the educational intervention needed by the student (Algozzine, Sacca, & Maheady, 1986; Galagan, 1985; Salvia & Ysseldyke, 1991).

While these and other concerns have been raised about the referral and identification process in general, additional problems more specific to students with LD have evolved. For instance, the use of discrepancy formulas to determine the degree of discrepancy between achievement and intelligence in one or more academic areas has been reproached as being technically unsound (Council for Learning Disabilities, 1986). In addition, the typical standardized test is timed and uses a multiple-choice format, which requires the ability to recognize and select the best answer. Students with LD often exhibit difficulty performing under time pressures and have problems in tracking and bubbling in their answer on test forms. Fuchs and Fuchs (1990) noted that multiple-choice responses reveal little about students' strategies in solving problems. Shepard (1989) added that multiple-choice formats lead to endless drill on decontextualized skills.

Though test authors are moving toward the inclusion of special populations in their standardization procedures, the validity and other statistical measures of traditional, norm-referenced tests are primarily standardized on groups of normal children. The response limitations of many children with disabilities seriously interfere with their performance on such measures and make it difficult to obtain accurate estimates of their abilities and to acquire precise measures of test reliability and validity (Keogh & Sheehan, 1981). Assessment specialists must find ways to circumvent these difficulties in order to secure measures of functional status. While a number of options are available, a valid one is to include norm-referenced, criterion-referenced, and judgement-based assessment measures from multiple sources across multiple settings (Bagnato, Neisworth, & Munson, 1989).

These issues that characterize the status of students with LD have been spotlighted in the past 15 years. Difficulties associated with the referral and identification procedures for students with LD have prompted a search for procedures that also might result in the acquisition of information that is useful to teachers and child study teams in planning intervention strategies. Many specialists concerned with the education of students with disabilities believe that teachers should become more centrally involved in the assessment process. Previous investigations have found that teachers are able to identify LD early in a child's school career, particularly when rating scales and checklists based on teachers' judgements are used (Mercer, Algozzine & Trifiletti, 1979). The use of teachers' judgements as an integral part of the identification process is not a new practice. Myklebust (1973) and other researchers developed teacher ratings which utilize teachers'

judgements to assist in identifying students with LD. While teachers' judgements generally are considered useful during the referral process, their measures were informally constructed and lacked suitable psychometric properties. For these and other reasons, the majority of teacher judgement measures have fallen into disuse. Recently, however, interest has been rekindled due to the presence of sounder measures and a growing recognition of the need for better information from educators during the referral process (Oakland, Shermis, & Coleman, 1990).

Educators have the expertise to produce an accurate evaluation of student progress, especially in the area of written language. Studies show a direct relationship between teachers' judgements and students' academic achievement. The results of these studies suggest that teachers' judgements share a commonality with standardized assessments (DuPaul et al., 1991; Hammill & Hresko, 1994; Wright & Wiese, 1988). Additionally, because authentic assessment practices rely heavily on teachers' judgements (Gresham et al., 1987), teachers' judgements have the ability to augment, enhance, and corroborate data obtained from educational assessment measures in a sound and valid manner (Gerber & Semmel, 1984; Simeonsson et al., 1982).

Movement Toward the Integration of Students With LD

With the drive toward restructuring of the educational system, school reform emphasizes the special learning needs of individual students. Reconstruction of the entire education system is seen by many as the solution to preparing children with disabilities for the next century (Audette & Algozzine, 1992).

The terms integration, mainstreaming, supported education, inclusion, and least-restrictive environment seep into almost any conversation among educators today. This language does not refer just to students with severe disabilities, but to any special education student with a label (Biklen, Ferguson, & Ford, 1989).

Special education has been touched by the reform efforts. The first responsibility of education reformers, that of sounding the alarm, has occurred. Problems related to assessment, decision making, and intervention are among those criticized by special education's reformers (Ysseldyke, Algozzine, & Thurlow, 1992). The second task of reform, that of proposing solutions, is in development. Efforts to reintegrate students with disabilities, to develop partnerships with general education in meeting the needs of students with disabilities, and to challenge and debate the maintenance of current practices have dominated the interests of special educators for the past few years (Audette & Algozzine, 1992).

Because the American education system is under excessive scrutiny and pressure from essentially every educational, political, business, or special interest group in society for its alleged faults, the general public is demanding more and stronger evidence that such reforms are working to produce students who can think, communicate, and solve problems. The pressure has centered attention on an alternative approach to educational assessment, authentic assessment (National Commission on Excellence in Education, 1983; National Governors' Association, 1986). Moving toward an inclusive environment and developing partnerships with general education to meet the needs of students with special needs requires the inclusion of students with disabilities in authentic assessment

practices. Because authentic assessment is being applied to special populations, establishing quantitative measures (i.e., types of validity) and determining whether these measures vary between individuals with and without LD is necessary.

Conclusion

School reform, which is a dominant force in contemporary education, is an effort to ensure that all students in the United States receive a free and appropriate education. Efforts to reintegrate students with disabilities and assess their progress in a more realistic, naturalistic, and authentic manner, are resulting in the transformation of local, state, and national educational policies.

The ability of teachers to make accurate educational decisions is essential if students' progress is to be ensured. Research studies have substantiated the validity of teachers' judgements in the evaluation of academic achievement and have supported teachers' judgements as effective and defendable tests of educational achievement in written language. In addition, teachers' judgements can be used to augment, enhance, and even corroborate the strengths of students with LD in writing rather than focusing on their persistent writing difficulties. Because many authentic assessments center around writing, the use of teachers' judgements in the evaluation of students' progress is a valid and essential means of educational assessment.

Research Questions

As evidenced in this review of literature, a connection between the validity of teachers' judgements and authentic assessment practices in written language has been supported. The following research questions were generated to direct this investigation:

1. Is there a correlation between teachers' judgements of students' authentic assessments and scores on standardized assessments of written language?

2. Do measures of criterion-related concurrent validity differ between groups of students with and without LD?

3. Do measures of construct validity support the use of authentic assessment in written language? Because authentic assessment encompasses a basic school subject, students who do well in writing should do well in other areas of school. If this is true, student performance on writing samples should correlate with their performance in other academic areas.

CHAPTER 3

METHODOLOGY

This study was conducted to investigate the selected validity of authentic assessment in written language. Additionally, this study was designed to determine whether these types of validity vary between individuals with and without learning disabilities (LD). The following section describes the methodology used in this study. Organization for this section is as follows: (a) subject selection, (b) setting, (c) instrumentation, (d) research design, (e) data collection, and (f) data analysis.

Subject Selection

Permission to conduct this study was obtained from the director of special education and the administrators of the research and evaluation departments in a large, urban school district. The nature of the study was also reviewed and approved by the Human Subjects Institutional Review Board at the University of North Texas.

General educators and students with and without LD in the fourth- and fifth-grades were utilized in this study. Thirty-eight students with LD and 46 students without LD were included. In addition, 10 teachers from fourth- and fifth-grade regular education classes were included for the purpose of scoring authentic narrative writing assessments.

Setting

A large, local urban school district, located in North Central Texas, was included for this study. The district had 21 elementary schools, 7 middle schools, and 3 high schools, yielding a total enrollment of 24,822 students. In addition, the district-wide attendance percentages for 1992-1993 were 95%, with less than a 4% drop-out rate. The student ethnic composition at the time of the study was White (86%), Hispanic (7%), Black (4%), Asian/Pacific Islander (2%), and American Indian/Alaskan Native (.5%). The school district implemented authentic assessment practices and had teachers who were trained in scoring authentic assessments of writing according to the Texas Assessment of Academic Skills.

Instrumentation

The accurate evaluation of students' progress through school is basic to responsible education. Judgement-based assessment plays an important role in the assessment of writing (Neisworth & Bagnato, 1988). A review of studies shows a direct relationship between teachers' judgements and students' academic achievement, and provides evidence of the validity of teachers' judgements in evaluating students' educational achievement in written language (DuPaul, Rapport, & Perriello, 1991; Hammill & Hresko, 1994; Wright & Wiese, 1988).

In this research study teachers' judgements were employed in evaluating authentic assessment practices in written language. A holistic method was selected as a direct assessment of student writing skills because holistic scoring procedures have proven to be

a valid, reliable, and efficient method of rating students' writing samples (Elliott, Plata, & Zelhart, 1990). The teachers holistically scored student's authentic narrative writing samples using the Texas Achievement of Academic Skills (TAAS) focused holistic scoring process. This scoring process is holistic because the students' authentic writing samples are appraised as a whole. The scoring process is focused in that the individuals' writing is evaluated according to preestablished criteria. These criteria are as follows:

Objective 1: The student will respond appropriately in a written composition to the purpose or audience specified in a given topic.

Objective 2: The student will organize ideas in a written composition on a given topic.

Objective 3: The student will demonstrate control of the English language in a written composition on a given topic.

Objective 4: The student will generate a written composition that develops, supports, or elaborates the central idea stated in a given topic.

Each TAAS response was measured according to the extent to which it reflected mastery of these objectives. Individual responses were scored on a scale of 1 (low) to 4 (high). Students were given a rating of 0 if the response could not be scored.

The Test of Written Language-Revised (TOWL-2) (Hammill & Larsen, 1988) is a standardized, norm-referenced test. In order to obtain an estimate of students' functional writing ability, the Spontaneous Writing portion of the test was used in this study. The Contrived Writing portion of the TOWL-2, which focuses on the isolated

evaluation of the smallest units of written discourse, such as spelling, capitalization, punctuation, and word usage, was not used in this study.

Students wrote a narrative which was inspired by a picture prompt of prehistoric beasts and cavemen. This type of assessment focused on evaluating components of writing in terms of their relationship to an actual excerpt generated by the student. A student may be able to score well on tests of vocabulary, word usage, handwriting, spelling, capitalization, punctuation, and syntax that have contrived formats and still be unable to create a central idea that adequately communicates feelings, thoughts, and opinions. In other words, the expertise to write meaningfully in everyday life or school situations requires an integrated grasp of the components rather than mere competence in the components when they are measured in isolation.

The concept of reliability refers to the consistency with which any measuring instrument estimates various attributes of something. Sattler (1988) observed that for tests such as the TOWL-2 to be considered minimally reliable, their reliability must approximate or exceed .80 in magnitude. Coefficients of .90 or above are considered the most desirable (Salvia & Ysseldyke, 1991). Data related to four types of reliability were reported in the TOWL-2. The median reliability coefficients were interscorer (.96), internal consistency (.94), form equivalence (.78), and stability (.84). The TOWL-2 reports coefficients that are high enough to be accepted as evidence of the TOWL-2's reliability.

Most authors of current textbooks dealing with educational and psychological measurement (e.g., Anastasi, 1988; Gronlund, 1985; Salvia & Ysseldyke, 1991) suggest

that those who develop tests should provide evidence of at least three types of validity: content validity, criterion-related validity, and construct validity. The TOWL-2 addresses each of these types of validity.

Content validity involves the systematic examination of content to determine whether a representative sample depicts the domain to be measured (Anastasi, 1988). The determination of content validity is a matter of judgement and is closely tied to the procedures used to construct the assessment tool. Obviously, this type of validity has to be incorporated into the test at the time of item selection. By determining the rationale underlying the selection of the testing formats and items and of the statistical procedures used to choose good items, a test has relevance for content validity as well as for item and format selection. For these reasons, evidence for content validity is supported in the TOWL-2 (Hammill & Larsen, 1988).

Criterion-related validity concerns the relationship of the measurement criteria to some outside criterion (McLoughlin & Lewis, 1990). A test such as the TOWL-2, which is presumed to measure writing ability, should correlate well with other tests that are also known to measure writing. When the standard scores between the TOWL-2 and SRA Achievement Series were compared, the coefficients ranged from .30 to .70. Second, when teachers' judgements of students' stories and results of the TOWL-2 were compared, the coefficients ranged from .33 to .61. The coefficients reported are large enough to lend support to the contention that the TOWL-2 has criterion-related validity.

Construct validity is the ability to measure a particular construct or trait underlying a test (Salvia & Ysseldyke, 1991). Eight basic constructs thought to underlie

the TOWL-2 were delineated. Through measures of (a) age differentiation, (b) group differentiation, (c) grade differentiation, (d) interrelationships among test items, and the (e) relationship of the TOWL-2 to tests of achievement and intelligence, substantial construct validity was supported.

The Written Language Assessment (WLA) (Grill & Kirwin, 1988) was also used in this research study. The WLA is a standardized, norm-referenced test of writing ability that yields valid and reliable scores for general writing ability, productivity, word complexity, and readability. Each of the scores can be converted to a scaled score and summed to arrive at a written language quotient.

Narrative writing samples were used in order to investigate selected types of validity. Because the creative writing portion of the WLA is a narrative writing prompt of a girl and her cat, it was the only section administered. The score received on the creative writing portion was multiplied by three in order to calculate the quotients. As with the TOWL-2, the WLA provides a picture prompt to encourage students' writing. The WLA is a direct assessment, product evaluation instrument. The WLA is not a test with contrived tasks that isolate subskills of writing from composition, but rather a means of evaluating writing based exclusively on students' actual compositions.

Data about two types of reliability, internal consistency and inter-rater, were reported in the WLA. Internal consistency reliability is a measure of how well a test or subtest measures one skill from beginning to end of the test or subtest. The written language quotient has a median internal consistency reliability coefficient ranging from .90 to .92, depending on whether three or four scores are being compared. Inter-rater

reliability measures the extent to which separate raters agree on the ratings of the same piece of writing. Three studies were conducted which yielded inter-rater reliability coefficients of .81, .75, and .75, respectively. The WLA yields coefficients that are acceptable evidence of the test's reliability.

In order to investigate the criterion-related validity of the WLA, students were given the Picture Story Language Test (PSLT) (Mykelbust, 1965). The two standardized tests are similar because both use direct product evaluation of students' writing. All correlations were significant beyond the .01 level of confidence except for PSLT syntax quotient correlations with WLA scores.

In determining construct validity, two theoretical assumptions were constructed for the WLA which were thought to underlie the test. First, the authors assumed that the four WLA scores represent measures of different aspects of writing. Second, they speculated that young students' writing performance is better among older students than among younger students. To investigate the first assumption, intercorrelations of WLA raw scores yielded median correlations ranging from .22 to .86. To investigate the second assumption, that older students' writing performance is better than that of younger students as measured by WLA scores, a multiple correlation with chronological age and the four WLA raw scores yielded a multiple correlation coefficient of .54. This result indicates a moderate relationship between WLA scores and students' age.

Research Design

After permission to conduct this study was obtained, parent/participant consent forms were sent to the parents of students with and without LD. Eighty-four consent forms were returned for participation in the research study.

A narrative writing sample was randomly selected and removed from each individual student's work folder. The students' authentic writing samples were divided evenly so that each anonymous teacher received the same number of writing samples from regular education students as from students with LD. In addition, an anchor writing sample was used to measure the consistency of teachers' judgements. This anchor writing sample was a narrative written during the 1993 TAAS test and scored by the Texas Education Agency. Furthermore, each teacher received the anchor writing sample as the second writing sample in their collection of writing samples.

During the same time period, additional information, such as report card grades and supplementary standardized assessment scores, was collected from the students' cumulative folders, special education records, and teacher input. Also, students were administered the creative writing section of Grill and Kirwin's (1988) Written Language Assessment (WLA) and the spontaneous writing portion of Hammill and Larsen's (1988) Test of Written Language (TOWL-2). After the standardized test data were accumulated, these scores were correlated with those of the students' authentic writing samples.

First, correlating two standardized tests of the same domain, one should expect a high correlation; however, because authentic assessment and standardized tests are

similar, but of different paradigms, a moderate correlation between the two was anticipated.

Second, the TOWL-2 and WLA already have established criterion-related concurrent validity; therefore, a substantial correlation between standardized achievement in written language and authentic assessment of written language was expected. Investigation was also undertaken to determine whether a significant difference occurs between the criterion-related concurrent validity coefficients of students with and students without LD.

In investigating the construct validity of authentic assessment, a construct which was thought to underlie authentic assessment was developed. This construct was analyzed through the use of standardized tests (e.g., TOWL-2, WLA), authentic writing assessments, and report card grades. The construct addressed was "Because authentic assessment encompasses a basic school subject, students who do well in writing should do well in other areas of school. If true, student performance on student's writings should correlate with other academic areas."

Data Collection

The data for this research study were collected during the spring of 1994 at a large, urban school district. The first step in developing a random sample of subjects who were in regular education was to select students who were in fourth or fifth grade, who had never been retained, and who were receiving no special education services. This procedure resulted in a pool of 106 regular education fourth- and fifth-grade students from 7 classrooms in two elementary schools in the North Central Texas Area. The

second step in the development of the subject sample was to send a parent/participant consent form home with each student. Of the 106 students, 49 returned the consent forms. In the final step, students were eliminated who did not meet all of the selection criteria. This yielded a final subset of 47 students who met all of the selection criteria.

In order to gather a large enough sample of students with LD, participant consent forms were distributed to 57 fourth- and fifth-grade students with LD from 16 classrooms in 2 additional elementary schools in the North Texas Area. For the purposes of this study, students who were identified by the school district as having a LD in the language arts areas were eligible to participate in the research. Additional information, such as supplementary standardized assessment scores and report card grades, were obtained from school cumulative records, special education records, and classroom teacher input.

The first step in developing a sample of subjects with LD was to select fourth- and fifth-grade students who were identified as having an academic deficity in the language arts areas based on Texas state guidelines. This procedure resulted in a pool of 48 subjects. The second step in the development of the sample of students with LD was to send a parent/participant consent form home with each student. Of the 48 students, 44 students with LD returned the consent forms. In the final step, students were eliminated who did not have a learning disability in language arts; therefore, if students qualified as having a learning disability in mathematics only, they were eliminated from the study. The final subset resulted in 42 students who met all of the selection criteria.

Eighty-nine students met the criteria for participation in the research study. Five students were eliminated from the sample because the students moved before the end of

the study, yielding a total sample size of 46 students without LD and 38 students with LD.

Data Analysis

In order to support criterion-related concurrent validity and construct validity as each pertains to authentic assessment, the administering of several standardized tests, collecting authentic of writing assessments, and accumulating of report card grades in language arts was required. The standardized written language assessments were scored according to the guidelines outlined in the testing manuals.

Much emphasis has been placed on the reliability of holistic scoring procedures of authentic assessment practices in written language (Huot, 1990). The concept of reliability refers to the consistency with which any measuring instrument estimates various attributes of something. In supporting the inter-rater reliability of teacher judgements of authentic writing samples, an anchor writing sample was given as the second writing sample to each teacher. Percentages were calculated to support the inter-rater reliability.

The first research question, "Is there a correlation between teachers' judgements of students' authentic samples and scores on standardized assessments in written language?" was analyzed by correlating the holistic scores of individual students' authentic writing samples and the standardized written language results collected from the Test of Written Language-Revised and Written Language Assessment. Because the authentic assessment data are ordinal and the standardized assessments are interval in nature, the biserial

correlation coefficient was calculated to determine the relationship between the (a) TOWL-2 and teachers' judgements of authentic writing samples and (b) WLA and teachers' judgements of authentic writing samples. Correlating two standardized tests of the same domain, one should expect a high correlation. However, because authentic assessment and standardized tests are similar, but of different paradigms, a moderate correlation between the two was anticipated.

The second research question "Do measures of criterion-related concurrent validity differ between groups of students with and without LD?" was measured by analyzing the validity coefficients of students with and without LD.

Standardized written language assessments such as the TOWL-2 and the WLA, which measure writing, correlate well with other tests that are also known to measure writing. This relationship validates evidence of criterion-related concurrent validity. The TOWL-2 and WLA already have established criterion-related concurrent validity; therefore, a substantial correlation between standardized achievement in written language and authentic assessment of written language was expected. Data from the TOWL-2, the WLA, and teachers' judgements of authentic writing samples were correlated between students with and without LD. The biserial correlation coefficient was used to determine the relationship between students with and without LD and (a) the TOWL-2 and teachers' judgements of students' authentic writing samples and (b) the WLA and teachers' judgements of students' authentic writing samples.

To determine whether the test statistics differed between students with and without LD on the standardized and authentic written language assessments, various

calculations were used. Fisher's z transformations, estimated standard error of the difference between independent transformed correlation coefficients, and test statistics were calculated.

The last research question is "Do measures of construct validity support the use of authentic assessment in written language?" In investigating the construct validity of authentic assessment, a construct which was thought to underlie authentic assessment was analyzed. This construct was analyzed through the use of standardized tests (e.g., TOWL-2, WLA), authentic writing samples, and report card grades. The construct addressed was: "Because authentic assessment encompasses a basic school subject, students who do well in writing should do well in other areas of school. If true, student performance on student's writings should correlate with other academic areas."

The assumption thought to underlie the construct validity of authentic assessment of written language used data from teachers' judgements of authentic writing samples and individual students' grades in language arts. The data were calculated using Spearman's rho correlation coefficient because both variables being correlated were ranks.

CHAPTER 4

RESULTS

The purpose of the present study was to investigate the selected validity of authentic assessment measures in written language. In addition, the study sought to determine if the various types of validity differ between individuals with and without learning disabilities (LD). Specific standardized written language assessments, students' authentic writing samples and academic grades from the classroom were used for the purpose of analyzing various types of validity.

Student Demographics

Information related to gender, age, and ethnicity was available for all 84 students, 38 students with LD and 46 without LD, in the research sample (see Appendix B). The mean age of students participating in the study was 11 years 2 months, with a standard deviation of 8.3 months and a range of 10 years 7 months to 12 years 8 months. An analysis of males to females showed the division to be almost even (i.e., M = 51%; F = 49%). Examining the ethnicity, showed that most of the students were identified as White (79%) or Black (14%). Hispanics (3%), Asian (1%) and other minorities (3%) made up a relatively small portion of the total sample size. Both gender and ethnicity data parallel the current 1990 census data (Table 1).

Table 1

Characteristics of Students With and Without Learning Disabilities Who Participated in the Study

Variable	N	Percent
	Gender	
Male	43	51
Female	41	49
	Ethnicity	
Caucasian	66	79
Afro-American	12	14
Hispanic	2	3
Asian	1	1
Other	3	3

Teacher Characteristics

Initially, 10 regular education content area elementary school teachers of fourth- and fifth-grade students were requested to participate in the study. The teachers were from one North Central Texas area school and were trained according to the TAAS focused holistic scoring rubric. All 10 of the teachers agreed to participate and did so until the end of the study.

Information related to gender, ethnicity, teaching background, educational background, training in written language and special education, experience teaching students with LD, and amount of time required to score holistically was available for all

10 teachers participating in the study (see Appendix C). All of the teachers in the research study were female. An analysis of ethnicity showed that the teachers were White (80%), Black (10%), or Hispanic (10%). The teachers had 112 years of experience among them, with a mean of 11.2 years of experience per teacher; however in teaching fourth- or fifth-grade students, the teachers had 48 years of experience, with a mean of 4.8 years of experience. All of the teachers had a bachelors' degree in Education. Only 20% held higher degrees or another certification area (i.e. speech pathology). Teachers averaged 5.10 semester hours of education and 31.6 in-service clock hours related to written language; whereas in the area of special education, the teachers averaged 6.30 semester hours of special education and 3.3 in-service clock hours. Approximately one-third of the teachers reported that no students with LD had been mainstreamed into their classrooms during the past 3 years. On the other hand, the remaining two-thirds averaged 10.4 students with LD who were mainstreamed into their classrooms during the past 3 years. Thus, a mean of 7.3 students with LD were included in regular classrooms for the 10 classroom teachers combined (Tables 2 and 3).

As part of the research study, teachers were asked to score narrative writing samples holistically and according to the TAAS focused holistic scoring rubric for which they had been trained. The teachers had been scoring according to this rubric for an average of 18.2 months. An analysis of the writing samples

Table 2

Characteristics of Regular Education Elementary Content Area Teachers Who Participated in the Study

Variable	N	Percent
	Gender	
Male	0	0
Female	10	100
	Ethnicity	
White	8	80
Black	1	10
Hispanic	1	10
	Education	
Certification in teaching (degree outside of teaching)	0	0
Bachelors' in education	10	100
Masters' in education	2	20

showed that most of the teachers were able to score one narrative writing sample in 2.90 minutes, or 2 minutes and 54 seconds. When analyzing the inter-rater reliability of these teachers, 80% (8) of the teachers scored the anchor sample of writing perfectly; the remaining 20% (2) of the teachers scored the sample as a 3 instead of a 2. Collectively, the teachers were consistent with one another when holistically scoring the fourth- and fifth-grade narrative writing samples.

Table 3

Characteristics of Regular Education Elementary Content Area Teachers Who Participated in the Study

Variable	Mean	SD	Range
Years of teaching experience	11.2	9.8	0-30
Years teaching 4th and 5th grade	4.8	4.4	0-15
Semester hours in written language	5.1	6.2	0-15
In-service clock hours within past 3 years in written language	31.6	29.3	0-100
Semester hours in special education	6.3	15.6	0-50
In-service clock hours within past 3 years in special education	3.3	9.4	0-30
Number of students with learning disabilities mainstreamed into own classroom within past 3 years	7.3	6.3	0-15
Number of months scoring holistically	18.2	9.6	6-30
Number of minutes to score 1 writing sample	2.9	1.9	0-5

Research Questions and Results

Three research questions were generated to guide this study. In this section each research question is addressed individually, as are the statistical procedures utilized, and the results found. Each research question is discussed. Where appropriate, all correlations were adjusted for restricted range and attenuation (Guilford & Fruchter, 1978).

Research Question 1

Is there a correlation between teachers' judgements of students' authentic samples and scores on standardized assessments in written language?

Data from two standardized testing instruments designed to measure written language were correlated with teachers' judgements of student authentic writing samples (see Appendices D & E). The data from the standardized assessments were quantitative and measured on an interval scale; whereas, the data from the authentic assessments were ordinal. Therefore, the biserial correlational coefficient was calculated to determine the relationship between the (a) TOWL-2 and teachers' judgements of student authentic writing samples and (b) WLA and teachers' judgements of student authentic writing samples.

Correlating standardized written language assessments of the same domain, one should expect a high correlation; however, because authentic assessment and standardized assessments of written language are similar but of different paradigms, a moderate correlation between the two was anticipated.

Analyzing the relationship between the spontaneous writing quotient of the TOWL-2 and individual students' authentic writing samples yielded a correlation of .45, significant at the .001 level. Correlation of the subtests of the TOWL-2 and teachers' judgements of student authentic writing samples produced correlations between .33 and .45, significant at the .001 level.

Examining the correlation between the written language quotient of the WLA yielded a correlation of .46, significant at the .001 level. Positive correlations of .27 to

.42 between the subtests of the WLA and teachers' judgements of students' authentic writing samples were also found. Using Anastasi's (1988) recommendations governing the interpretation of validity coefficients, only coefficients that were statistically significant ($p < .05$) and greater than .30 were considered substantial. These correlation coefficients represent a moderate relationship, and the correlations are statistically significant (Table 4).

Research Question 2

Do measures of criterion-related concurrent validity differ between groups of students with and without LD?

Criterion-related validity concerns the relationship of the measurement criteria to some outside criterion (McLoughlin & Lewis, 1990). Standardized written language assessments such as the TOWL-2 and WLA, which measure writing, correlate well with other assessment tools that are also known to measure writing. As previously reported, the TOWL-2 and WLA have established criterion-related concurrent validity; therefore, a substantial correlation between standardized achievement in written language and authentic assessment of written

language was expected. Additionally, determination of whether a significant difference occurs between the criterion-related concurrent validity coefficients of students with and without LD was analyzed.

Table 4

Correlation of Students' Authentic and Standardized Written Language Assessments

Variable	Correlation	Significance
Spontaneous writing quotient on the TOWL-2 and student authentic writing samples	.45**	.001
Subtests of TOWL-2 & student authentic writing samples		
Thematic maturity	.35**	.001
Contextual vocabulary	.44**	.001
Syntactic maturity	.33**	.002
Contextual spelling	.38**	.001
Contextual style	.45**	.001
Written language quotient on the WLA & student authentic writing samples	.46**	.001
Subtests of WLA & student authentic writing samples		
General writing ability	.42**	.001
Productivity	.34**	.002
Word complexity	.42**	.001
Readability	.27	.010

Note: *p < .05 **p < .01

Data from the TOWL-2, WLA, and teachers' judgements of students' authentic writing samples were correlated between students with and without LD. Because the variables on the standardized assessments were measured on interval scales and the underlying distributions of the variables were normal, and both variables on the authentic assessment measure or teachers' judgements of student authentic writing samples were

ordinal, the biserial correlation coefficient was utilized. The relationship between students with and without LD and (a) the TOWL-2 and teachers' judgements of students' authentic writing samples and (b) the WLA and teachers' judgements of student's authentic writing samples was investigated.

The size of the correlation is directly related to the variability of the sample. When the entire sample was utilized (e.g., 84 students with and without LD), variability was high; however, when the sample was divided between students with and without LD, the variability was reduced. In order to compensate for this restricted range, a correction for restriction of range was calculated (Guilford & Fruchter, 1978).

When two measured variables are correlated, the errors of measurement serve to lower the coefficient of correlation as compared with what it would be if the two measures were perfectly reliable. In order to counterbalance this imperfect measure, a correction for attenuation was calculated (Guilford & Fruchter, 1978).

The correlation between the spontaneous writing quotient of the TOWL-2 and teachers' judgements of writing samples of students with LD yielded a correlation of .39, significant at the .037 level. A positive correlation of .33, significant at the .047 level, was calculated between the written language quotient on the WLA and students with LD (Table 5).

Table 5

Correlation of Students With Learning Disabilities Authentic and Standardized Written Language Assessments

Variable	Correlation	Significance
Spontaneous writing quotient and authentic writing samples	.39*	.037
Subtests of TOWL-2 and authentic writing samples		
Thematic maturity	.43*	.029
Contextual vocabulary	.33	.091
Syntactic maturity	.36*	.040
Contextual spelling	.36*	.049
Contextual style	.38*	.066
Written language quotient and authentic writing samples	.33*	.047
Subtests of the WLA and student authentic writing samples		
General writing ability	.13	.222
Productivity	.23	.099
Word complexity	.24	.065
Readability	.19	.216

Note: *p < .05 **p < .01

The correlations of the scores of students without learning disabilities on standardized assessment tools and teachers' judgements of authentic writing samples yielded very poor correlations. From these results, it appears that there is no relationship

between the scores of students without learning disabilities on standardized tests and teachers' judgements of their writing samples (Table 6).

Table 6

Correlation of Students Without Learning Disabilities Authentic and Standardized Written Language Assessments

Variable	Correlation	Significance
Spontaneous writing quotient and authentic writing samples	.19	.164
Subtests of the TOWL-2 and authentic writing samples		
Thematic maturity	.24	.129
Contextual vocabulary	.14	.247
Syntactic maturity	-.10	.319
Contextual spelling	.07	.326
Contextual style	.39*	.025
Written language quotient and authentic writing samples	.27	.077
Subtests of the WLA and student authentic writing samples		
General writing ability	.29	.052
Productivity	.17	.193
Word complexity	.30	.065
Readability	.10	.343

Note: *$p < .05$ **$p < .01$

The majority of these correlations have little, if any, consistent relationship. Of the correlations that demonstrated a low positive correlation, some statistical significance was represented. Even though the correlations between standardized and authentic assessments in written language for students without LD were not significant, the correlations were compared to determine if the group characteristics differed between students with and without LD.

Various calculations were used to determine whether differences occurred between the test statistics of students with and without LD and standardized and authentic written language assessments. First, as the absolute value of the correlation coefficient increased, the sampling distribution became more skewed. Thus, the normal distribution cannot be used as the underlying distribution for this test statistic. In order to overcome this problem, the Fisher's z transformation was applied to the correlation coefficients. This transformation produced a sampling distribution that was nearly normal for any value of the correlation coefficient. Second, the estimated standard error of the difference between independent transformed correlation coefficients and test statistics was calculated.

Analysis of the correlations between the spontaneous quotient of the TOWL-2 and teachers' judgements of student's authentic writing samples between students with and without LD produced a z testing for significance of .87. Furthermore, examination of the correlations between the written language quotient on the WLA and teachers' judgements of students' authentic writing samples between students with and without LD produced a

z testing for significance of .27. These statistics fail to present evidence that the test statistics differ for students with and without LD.

Research Question 3

Do measures of construct validity support the use of authentic assessment in written language?

In investigating the construct validity of authentic assessment, a construct which underlies authentic assessment in written language was examined. This construct was analyzed using data from students' report cards and teachers' judgements of authentic writing samples. The construct addressed was: "Because authentic assessment encompasses a basic school subject, students who do well in writing should do well in other areas of school. If true, student performance on student's writings should correlate with language arts."

This assumption, thought to underlie the construct validity of authentic assessment of written language, was calculated using Spearman's rho correlation coefficient because the variables being correlated were ranks. The correlation between teachers' judgements of authentic writing samples and students' grades in language arts yielded a correlation of .40, significant at the .001 level. Because authentic assessment encompasses a basic school subject, students who do well in writing do well in other areas of school (e.g. integrated language arts). This positive correlation substantiates the underlying construct validity of authentic assessment measured in written language.

The results of this study provide minimal evidence of criterion-related concurrent and construct validity as they pertain to authentic assessment practices in written language for students with and without LD. The correlations between teachers' judgements of students' authentic samples and scores on standardized assessments in written language are supported. In addition, the correlation between teachers' judgments of authentic writing samples and students' grades in language arts is validated. Nevertheless, the findings fail to present evidence that test statistics differ for students with and without LD.

CHAPTER 5

SUMMARY, IMPLICATIONS, AND RECOMMENDATIONS

This research study was designed to investigate whether authentic assessment in written language is a valid assessment tool for use with students with and without learning disabilities (LD). In examining the validity of authentic assessment, the construct validity and concurrent validity for narrative writing samples of students with and without LD were explored. The results of this study, implications, and recommendations for future research are provided in this chapter.

Results of the Study

Based on a review of literature, this study was directed by the following research questions:

1. Is there a correlation between teachers' judgements of students' authentic samples and scores on standardized assessments in written language?

2. Do measures of criterion-related concurrent validity differ between groups of students with and without LD?

3. Do measures of construct validity support the use of authentic assessment in written language? Because authentic assessment encompasses a basic school subject,

students who do well in writing should do well in other areas of school. If true, student performance on student's writings should correlate with other academic areas.

General educators and students with and without LD at the fourth- and fifth-grade levels were participants in this study. Forty-six students without LD and 38 students with LD were included, yielding a total sample size of 84 students. Additionally, 10 fourth- and fifth-grade content area regular education teachers were included for the purpose of scoring authentic writing samples.

One narrative authentic writing sample was randomly selected and removed from each student's work folder. An anchor narrative writing sample, which was scored by the Texas Education Agency, was used to measure the consistency of the teachers' judgements. Each teacher scored an equal number of authentic writing samples of students with and without LD and the anchor item according to the Texas Achievement of Academic Skills focused holistic scoring process. The students in the study were administered the spontaneous writing portion of the Test of Written Language-Revised and the creative writing portion of Written Language Assessment, which are two standardized written language assessments. Finally, language arts grades were gathered from students' report cards.

The results of the construct and criterion-related concurrent validity coefficients of authentic assessment in written language for students with and without LD are questionable. The teachers were capable of successfully evaluating student progress. However, the results fail to present evidence that the test statistics differed for students with and without LD.

Implications

Several implications for authentic assessment in written language of students with and without LD are evident from the results of this research study. The discussion and implications of these results are addressed according to the research questions.

Research Question 1: Is there a correlation between teachers' judgements of students' authentic assessments and scores on standardized assessment in written language?

The results of the correlation coefficients between teachers' judgements of students' authentic samples and scores on standardized writing samples indicate that, in the area of written language, teachers have the expertise to produce an accurate evaluation of students' progress. The correlation of scores on students' authentic writing samples and current standardized written language assessments yielded correlations of .45 and .46 on the TOWL-2 and WLA, respectively. This finding demonstrates that the best practices for evaluating students' educational achievement in written language should include the use of teachers' judgements.

The correlations between standardized written language assessments and teachers' judgements of authentic written language samples are supportive of previous research on teachers' judgements and scores on standardized assessments (e.g. DuPaul, Rapport, & Perriello, 1991; Egan & Archer, 1985; Hammill & Hresko, 1994; Wright & Wiese, 1988). The correlation coefficients produced are acceptable; moreover, each correlation is statistically significant, thus validating the use of teachers' judgements in evaluating students' educational achievement in written language.

These data are consistent with earlier findings which support teachers' judgements in expanding, strengthening, and confirming data obtained from educational assessment measures in a sound and valid manner (Gerber & Semmel, 1984; Simeonsson, Huntington, Short, & Ware, 1982). These research results demonstrate that teachers are capable of successfully assessing students' educational achievement in written language.

Research Question 2: Do measures of criterion-related concurrent validity differ between groups of students with and without LD?

The results of the criterion-related concurrent validity indicate that the relationship between students' scores on standardized and authentic written language assessments has poor to moderate validity, depending upon whether or not the students have LD. The results are not consistent between the two samples. Some of the correlations obtained using students with LD provided a moderate correlation and some statistical significance was represented; however the correlations acquired using students without LD yielded poor correlations and little, if any, statistical significance. These criterion-related concurrent validity results are not consistent with the results of recent research conducted by Hammill and Hresko (1994) in which correlation coefficients of .89, .70, .59 and .72 were obtained for investigating the relationship between standardized written language assessments and teachers' judgements of written language ability.

Analysis of the correlation coefficients between the TOWL-2 and WLA suggests that these two written language assessments are more similar than first believed. Correlation of the spontaneous writing quotient of the TOWL-2 with the creative writing portion of the WLA of students with LD yielded a correlation coefficient of .47;

moreover, correlating these same tests with students without LD yielded a correlation coefficient of .58.

Even though the correlations of standardized written language assessments were not significant for students without LD, the two groups were analyzed to determine if differences occurred between the test statistics. Analyis of the correlations of the spontaneous quotient of the TOWL-2 and teachers' judgements of students' authentic writing samples for students with and without LD produced a significant difference of .87. Furthermore, examination of the correlations between the written language quotient on the WLA and teachers' judgements of students' authentic writing samples between students with and without LD produced a significant difference of .26. These statistics fail to present evidence that the test statistics differ for students with and without LD.

The TOWL-2 and WLA already have established criterion-related concurrent validity. These results show some evidence of criterion-related concurrent validity in authentic written language assessments for students with LD. However, these data do not provide evidence of criterion-related concurrent validity in standardized and authentic written language assessments for students without LD. These results are not consistent with the findings of Mercer, Algozzine and Trifiletti (1979), who found that rating scales based on teachers' judgements were legitimate tools for identifying students with and without LD, or the findings of Ysseldyke, Algozzine, and Richey (1982), who found that teachers' judgements were essential in determining students' eligibility for special education services.

The data demonstrate that criterion-related concurrent validity in authentic assessment in written language is lacking. Although there is evidence of criterion-related concurrent validity of authentic assessment in written language with students with LD, the evidence is minimal. This shortage of evidence could be a result of comparing two separate paradigms; however, problems with the TAAS focused holistic scoring rubric (e.g., extent of training, established quantitative measures) are more likely an underlying factor.

Research Question 3: Do measures of construct validity support the use of authentic assessment in written language?

An assumption thought to underlie the construct validity of authentic assessments in written language was, "Because authentic assessment encompasses a basic school subject, students who do well in writing should do well in other areas of school. If true, students performance on their writings should correlate with language arts." The correlation of .40 significant at the .001 level between teachers' judgements of authentic writing samples and students' grades in language arts supports the assumption which underlies the construct validity of authentic assessment of written language. A precise evaluation of students' progress through their educational years is basic to responsible education. Perhaps the immense amount of time teachers spend with students and the very meaningful role they play in the development of students' academic achievement allow teachers to be accurate evaluators of student progress (Sharpley & Edgar, 1986). Because authentic assessment encompasses a basic school subject, students who do well in writing do well in other areas of school (e.g. integrated language arts); therefore, this

finding substantiates the second assumption which underlies the construct validity of authentic assessment in written language.

In conclusion, these results, based upon the delineated research questions, suggest that the various types of validity related to authentic assessment practices in written language are only minimally supported. Results of this study indicate that teachers have the expertise to produce an accurate evaluation of students' progress; however, teachers are not able to differentiate between students with and without LD. As the movement toward inclusion of students with LD into regular classrooms and authentic assessment practices becomes more widespread, the investigation of validity as it pertains to authentic assessment of written language should be further investigated in order to substantiate the utilization of authentic assessment in written language with students with and without LD.

Recommendations for Further Study

This research study contributes to educators' limited knowledge of the validity of authentic assessment in written language of students with and without LD. When attempting to generalize the results of this study, the nature of the sampling methods should be considered; only students in the fourth and fifth grades and their narrative writing samples were included in this research. Furthermore, this study is a validity correlation study and does not allow for causality to be concluded. Considering these limitations, the following recommendations are made for further research:

First, although the inter-rater reliability (i.e., anchor writing sample) obtained in the study was quite important, the study does not provide any other evidence of reliability of authentic assessment. Further investigation into the reliability of authentic assessment practices is recommended.

In order to increase one of the assumptions which underlies the construct validity of authentic assessment in written language, the report cards of students with LD must be taken into consideration. The report cards of students with LD are dependent on students' individualized education plans. According to the teachers, the report card grades of students with LD are not reflective of individual work, but, rather, are a procedure that must be completed. Identifying a sample of students with LD who receive grades reflective of their academic work might produce a higher correlation with teachers' judgements of authentic assessment.

The lack of validity studies in the area of authentic assessment has been noted. If writing is to be emphasized in classrooms for students with LD, it would be revealing to compare students progress over time and determine the predictive validity of authentic assessment practices.

This research study marks the beginning of research in the area of validity of authentic assessment in written language for students with and without LD. Future research should be focused on evidence of reliability, as well as to establishing the validity of authentic assessment. Moreover, identifying students with LD who receive grades reflective of their academic work would extend the current study. If authentic assessment practices are to continue in the educational reform movement, further studies

are recommended in order to ensure the validity of authentic assessment in written language for students with and without LD.

APPENDIX A

EXISTING RESEARCH EXAMINING THE RELATIONSHIP BETWEEN TEACHER JUDGEMENTS AND WRITTEN LANGUAGE

Existing Research Examining the Relationship Between

Teacher Judgements and Written Language

Research Study	**Achievement Area**	**Results**
DuPaul, Rapport, & Perriello (1991)	Language Arts	.61
Egan & Archer (1985)	English	.65
Hammill & Hresko (1994)	Writing	
Diagnostic Achievement Test for Adolescents		.89
Test of Adolescent Language-2		.70
Test of Written Language (Spontaneous Writing Quotient)		.59
Woodcock-Johnson Psycho-Educational Battery-Revised		.72
Hopkins, George, & Williams (1985)	Language Arts	.74
Wright & Wiese (1988)	Language Arts	.76

APPENDIX B

STUDENT DATA INFORMATION FORM

STUDENT DATA INFORMATION

Name _____ Gender: Male/Female
Birthday _____ Ethnicity _____
Grades _____ 87-88 _____ 88-89 _____ 89-90 _____ 90-91 _____ 91-92 _____ 92-93 _____ 93-94

TAAS: Writing Date _____ Grade _____ EXEMPT

Writing Met Expectations: yes no

Narrative Composition Rating: 1 2 3 4 Total Multiple Choice Objectives Mastered _____
Sentence Construction: yes no _____/10 Total Items _____
English Usage: yes no _____/8 Scale Score _____
Use of Spelling, Cap & Punc: yes no _____/10

TAAS: Reading Date _____ Grade _____

Reading Met Expectations: yes no

Word Meanings: yes no _____/6 Total Multiple choice Objectives mastered_____
Supporting Ideas: yes no _____/8 Total Items _____
Summarization: yes no _____/6 Scale Score _____
Relationships & Outcomes: yes no _____/6
Inferences & Generalizations: yes no _____/10
Point of View: yes no _____/4

SPECIAL EDUCATION RECORDS

WISC-R/WISC-III Date _____
Verbal _____
Performance _____
Full Scale _____

WOODCOCK JOHNSON (SS) Date _____
Basic Skills _____ Spelling _____
Letter-Word Identification_____ Proofing _____
Passage Comprehension _____ Word Attack _____
Dictation _____ Broad Reading _____

Eligibility _____
Services Received _____

OTHER INFORMATION

_____

_____

_____

APPENDIX C

TEACHER QUESTIONNAIRE

CONFIDENTIAL TEACHER QUESTIONNAIRE

Name _____ Gender _____

School _____ Ethnicity _____

Number of years teaching _____

Number of years teaching fourth/fifth grade _____

Number of years/months scoring holistically _____ years _____ months

When were you trained from the Writing Collection (Mth/Yr) _____

Education: (Please check and write the year in which degree was received)

_____ Certification in Teaching (degree is from an area other than teaching)

_____ Bachelors' in Education

_____ Masters' in Education

_____Other (please explain) _____

Number of semester hours completed pertaining to written language _____

Number of in-service clock hours completed in the past 3 years in written language _____

Number of semester hours completed in the area of special education _____

Number of in-service clock hours completed in the past 3 years in the area of special education ___

Average amount of time to score one writing sample _____

Amount of time to score all writing samples _____

Number of students with learning disabilities mainstreamed into your classroom during the past 3 years _____

APPENDIX D

DESCRIPTION OF THE TOWL-2

Scores for the spontaneous subtests of the TOWL-2 are obtained by analyzing the quality of the student's freely written story. Below are the five subtests that comprise the spontaneous writing quotient of the TOWL-2:

Thematic maturity

The student writes a story in response to a stimulus picture. Points are earned for each instance in which the student mentions a predetermined element in the story content. For example, (a) Does the student write in paragraphs? (b) Are personal names given to characters? (c) Is dialogue or monologue used? (d) Is time set for the story? and (e) Are events related that occurred prior to those events shown in the picture? are a few instances of the predetermined criteria.

Contextual vocabulary

The vocabulary level of the student's story is evaluated by applying the long unduplicated word method. In other words, the contextual vocabulary score is the number of different words used in the story that have seven or more letters. Made-up words, correctly hyphenated words, addition of a suffix or prefix to a root word, and misspelled words that, if spelled correctly would contain seven or more letters, constitute a scorable word.

Syntactic maturity

The syntactic maturity score is the number of words in the composition that are used to form grammatically acceptable sentences. Unacceptable grammar would include problems in tense and plural agreements; illiterate usages; confusion between "lie" and "lay," "like" and "as," etc.

Contextual spelling

The score for contextual spelling is the number of correctly spelled words in the story. The scorer counts the number of different words in a story that are spelled correctly. Words are counted only once regardless of how many times they are misspelled in the story. Punctuation or capitalization errors are not counted as misspellings. Errors such as "cant" (can't), "Johns" (John's), "zeus" (Zeus), and "ill fated" (ill-fated) would be acceptable.

Contextual style

The student's story is scored for the number of instances in which different punctuation and capitalization rules are used. For example, (a) period at end of a statement, (b) a comma between the day of the month and the year, (c) a comma before the conjunction in a compound sentence, (d) an apostrophe in contractions, (e) capitalization of the word I, and (f) capitalization of titles used with names of persons are instances of contextual style which would scored as correct.

APPENDIX E

DESCRIPTION OF THE WLA

The Written Language Quotient of the WLA is comprised of the General Writing Ability, Productivity, Word Complexity, and Readability scores. The following sections contain explanations of for the use of these scores.

The General Writing Ability is comprised of ratings of the writing samples in three categories: rhetoric, legibility, and overall quality. For the WLA, rhetoric refers to the quality of writing: its style, fluency, use of vocabulary and figures of speech, humor, literary qualities, and eloquence. The WLA standard for handwriting is legibility. Finally, overall quality is a rater's global impression of the writing. Each of these categories is scored holistically, which permits rating a writing sample for the global impression it makes on the rater; thus, any and all aspects of judging writing figure into the rating.

The WLA productivity, word complexity, and readability scores each require counts of various elements of writing. The productivity score of the WLA is simply a count of the total words used in the story. The word complexity score is the number of multisyllabic words in the context of the story. The WLA readability score reveals the average reading difficulty level of the writing by using the number of words, syllables, and sentences in the writing. Each writing sample score is reported as a number that represents a grade-level equivalent.

REFERENCES

Adams, D. M., & Hamm M. E. (1992). Portfolio assessment and social studies: Collecting, selecting, and reflecting on what is significant. Social Education, 56(2), 103-105.

Algozzine, R. Sacca, M., & Maheady, L. (1986). Assessment in remedial and special education: Turn up the lights--The party's not over. The Pointer, 30(2), 508.

Anastasi, A. (1988). Psychological testing (6th ed.). New York: Macmillan.

Anderson, P. L. (1982). A preliminary study of syntax in the written expression of learning disabled children. Journal of Learning Disabilities, 15, 359-362.

Archibald, D. A., & Newmann, F. M. (1988). Beyond standardized testing: Assessing authentic academic achievement in the secondary school. Reston, VA: National Association of Secondary School Principals.

Audette, B., & Algozzine, B. (1992). Free and appropriate education for all students: Total quality and the transformation of American public education. Remedial and special Education, 13(6), 8-18.

Bagnato, S. J., Neisworth, J. T., & Munson, S. M. (1989). Linking developmental assessment and early intervention: Curriculum-based prescriptions. Rockville, MD: Aspen.

Baker, E. L. (1991). Expectation and evidence for alternative assessment. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Barenbaum, E. M., Newcomer, P. L., & Nodine, B. F. (1987). Children's ability to write stories as a function of variation in task, age, and developmental level. Learning Disability Quarterly, 10, 175-188.

Barrett, T. J. (1992). Implementation of an integrated language arts performance assessment in a large urban school district: Technical issues in aggregating and reporting results. San Francisco, CA: American Educational Research Association. (ERIC Document Reproduction Service No. 352 371).

Biklen, D., Ferguson, D., & Ford, A. (1989). Schooling and disability. Chicago: Chicago Press.

Borg, W. R., Worthen, B. R., & Valcarce, R. W. (1988). Teachers' perceptions of the importance of educational measurement. Journal of Experimental Education, 55(1), 9-14.

Bracey, G. W. (1993). Assessing the new assessments. Principal, 72(3), 34-36.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). Assessing writing skill (Research Monograph No. 11). New York: College Entrance Examination Board.

Camp, R. (1993). The place of portfolios in our changing view of writing assessment. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 183-212). Hillsdale, NJ: Lawrence Erlbaum Associates.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. Research in the Teaching of English, 18, 65-81.

Clark, C. M., & Peterson, P. L. (1986). Teacher's thought processes. In M. C. Wittrock (Ed.), Third handbook of research on teaching (pp. 255-296). New York: Macmillan.

Cole, N., & Moss, P. (1989). Bias in test use. In R. Linn (Ed.), Educational measurement (pp. 201-219). New York: Macmillan.

Cooper, C. R. (1977). Holistic evaluation of writing. In C. R. Cooper & L. Odell (Eds.). Evaluating writing: Describing, measuring, Judging (pp. 3-32). Urbana: NCTE.

Cooper, C. R. (Ed.). (1981). The nature and measurement of competency in English. Urbana, IL: National Council of Teachers of English.

Council for Learning Disabilities (1986). Use of discrepancy formulas in the identification of learning disabled individuals: A position statement by the Board of Trustees of the Council for Learning Disabilities. Learning Disability Quarterly, 8(3), 245.

Cronbach, L. J. (1961). Essentials of psychological testing. New York: Harper & Row.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 3-17). Hillsdale, NJ: Lawrence Erlbaum Associates.

Davis, B. G., Scriven, M., & Thomas, S. (1981). The evaluation of composition instruction. Inverness: Edgepress.

Deno, S., Marston, D., & Mirkin, P. (1982). Valid measurement procedures for continuous evaluation of written expression. Exceptional Children, 48, 358-371.

Department of Education and Science. (1987). National curriculum: Task group on assessment and testing. London: Author.

Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. Applied Measurement in Education, 4(4), 289-303.

DuPaul, G. J., Rapport, M. D., & Perriello, L. M. (1991). Teacher ratings of academic skills: The development of the academic performance rating scale. School Psychology Review, 20(2), 284-300.

Egan, O., & Archer, P. (1985). The accuracy of teachers' ratings of ability: A regression model. American Educational Research Journal, 22, 25-34.

Eisner, E. W. (1991). What really counts in schools. Educational Leadership, 48(5), 10-17.

Elbow, P., & Belanoff, P. (1986). Portfolios as a substitute for proficiency examinations. College composition and communication, 37, 336-339.

Elliott, N., Plata, M., & Zelhart, P. (1990). A program development handbook for the holistic assessment of writing. Lanham, MD: University Press of America.

Englert, C. S., & Thomas, C. C. (1987). Sensitivity to text structure in reading and writing: A comparison between learning disabled and non-learning disabled students. leaning Disability Quarterly, 10, 93-105.

Englert, C. S., Raphael, T., Anderson, L., Anthony, H., Fear, K., & Gregg, S. (1988). A case for writing intervention: Strategies for writing informational text. Learning Disabilities Focus, 3(2), 98-113.

Farr, R. (1992). Putting it all together: solving the reading assessment puzzle. The Reading Teacher, 46(1), 26-37.

Farr, R., & Carey, R. (1986). Reading: What can be measured? Newark, DE: International Reading Association.

Frederiksen, J. R., & Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18, 27-32.

Fuchs, L. S., & Fuchs, D. (1990). Traditional academic assessment: An overview. In R. Gable and J. M. Hendrickson (Eds.), Assessing students with special needs (pp. 1-13). New York: Longman.

Gajar, A. (1989). A computer analysis of written language variables and a comparison of compositions written by university students with and without learning disabilities. Journal of Learning Disabilities, 22, 125-130.

Galagan, J. E. (1985). Psychoeducational testing: Turn out the lights, the party's over. Exceptional Children, 52(3), 288-299.

Gerber, M. M., & Semmel, M. I. (1984). Teacher as imperfect test: Reconceptualizing the referral process. Educational Psychologist, 19, 137-148.

Graham, S., & Harris, K. R. (1989). Improving learning disabled students' skills at composing essays: Self-instructional strategy training. Exceptional Children, 56, 201-214.

Graves, A., Montague, M., & Wong, Y. (1990). The effects of procedural facilitation on the story composition of learning disabled students. Learning Disabilities Research, 5, 88-93.

Greenberg, K., Weiner, H. S., & Donovan, R. A. (Eds.). (1986). Writing assessment: Issues and strategies. New York: Longman.

Gregg, K. N. (1983). College learning disabled writer: Error patterns and instructional alternatives. Journal of Learning Disabilities, 16, 334-338.

Gregg, N., & Hoy, C. (1989). Coherence: The comprehension and production abilities of college writers who are normally achieving, learning disabled, and underprepared. Journal of Learning Disabilities, 22, 370-372.

Gresham, F. M., Reschly, D. J., & Carey, M. P. (1987). Teachers as "tests": Classification accuracy and concurrent validation in the identification of learning disabled children. School Psychology Review, 16, 543-553.

Grill, J. J., & Kirwin, M. M. (1988). Written language assessment. Novoto, CA: Academic Therapy.

Gronlund, N. E. (1985). Measurement and evaluation in teaching. New York: MacMillan.

Gronlund, N. E., & Linn, R. L. (1990). Measurement and evaluation in teaching (6th ed.). New York: Macmillan.

Guilford, J. P., & Fruchter, B. (1978). Fundamental statistics in psychology and education. New York: McGraw-Hill.

Hacker, J., & Hathaway, W. (1991). Toward extended assessment: The big picture. Chicago: Paper presented at the Annual Conferences of the American Educational Research Association and the National Council on Measurement in Education. (ERIC Document Reproduction Service No. 337 494)

Hambleton, R. K., & Murphy, E. (1992). A psychometric perspective on authentic measurement. Applied Measurement in Education, 5(1), 1-16.

Hammill, D. D., & Hresko, W. P. (1994). Comprehensive scales of student abilities: Quantifying academic skills and school-related behavior through the use of teacher judgements. Austin: Pro-Ed.

Hammill, D. D., & Larsen, S. (1978). Test of written language--Revised. Austin: Pro-Ed.

Hammill, D. D., & Larsen, S. (1988). Test of written language--Revised. Austin: Pro-Ed.

Haney, W., & Madaus, G. (1989). Searching for alternatives to standardized tests: Whys, whats, and whithers. Phi Delta Kappan, 70(9), 683-687.

Hartle, T. W., & Battaglia, P. A. (1993). The federal role in standardized testing. In R. E. Bennett & W. C. Ward (Eds.), Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment (pp. 291-311). Hillsdale, NJ: Lawrence Erlbaum.

Herman, J, Aschbacher, P. R., & Winters, L. (1992). A practical guide to alternative assessment. Alexandria, VA: Association for Supervision and Curriculum Development. (ERIC Document Reproduction Service No. ED 352 389)

Hoge, R. D. (1983). Psychometric properties of teacher-judgement measures of pupil aptitudes, classroom behaviors, and achievement levels. The Journal of Special Education, 17, 401-429.

Hopkins, K. D., & George C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. Journal of Educational Measurement, 22, 177-182.

Houck, C., & Billingsley, B. (1989). Written expression of students with and without learning disabilities: Differences across the grades. Journal of Learning Disabilities, 22, 561-572.

Hresko, W. (1988). Test of early written language. Austin: Pro-Ed.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. College Composition and Communication, 41(2), 201-213.

Isaacson, S., & Mattoon, C. (1990). The effect of goal constraints on the writing performance of urban learning disabled students. Learning Disability Research, 5, 94-99.

Jongsma, K. S. (1989). Portfolio assessment. The Reading Teacher. 43(3), 264-265.

Keech-Lucas, C. (1988). Toward ecological evaluation. The Quarterly of the National Writing Project and the Center for the Study of Writing, 10(1), 1-17.

Kellaghan, T., Madaus, G., & Airasian, P. (1982). The effects of standardized testing. Boston: Kluwer-Nijhoff.

Keogh, B. K., & Sheehan, R. (1981). Strategies for documenting progress of handicapped children in early education programs. Educational Evaluation and Policy Analysis, 3(6), 58-67.

Killoran, J. (1992). In defense of the multiple-choice question. Social Education, 56(2), 106-108.

Laughton, J., & Morris, N. (1989). Story grammar knowledge of learning disabled students. Learning Disabilities Research, 4, 87-95.

Levi, (1990). Assessment and educational vision: Engaging learners and parents. Language Arts, 67(3), 269-273.

Linn, R. L. (1993). Educational assessment: Expanded expectation and challenges. Educational Evaluation and Policy Analysis, 15(1), 1-16.

Linn, R. L., Baker, E. L., & Dunbar, S. B. (1991). Complex, performance-based assessment: Expectations and validation criteria. Educational Researcher, 20(8), 15-21.

Lucas, C., & Carlson, S. B. (1989). Prototype of alternative assessment strategies for new teachers of English. San Francisco, CA: Report to the California New Teacher Project.

MacArthur, C. A., & Graham, S. (1987). Learning disabled students' composing under three methods of text production: Handwriting, word processing, and dictation. The Journal of Special Education, 21, 22-42.

Maeroff, G. I. (1991). Assessing alternative assessment. Phi Delta Kappan, 73(4), 272-281.

Marzano, R. J. (1990). Standardized tests: Do they measure general cognitive abilities? NASSP Bulletin, 74(526), 93-101.

Marzano, R. J., & Costa, A. L. (1988). Question: Do standardized tests measure cognitive skills? Answer: No. Educational Leadership, 45, 66-73.

Marzano, R. J., & Jesse, D. M. (1987). A study of general cognitive operations in two achievement test batteries and their relationship to item difficulty. Aurora, CO: Mid-continent Regional Educational Laboratory.

McClennon, M. C. (1988). Testing and reform. Phi Delta Kappan, 69, 766-771.

McKendy, T. (1992). Locally developed writing tests and the validity of holistic scoring. Research in the teaching of English, 26(2), 149-165.

McLoughlin, J., & Lewis, R. B. (1990). Assessing special students. Columbus, OH: Merrill.

Mehrens, W. A., & Lehmann, I. J. (1987). Using teacher-made measurement devices. NASSP Bulletin, 71(496), 36, 38-44.

Mercer, C., & Corbett, N. L. (1991). Enhancing assessment for students at risk for school failure. Contemporary Education, 62(4), 250-265.

Mercer, C. F., Algozzine, R., & Trifiletti, J. J. (1979). Toward defining discrepancies for specific learning disabilities: An analysis and alternatives. Learning Disability Quarterly, 2(4), 25-31.

Messick, S. (1989). Meaning and values of test validation: The science and ethics of assessment. Educational Researcher, 18(2), 5-11.

Miller, J. A. (1991). Bush strategy launches "crusade" for education. Education Week X, 31(1), 26.

Mitchell, R. (1989). A sampler of authentic assessment: What it is and what it looks like. Paper prepared for the California Assessment Program Conference, Sacramento, CA.

Moody, D. (1991). Strategies for statewide student assessment. Policy Briefs, Number 17. Washington, DC: Office of Educational Research and improvement. (ERIC Document Reproduction Service No. ED 342 798)

Moran, M. R. (1981). Performance of learning disabled and low achieving secondary students on formal features of a paragraph-writing task. Learning Disability Quarterly, 4, 271-280.

Moran, M. R. (1987). Options for written language assessment. Focus on Exceptional Children, 19(5), 1-10.

Morris, N., & Crump, W. D. (1982). Syntactic and vocabulary development in the written language of learning disabled and non-learning disabled students at four age levels. Learning Disability Quarterly, 5, 163-172.

Myklebust, H. R. (1965). Picture story language test. New York: Grune & Stratton.

Myklebust, H. R. (1973). Development and disorders of written language: Studies of normal and exceptional children. New York: Grune & Stratton.

National Commission on Excellence in Education. (1983). A nation at risk: The imperative for educational reform. Washington: U.S. Department of Education.

National Commission on Testing and Public Policy (1990). From gatekeeper to gateway: Transforming testing in America. Chestnut Hill, MA: Author.

National Governors' Association (1986). A time for results: The governor' 1991 report on education. Washington D.C.: National Governor's Association (ERIC Document Reproduction Service No. ED 279 603)

Neill, M., & Medina, N. J. (1989). Standardized testing: Harmful to educational health. Phi Delta Kappan, 70(9), 688-697.

Neisworth, J. T., & Bagnato, S. J. (1988). Assessment in early childhood special education: A topology of dependent measures. In S. M. Odom & M. B. Karnes (Eds.), Early intervention for infants and children with handicaps (pp. 23-39). Baltimore: Brookes.

Nevo, B. (1985). Face validity revisited. Journal of Educational Measurement, 22, 287-293.

Newcomer, P. L., & Barenbaum, E. M. (1991). The written composing ability of children with learning disabilities: A review of literature from 1980 to 1990. Journal of Learning Disabilities, 24(10), 578-593.

Nodine, B. F., Barenbaum, E. M., & Newcomer, P. L. (1985). Story composition by learning disabled, reading disabled, and normal children. Learning Disability Quarterly, 8, 167-179.

Not as easy as A, B, C. (1990, January 8). Newsweek.

Oakland, T., Shermis, M. D., & Coleman, M. (1990). Teacher perceptions of differences among elementary students with and without learning disabilities in referred samples. Journal of Learning Disabilities, 23(8), 499-504, 520.

O'Neil, J. (1992). Putting performance assessment to the test. Educational Leadership, 49(8), 14-19.

Paulson, F. L., Paulson, P. R., & Meyer, C. A. (1991). What makes a portfolio a portfolio? Educational Leadership, 48(5), 60-63.

Pedulla, J. J., Airasian, P. W., & Madaus, G. F. (1980). Do teacher ratings and standardized test results of students yield the same information? American Educational Research Journal, 17, 303-307.

Peterson, P. L. (1988). Teachers' and students' cognitional knowledge for classroom teaching and learning. Educational Researcher, 17(5), 5-14.

Popham, J. W. (1981). Modern educational measurement. Englewood: Prentice.

Poplin, M., Gray, R., Larsen, S., Banikowski, A., & Mehring, T. (1980). A comparison of components of written expression abilities in learning disabled and non-disabled children at three grade levels. Learning Disability Quarterly, 3, 46-53.

Quinta, F., & McKenna, B. (1991). Alternatives to standardized testing. Washington, DC: National Education Association.

Resnick, L. B. (1987). Learning in school and out. Educational Researcher, 16(9), 13-20.

Resnick, L. B. (1989). Tests as standard of achievement in schools. Paper prepared for the Educational Testing Service Conference. The Uses of Standardized Tests in American Education, New York.

Ripich, D., & Griffith, P. (1988). Narrative abilities of children with learning disabilities and non-disabled children: Story structure, cohesion, and proposition. Journal of Learning Disabilities, 21, 165-173.

Rogers, V. (1989). Assessing the curriculum experienced by children. Phi Delta Kappan, 70(9), 714-718.

Roth, F., & Speakman, N. (1986). Narrative discourse: Spontaneously generated stories of learning disabled and normally achieving students. Journal of Speech and Hearing Disorders, 51, 8-23.

Ruth, L,, & Murphy, S. (1988). Designing writing tasks for the assessment of writing. Norwood, NJ: Ablex.

Salvia, J., & Ysseldyke, J. E. (1991). Assessment in special and remedial education (5th ed.). Boston: Houghton Mifflin.

Sattler, J. (1988). Assessment of children's intelligence and special abilities. San Diego: Jerome M. Sattler.

Schmidt, J., Deshler, D., Schumaker, J., & Alley, G. (1988). The effects of generalization instruction on the written performance of adolescents with learning disabilities in the mainstream classroom. Reading Writing, and Learning Disabilities, 4, 291-309.

Secretary's Commission on Achieving Necessary Skills (1991). What work requires of school: A SCANS report of America 2000. Washington, DC: U.S. Department of Labor.

Sharpley, C. F., & Edgar, E. (1986). Teachers' ratings vs. standardized tests: An empirical investigation of agreement between two indices of achievement. Psychology in the Schools, 23, 106-111.

Shavelson, R. J., & Stern, P. (1981). Research on teachers' pedagogical thought, judgements, decisions, and behavior. Review of Educational Research, 51, 455-498.

Shepard, L. A. (1989). Why we need better assessments. Educational Leadership, 46(7), 4-9.

Shepard, L. A. (1991). Psycholmetricians' beliefs about learning. Educational Researcher, 20(6), 2-16.

Simeonsson, R. J., Huntington, G. S., Short, R. J., & Ware, W. B. (1982). The Carolina record of individual behavior: Characteristics of handicapped infants and children. Topics of Early Childhood in Special Education, 2(2), 43-55.

Simmons, J. (1990). Portfolios as large-scale assessment. Language Arts, 67(3), 264-273.

Snow, R. E. (1989). Toward assessment of cognitive and conative structures in learning. Educational Researcher, 18, 8-14.

Statistical Abstract of the United States (1990). Washington, D.C.: U.S. Bureau of the Census.

Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. In R. O. Friedle (Ed.), New directions in discourse processing (pp. 53-120). Norwood, NJ: Ablex.

Stiggins, R. J. (1991). Facing the challenges of a new era of educational assessment. Portland: Northwest Regional Educational Laboratory.

Swartz, R., Patience, W., & Whitney, D. R. (1985). Adding an essay to the GED writing skills test: Reliability and validity issues (GED Testing Service Research Studies No. 7). Washington, D.C.: American Council on Education (ERIC Document Reproduction Service No. ED 266 288)

TAAS and the writing process: A composition handbook (Grades 3 through 5). Austin, TX: Texas Education Agency.

Tindal, G., & Parker, R. (1989). Assessment of written expression for students in compensatory and special education programs. The Journal of Special Education, 23, 169-183.

U.S. Department of Education. (1991). Thirteenth annual report to Congress on the implementation of the Education of the Handicapped Act. Washington, DC: Division of Educational Services, Special Education Programs.

Vallecorsa, A., & Garriss, C. (1990). Story composition skills of middle-grade students with learning disabilities. Exceptional Children, 57, 48-53.

Valencia, S. (1990). A portfolio approach to classroom reading assessment: The whys, whats, and hows. The Reading Teacher, 43, 338-339.

Valencia, S., & Pearson, P. D. (1987, April). Reading assessment: Time for a change. The Reading Teacher, 40(8), 726-732.

Valencia, S., McGinley, W., & Pearson, P. D. (1990). Assessing literacy in the middle school. In G. Duffy (Ed.), Reading in the middle school (2nd ed., pp. 124-141). Newark, DE: International Reading Association.

Vavrus, L. (1990). Put portfolios to the test. Instructor. 100(1), 48-53.

Vogel, S. A., & Moran, M. (1982). Written language disorders in learning disabled college students: A preliminary report. In W. Cruickshank & J. Learner (Eds), Coming of age: The best of ACLD (pp. 137-157). Syracuse, NY: Syracuse University.

Wallace, G., & Larsen, S. C. (1978). Educational assessment of learning problems. Boston: Allyn & Bacon.

Warner, M. M., Alley, G. R., Deshler, D. D., & Schumaker, J. B. (1980). An epidemiology study of learning disabled adolescents in secondary schools: classification and discrimination of learning disabled and low-achieving adolescents (Research Report No. 20). Lawrence, KS: University of Kansas Institute for Research in Learning Disabilities.

Weiderholt, J. L., Hammill, D. D., & Brown, V. L. (1983). The resource teacher: A guide to effective practices. Austin: Pro-Ed.

White, E. (1985). Teaching and assessing writing. San Francisco: Jossey-Bass.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. Phi Delta Kappan, 70(9), 703-713.

Wilson, V. (1991). Portfolio assessment, psychometric theory, cognitive learning theory: Ships crossing in the night. Contemporary Education, 62(4), 25

Wolf, D. P., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of students assessment. In G. Grant (Ed.), Review of Research In Education, (pp. 31-73). Itasca, IL: Peacock..

Worthen, B. R. (1993). Critical issues that will determine the future of alternative assessment. Phi Delta Kappan, 74(6), 444-454.

Worthen, B. R., & Spandel, V. (1991). Putting the standardized test debate in perspective. Educational Leadership, 48(5), 65-69.

Wright, D., & Wiese, M. J. (1988). Teacher judgement in student evaluation: A comparison of grading methods. Journal of Educational Research, 82, 10-14.

Ysseldyke, J. E., Algozzine, B., & Richey, S. (1982). Judgement under uncertainty: How many children are handicapped? Exceptional Children, 48(6), 531-534.

Ysseldyke, J. E., & Algozzine, B., & Thurlow, M. (1992). Critical issues in special and remedial education (2nd ed.). Boston: Houghton Mifflin.