This is to certify that the

dissertation entitled

EXAMINING THE VALUE OF A PERFORMANCE-BASED ASSESSMENT: A SOCIAL VALIDITY STUDY

presented by

Tanja Lynne Bisesi

has been accepted towards fulfillment
of the requirements for

Ph. D. degree in Educational Psychology

Jaffy C. Raphael
Major professor

Date August 15, 1997

EXAMINING THE VALUE OF A PERFORMANCE-BASED ASSESSMENT: A SOCIAL VALIDITY STUDY

By

Tanja Lynne Bisesi

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology, and Special Education

1997

ABSTRACT

EXAMINING THE VALUE OF A PERFORMANCE-BASED ASSESSMENT:
A SOCIAL VALIDITY STUDY

By

Tanja Lynne Bisesi

I conducted this study to explore the value of a literacy assessment program for meeting the needs of consumers and to examine the potential value of a performance-based assessment for addressing the information gaps in the established assessment program. In Chapter 1, I establish the need for this work by exploring the expanding role of assessment in education and the inadequacies in traditional approaches to studying the value of assessments. In Chapter 2, I present an historical account of assessment in education, including the forces that led to the recent proliferation and diversity of assessments. I also discuss assessment validity lenses for examining the value of assessments, and examine the construct of social validity and its potential value in providing a theoretical framework for studying the value of assessments from the perspective of assessment consumers.

In Chapter 3 through 6, I present a description of my study. I describe the school and classroom where I focused my work, the major participants in the study, and the approach I used to study assessment value, in Chapter 3. In Chapter 4, I describe Highmeadow's literacy assessment program in terms of its constituent tools and available information. The purpose of this chapter is to provide a context for understanding assessment tool use and value. I also

establish that the evolution of Highmeadow's dual-system literacy assessment program was typical of the trend toward expanding, additive, assessment programs in education. In Chapter 5, I analyze patterns of assessment use both across and within consumer groups, by evaluating the tools making up the assessment and identifying the dimensions and properties of assessment tools valued by assessment consumers. In Chapter 6, I explore the value of a performance-based assessment in terms of its potential for meeting the assessment needs of consumers.

In Chapter 7, I discussed the implications of my work. My findings have practical implications for the integration of Highmeadow's literacy assessment program as well as the design of literacy assessment programs more generally. Findings also have theoretical implications for how we study and evaluate the assessment tools and programs we develop.

ACKNOWLEDGEMENTS

I would like to express thanks to those who made work and life possible over the course of this project. First, I would like to thank the faculty and staff in the College of Education at MSU. Specifically, I want to thank my dissertation committee, Taffy E. Raphael, P. David Pearson, Laura Roehler, James Gavelek, and Carol Sue Englert for their helpful feedback on drafts of this paper. Special thanks go to my dissertation co-chairs Taffy Raphael and P. David Pearson. Without Taffy, this project would never have started, more less been completed. Her expertise as a researcher can only be matched by her skill as a mentor. Her guidance, support, and enthusiasm allowed me to develop both competency and confidence as a researcher. Throughout this project and others, she taught me research methodology and more importantly, how to think and write. And despite the detours I made in completing this work, Taffy continued to support me in both my research and my life. I would also like to thank my co-chair P. David Pearson. His timely arrival at MSU during the final stage of my dissertation analysis and writing was extremely fortuitus. His expertise in literacy and assessment are only rivaled by his caring and supportive nature, without which I would have

been lost. His broad disciplinary knowledge offered me much insight into my research data and the field of assessment. I will always treasure our many discussions and our friendship. Finally, I want to thank the College of Education for rewarding my efforts with a fellowship which not only encouraged my timely completion of this project but inspired my confidence in the quality of my work.

Second, this project would not have been possible without the faculty and staff at Highmeadow. They made me feel welcome at all times, allowing me to borrow equipment and supplies and tolerating my inquiring presence. Sincere thanks also go to project participants Joan, June, and the 26 fifth-grade students and their parents. Without their commitment to filling out surveys, making themselves available, and answering questions, I would never have been able to conduct this research. Special thanks go to June for being both colleague and friend. Our many discussions about this project fueled my enthusiasm. We were both pregnant during the course of data collection which provided an additional degree of camaraderie.

This acknowledgment would not be complete without expressing my appreciation to my family. My undying gratitude goes to my husband, Mark, for tolerating my "moods," listening to all my crazy ideas, sharing his expert editorial eye, and enduring the journey of this project with me. He has been my life-line and anchor: my colleague, friend and spouse. I also want to thank my beautiful daughter, Abigail, who provided me with a refuge of joy

and fun when the pressure of my work became unbearable. And finally, my deepest and sincere thanks go to my Father and Mother for encouraging my love of learning and giving me the self-confidence to pursue and achieve this accomplishment.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

CHAPTER ONE

ESTABLISHING THE RESEARCH PROBLEM

One of the most important issues in literacy education today is assessment. There is a great deal of literacy assessment taking place in our schools, and the amount of testing is increasing. Past disappointment with student literacy achievement and demands for greater accountability have been partially responsible for this increase, as policy makers and the public desire more information for judging the quality of literacy education (Farr, 1992). Standardized tests have typically been used to fulfill this accountability need.

The value of standardized assessments has been established, historically, through psychometrically-grounded validity methods. Psychometric validity frameworks have emphasized the value of assessment in terms of how accurately (i.e., construct validity) and consistently (i.e., reliability) it reflected some trait and/or domain of interest (i.e., content validity) for some particular purpose (e.g., accountability). An additional feature of standardized tests has been their cost-effectiveness in terms of data collection and scoring (e.g., efficient, objective). Because these tests provided trustworthy (i.e., reliable), objective (i.e., constrained-response), and efficient (i.e., machine-scored) measures of achievement and accountability, high-stakes standardized test use continues to proliferate.

Despite benefits, the proliferation of test use has had adverse

consequences. Standardized tests have been found to align poorly with curriculum (e.g., Bisesi & Raphael, 1997; Raphael, Wallace, & Pardo, 1996), narrow the scope of content covered during instruction (Smith, 1991; Shepard, 1989) and result in poor learning and performance motivation (Paris, Lawton, Turner, & Roth, 1991). Thus, in response to these limitations validity frameworks have recently been expanded to include criteria for the evaluation of tests in terms of their consequences (e.g., Messick, 1989b; Cronbach, 1988).

Further, the growing awareness of the negative impact of assessments (e.g., standardized tests) used for high-stakes purposes (i.e., accountability) and shifts in the prevailing theories and assumptions underlying literacy have driven the search for alternative assessments that better reflect current literacy theory and curricula. The proliferation of performance-based assessment use is one manifestation of the search for a better alternative. But what do these assessments have to contribute to literacy assessment in our schools? This question has only been partially answered by the expanding literature on performance-based assessments.

Recently developed performance-based assessments have been designed to remedy many of the limitations of standardized tests (Baker, O'Neil, & Linn, 1993) by providing an "authentic and direct appraisal of educational competence" toward the improvement of teaching and learning (Messick, 1994; p. 13). Researchers have found that these kinds of assessments

better reflect current views of literacy and school curricula (e.g., Bisesi & Raphael, 1997). Research also suggests that these assessments empower teachers to take control of their classroom practice (e.g., instruction, assessment) and professional development (e.g., Stewart, Paradis, & Aegerter, 1992), and involve students in meaningful learning and reflection (e.g., Tierney, Carter, & Desai, 1991). Nevertheless, the popularity and proliferations of performance-based assessment use has contributed to a further increase in the overall assessment of literacy students (Farr, 1992).

As school-based literacy assessment programs become larger and more complex, and begin seriously to impose on the resources of schools and the instructional time of teachers and students, their value must be appraised. We must decide which assessments are worth including in an assessment program and discard those that are not. As we develop and implement alternative assessment tools (e.g., performance-based assessments) and assessment programs that include diverse information sources (e.g. standardized tests, performance-based assessments), it becomes imperative to have guidelines for judging the value of assessment programs as a whole and their constituent parts (i.e., individual tools).

Despite the enthusiasm for performance-based assessments in the evaluation of literacy learning and performances, it is not yet clear whether their potential contribution justifies their effect of expanding school-based literacy assessment programs. Psychometrically-grounded validity

frameworks developed specifically for evaluating performance-based assessments (e.g., Haertel, 1991; Linn, Baker, & Dunbar, 1991; Frederiksen, & Collins, 1989) have provided guidelines for judging the value of these assessments. Nevertheless, these frameworks continue to emphasize scientific, theory-oriented, hypothesis testing as a basis for warranting interpretations (i.e., construct validation) and remain limited in that they focus on the use of psychometric procedures (e.g., statistical analyses) and evidence toward the understanding of scientific constructs and researcher interpretations. And while these frameworks consistently include guidelines for examining the consequences of assessment use (i.e., consequential validity), they continue to stress the value of assessments in terms of their "technical" psychometric features including reliability (e.g., generalizability), objectivity, and efficiency, features on which performance-based assessments have been found lacking (e.g., Wainer & Thissen, 1993).

As my discussion makes clear, particular forms of assessment are favored over others depending on the frame of reference (i.e., validity perspective) one takes in judging assessment value. Technical criteria tend to favor standardized tests. Consequential criteria often favor alternative assessments including those that are performance-based. The question then becomes, how should we go about determining the value of various assessment tools when designing literacy assessment programs? How should we decide which assessment tools ought to be included in our program and

which ones should not?

Social validity (Wolf, 1978) provides us with an alternative viewpoint for studying the value of assessments and justifying their inclusion in assessment programs. The construct of social validity emerged from the discipline of behavioral analysis. Behavior analysts used the construct and methods of social validity to develop intervention programs that were considered socially important (i.e., valuable), namely, appropriate and worthwhile to those who used them (e.g., students, parents, teachers). The most commonly used method for collecting social validity data is questioning (i.e., in the form of surveys or interviews) consumers of a given intervention program about whether they approve of the program including its goals, procedures, and outcomes. For example, consumers are asked, "Do you think this program is of value?" and "What exactly are your likes and dislikes? In other words, the criteria for judging the value of a program from this perspective is simply consumer satisfaction. Thus, a social validity perspective on assessment programs would mandate the exploration of assessment value from the point of view of the consumer.

Thus far, assessments have not been evaluated in terms of their value from the perspective of assessment consumers (i.e., school administrators, teachers, students, parents). Several authors (e.g., Valencia, Hiebert, & Afflerbach, 1994; Farr, 1992) have noted that different assessment consumers have different needs. Others (e.g., Shepard & Bliem, 1995) have examined

consumer valuing of different types of assessments. These authors discuss the importance of addressing consumer values in the selection and development of assessments, yet they have not actually asked consumers what their needs are or whether various assessments can be used to address these needs. Farr (1992), for example, stated that school administrators preferred standardized tests to meet their need for making decisions on the school level, while students and teachers preferred performance-based assessments that were grounded in classroom activity. Actual uses and needs for assessments have been assumed, not validated through empirical study. Furthermore, none of these authors suggested that an analysis of consumer uses and needs be applied in the process of determining an assessment's value.

Most assessment consumers (e.g., administrators, students, parents) are consumers of assessment *information* (e.g., scores, descriptive interpretations, standards) not the assessment tools themselves. Assessment information is a logical focus for exploring the value of assessments from the perspective of consumers. The literature on information value frames worth of assessments in terms of the degree to which resulting data improves necessary decision making (Pearson & Garavaglia, 1997). And while researchers have defined and studied information value from both psychometric, and psychological-construct perspectives, the perspective of relevant assessment consumers have not been considered. The assessment

information needs and uses of those individuals who are the primary consumers of that information have not been recognized in either the development or evaluation of assessments and assessment programs (Pearson, 1997).

Social validity offers an important perspective for establishing the information value of literacy assessment tools and justifying their inclusion in literacy assessment programs. This perspective compliments psychometric, construct-oriented, and consequential lenses that are reflected in the assessment validity literature. Assessment validity frameworks focus almost exclusively on the agendas and needs of assessment developers and the validation of scientifically-oriented interpretations through psychometric methodologies, but ignore the values of assessment consumers.

The addition of the consequential validity concept to psychometric validity frameworks (e.g., Messick 1989b) recognized the social value of assessments in terms of their impact on the educational system (including assessment consumers), but failed to provide society with an active role in the development of assessment. The consequential validity perspective represents assessment consumers as relatively passive receivers of assessment interpretations, not active knowledgeable users of assessment information. Furthermore, social *consequences* can only be evaluated *after* a program has been implemented and used for a period of time, offering little guidance in the initial *design* of assessments and programs.

The social validity perspective, on the other hand, considers the values and needs of assessment consumers and encourages them to be actively involved in the selection and development of assessment tools and programs. Evaluating assessment from this perspective has the potential to remedy so-called "misinterpretations" and "abuses" of assessment information. If consumers receive assessment information that meets their needs, they will not be forced to use available yet inappropriate information for these purposes. Thus, social validation could provide an understanding of how assessment consumers use and value assessment tools and information made available to them. It could also provide insights on assessment-consumer information needs as well as unnecessary information redundancies within a literacy assessment program.

Performance-based assessments have demonstrated a unique potential to contribute to literacy assessment programs, by providing direct indexes of student performance on meaningful tasks relevant to curriculum and encouraging positive consequences for literacy instruction and learning. It is not yet clear, however, what role they might play in addressing the information needs of assessment consumers within the context of expanding assessment programs. In the present work, I set out to examine the value of an assessment program for meeting the needs of consumers and the potential value of a performance-based assessment for addressing the information gaps in the program. I begin in Chapter 4 by describing the focus of my case study

(Merriam, 1988), Highmeadow's literacy assessment program, in terms of the tools making up the program and information resulting from it. In Chapter 5, I use a social validation research design (e.g., Wolf, 1978), in conjunction with the constant comparative method of analysis (Glaser & Strauss, 1967) and other qualitative research procedures (e.g., Bogden & Biklen, 1992) to establish patterns of assessment-consumer (i.e., school administrator, teacher, parents, students) information use and valued assessment dimensions and properties, and to identify information redundancies in the literacy assessment program. In Chapter 6, I examine the value of the performance-based assessment for meeting the unaddressed information needs (i.e, gaps) of assessment consumers.

CHAPTER TWO

REVIEW OF THE LITERATURE

My desire to understand the information uses and values of assessment consumers led me toward two bodies of research. The first covers the history of assessment in education, including those forces which have resulted in increased assessment use. This work is important because it provides a context for understanding the proliferation of educational (including literacy) assessment programs. It also helps identify the consumers who have historically cared about and used educational assessments and assessment information. Finally, it highlights the need for useful, integrated literacy assessment programs.

The second body of research focuses on assessment validity. This work represents, both historically and conceptually, shifting perspectives on assessment development and evaluation. This literature is significant because it not only reflects the bases on which assessments have historically been valued, but it also contributes to our understanding of the paradigm in which the evaluation of assessment has been undertaken. Validity inquiry has been the primary means for systematically deciding which assessments should be used and which should not in a given context, a goal of the present study.

In this chapter, I first present an historical account of assessment in education including the forces that lead to the proliferation and diversity of

assessments. Next, I discuss assessment validity, perspectives on its study, what we have learned from this research about the value of assessment tools and information, and its limitations. In the latter part of this chapter, I explore the construct of social validity and its potential to provide a theoretical framework for exploring the value of assessments from the perspective of assessment consumers.

In Chapter 3 through 6, I present a description of my study. Chapter 3, describes the school and classroom where I focused my work, the major participants in the study, and the approach I used to study assessment information utility and value. Chapter 4 summarizes the school's established literacy assessment program including characteristics of the assessment tools used, when they were implemented and by whom, and information available to assessment consumers. Chapter 5 characterizes established assessment-tool use across and within consumer groups with the goal of identifying information redundancy in the literacy assessment program. I also present the dimensions of assessments that were critical to consumers in their valuation of assessments and how they decided to use information for particular purposes. In Chapter 6, I discuss the performance-based assessment as an example of how social validity inquiry can be used to identify valued assessment properties not addressed by the established assessment system. I also examine the potential value of the performance-based assessment for addressing those properties. I explore the information each consumer group

indicated that they needed and then examine the value of the performance-based assessment for addressing the information needs of assessment consumers.

Assessment in education

The literacy assessment program at Highmeadow represents the trend toward ever-expanding assessment programs in education. The dual assessment systems making up the program are represented by externally-mandated (i.e., outside the classroom) standardized testing and classroom-based teacher-initiated assessments. This dual-system program is typical of historical trends in educational assessment. Thus, the assessment program which is the focus of this study provided a rich context for exploring assessment consumer information uses and value and vividly illustrated the need to maximize the information provided by an assessment program while reducing the overall amount of assessment taking place.

Highmeadow's literacy assessment program has evolved for at least the last 10 years. The historical influences which are directing current trends in educational assessment were in place long before the initiation of Highmeadow's literacy assessment program. Examination of this history provides a frame for understanding the characteristics of Highmeadow's literacy assessment program described in Chapter 4. My overview of educational assessment focuses on two major historical trends: (1) the rise of large-scale standardized testing in education, and (2) the evolving features of

teacher-initiated classroom-based assessment of student literacy performance. The standardized testing trend is characterized by technical knowledge (i.e., psychometric theory) and empirical research generated by measurement experts. The classroom-based assessment trend is grounded in shifts in literacy theory, as well as subject-matter knowledge and logic applied by teachers in the everyday practice of teaching (see Table 1 for contrasting characteristics of standardized testing and classroom-based assessment).

Table 1-Characteristics of standardized testing & classroom-based assessment

CHARACTERISTICS	Standardized testing	Classroom-based assessment
Authority	Measurement experts	Teachers
Grounding principles	Psychometric theory	Curricular and instructional practice
Intended audience	Administrators, policy makers, researchers, (parents)	Teachers and students (parents)
Purpose	Sorting students and program evaluation	Instructional planning and student evaluation
Interpretative frame	Norm-referenced	Curriculum-referenced
Form of interpretation	Scores	Descriptions

Each trend is relatively distinct, but both contribute to our understanding of the diverse perspectives on educational (e.g., literacy) assessment, the increasing amount of assessment in education, and the expanding role of performance-based assessment in literacy assessment programs.

Standardized testing in education

The story of standardized testing in education began during the 1920s,

with compulsory education and the need to equitably distribute scarce resources for the purpose of educating large numbers of diverse students (Stiggins, 1991). With record numbers of students flocking to schools, the efficiency movement in education and the search for scientific solutions to emerging problems began. While the assembly-line organization of schools (e.g., linear progression of grades, standard curriculum) and the fixed school year provided a more efficient educational system, administrators needed assessments that allowed the sorting of students (according to their achievement in school and their potential to succeed in college) and the allocation of educational resources. In an attempt to make this sorting process fair and equitable (as well as efficient), schools called on the measurement community to develop "scientifically precise" tools (i.e., standardized multiple-choice tests) that would be more useful than the cumbersome and subjective judgments of teachers. Policy makers also wanted to evaluate the effectiveness of new school programs, as schools strove to become better through the application of scientific principles (Farr, & Carey, 1986). As a result, the science of educational testing exploded.

Because the science of educational testing originated in the psychological measurement community, it reflected their perspectives and methods. The empirically-oriented behavioral paradigm dominated the psychological measurement community in the early 1900s, and the science of educational testing was an instantiation of that paradigm. Educational tests

emphasized the objective (i.e., single correct answer), reliable (i.e., consistent) measurement of observable student academic behavior under controlled conditions (i.e., standardization). In the hands of measurement experts, educational assessment quickly became highly technical. Educational assessment required specialized knowledge and training (i.e., psychometric theory and methodology). It also became increasingly distant from the goals of instruction and the concerns of teachers and students in classrooms across the country (Stiggins, 1991).

As the science of educational testing became refined, policy makers became more reliant on the efficient new science and technology of assessment (Stiggins, 1991). Increasingly centralized (and expensive) assessment programs were implemented on the district, state, and the national levels for the purpose of accountability. This escalation is partially to blame for the over-use of assessment in education today, including the large standardized testing component of many school-based assessment program.

Classroom-based literacy assessment

The history of classroom-based assessment is a different story. Classroom-based assessments originated in the everyday practice of teachers in the schools. Consequently, they did not received the attention of researchers (until recently, with a shift in the educational research paradigm toward the study of educational phenomena in the social and historical contexts in which they occur), in contrast to standardized tests which were

researched from their conception. Nevertheless, we can see these assessments embedded in and reflecting literacy instructional practices through history.

Literacy instruction during the early 1900s focused on teaching oral reading. This instruction targeted skills such as decoding, fluency and other basic skills (e.g., spelling, handwriting). Teachers required a means to evaluate student performance of these skills. Teacher evaluation often consisted of informal judgments about student performance observed during the course of teaching. The scientific movement in education (beginning with 1909 publication of Thorndike's writing scale) lead to the publication of various performance scales (e.g., Gray's Standardized Oral Reading Paragraphs) which supplemented classroom observation and teacher evaluation (Smith, 1965). The availability of basal readers, graded by controlled vocabulary, also allowed teachers to evaluate student reading level.

As the scientific movement in education became increasingly predominant through the 1950s, so did the role of basal readers and their associated skill-management systems. Skills (e.g. sight vocabulary, decoding) were operationalized in the form of scope and sequence charts which teachers "checked off" as students demonstrated performance mastery (usually according to some quantitative criteria). Used in concert with basal readers, these tools provided teachers with an efficient way to evaluate students on

specific skills that were the focus of instruction. Because the scientific movement that these classroom-based assessments grew out of also launched the standardized testing movement, standardized tests constituted an effective means for assessing literacy skills that were the focus of instruction at this time. Thus, at this point in history, there was little malalignment (Bisesi & Raphael, 1997) between classroom instruction and standardized testing, a fact which facilitated the creation of standardized testing programs.

Later, the cognitive revolution in psychology impacted both reading research and subsequent instruction. As early as the 1970s, this revolution ushered in a period of reading instruction which focused on comprehension processes and strategies. Reading teachers assisted student in the comprehension of text through a series of pre-, during- and post-reading activities. For example, teachers had students use self-questioning strategies to encourage understanding. Oral and written summaries of text served as performance artifacts for the assessment of students' ability to use self-questioning to enhance comprehension. And while it was still possible to indirectly assess some comprehension strategies using standardized tests (e.g., identifying main ideas), the increasing focus of instruction on comprehension as a process (which is reflected in the pre, during, and post-reading strategies) reflected a growing malalignment between the activities assessed by teachers in literacy classrooms and the tasks characterizing standardized tests. This malalignment created a foundation for the

increasingly critical stance of teachers toward standardized testing.

Current literacy instruction, which draws on reader response (e.g., Langer, 1990) and socio-historical theory (e.g., Vygotsky, 1978), emphasizes the personal (e.g., opinions), social (e.g., discussion) and multidimensional (i.e., reading, writing, listening and speaking) nature of literacy. In today's literacy classrooms, students engage in a range of complex literacy tasks that can only be evaluated through direct observation of student performance. Performance on tasks of this nature is not easy to infer from scores on a constrained-response standardized test. The need to evaluate complex performances that reflect current classroom instructional practice has led to a growing recognition of performance-based assessments that are grounded in the instructional activities of the classroom. Despite the growing recognition and use of performance-based assessments, they have not replaced standardized testing in education. Performance-based assessment has become an "add on" to many assessment programs, contributing to the ongoing expansion of educational assessment and highlighting the need to evaluate their expanding role.

The two stories of assessment in education are unique. Nevertheless, when considered together they provide insight into historical forces which contributed to the proliferation of assessment in education including the push for increased educational accountability. This discussion also outlines characteristics of standardized testing and classroom-based assessment, factors

shaping current perspectives on assessment, and the expanding role of performance-based assessment in growing literacy assessment programs. In light of this historical perspective, I now turn to a consideration of assessment validity lenses and their role in the evaluation of educational assessments.

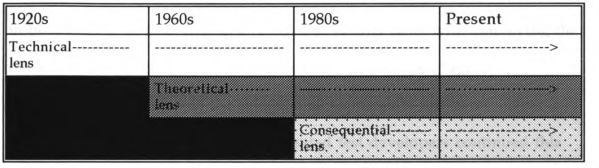
Validity lenses and the value of assessments

My interest in the value of assessment tools and information also led me to examine the literature pertaining to assessment validity. Because the value of assessments has historically been determined through the study of their validity, I was interested to see how other researcher had conceptualized value and evaluated assessment tools. I learned that the concept of validity was born in the field of psychological measurement in the last decade of the 19th century (Anastasi, 1993). And while the study of assessment validity has changed over the course of history, reflected in the differing validity lenses by which assessments have been explored, the concept was built upon and continues to reflect its psychometric roots. Grounding in the principles of classical test theory perpetuates a preoccupation with the technical procedures of science applied for the purpose of furthering scientifically-grounded psychological theory.

Nevertheless, the concept of assessment validity has become multi-dimensional and layered, as researchers become increasingly sensitive to emerging theoretical (i.e., constructs) and practical concerns (e.g., consequences). The additive nature of change in the conceptualization and

research on assessment validity is reflected in Figure 1.

Figure 1-The emergence of lenses on assessment validity



And while there have been notable attempts to create a more unified and integrated view of validity such as the model proposed by Messick (1989b), even his "progressive matrix" communicates an "additive," not evolutionary quality (e.g., the evidential basis of test interpretation is conceptualized as construct validity (CV), while the evidential basis of test use is CV+Relevance/Utility). Thus, this work on assessment validity has made a significant contribution toward broadening our perspective on the value of assessments. Yet, it does not provide insight on assessment value from the point of view of those who use them.

Technical lens: valuing assessments as scientific measurement

The technical lens on validity in educational assessment can be traced back to the psychological testing movement of the last decade of the 19th century (Anastasi, 1993). The movement was grounded in the psychophysical experiments of Wilhelm Wundt, James McKeen Cattell's interest in mental

measures, and the individual differences tradition of Sir Francis Galton (Resnick, 1982). Binet and Simon's intelligence test work in France at the turn of the century was the first application of this new science of testing to problems of education. Anastasi (1986) also credits Binet with employing the first scientific approach to the evaluation of tests, using an "age-differentiation" criterion in the selection of appropriate test items. Over the next several decades, educationally-oriented test developers continued this trend, applying increasingly complex technical procedures to the evaluation of tests. These procedures included statistical item analyses (e.g., internal reliability, factor analysis) as well as analyses for determining the relationships (e.g., predictive) between test scores and other external criteria like diagnostic category (e.g., mental retardation), and teacher judgment of performance (Anastasi, 1986).

During the first half of the 20th century as technical psychometric procedures became more sophisticated, there was no consensus on the recommended approach to the validation of assessments. Procedures applied at the time were diverse as test researchers attempted to establish the value of the tests they developed. Nevertheless, so-called "validity research" reported by test developers was confusing. Tests were evaluated in terms of their intrinsic validity, face validity, and logical validity (e.g., Gulliksen, 1949), to name only a few. Anastasi (1954) attempted to create order out of this chaos by organizing validity research into a three category framework (see Figure 2).

Figure 2-Traditional validity frameworks

Anastasi (1954)	Technical Recommendations (1954)	Standards (1966; 1974; 1985)
Content	Content	Content
Empirical	Predictive	Criterion-related
	Concurrent	
Factorial	Construct	Construct

Her framework included procedures and evidence relevant to content, empirical (similar to what we now call criterion-related validity), and factorial validity (similar to what we now know as construct validity).

While there was no consensus on the ideal approach to validation, there was one technical condition that was required of all valid measurement: reliability, or consistency in measurement. Reliability came in several varieties (e.g., test/retest, parallel-forms, split-half, inter-judge) depending on the nature of the measurement. Nevertheless, measurement reliability was a reflection of the dependability of the measure. Thus, all approaches to the evaluation of tests mandated the examination of reliability.

The publication of the Technical Recommendations for Psychological Tests and Diagnostic Techniques (1954) by the American Psychological Association and the 1955 Technical Recommendations for Achievement Tests by the American Educational Research Association and the National Council on Measurement in Education helped to establish some consensus. These recommendation documents outlined the types of validity that ought

to be addressed (i.e., content, predictive, concurrent, and construct validity) during test development (as well as the types of reliability), and the procedures for collecting and analyzing validity evidence.

The psychometric-grounded technical lens for valuing assessments was codified in the validity frameworks (including four types of validity) published in the Technical Recommendations (1954, 1955). The four types of validity outlined by the Technical Recommendations were believed to be relevant to the evaluation of any test depending on the testing purpose. While different types of validity were thought to be more critical to establish for particular kinds of tests used for specific purposes, technical aspects of the tests (as suggested by the term "technical recommendations") were the focus of evaluation. The systematic evaluation of the appropriateness of test items (i.e., content validity) was considered most relevant to academic achievement tests where a test's focus was curriculum content. Construct validity was an obscure form of validity reserved for psychological tests (e.g., of affect or personality) and involved the testing of scientific hypotheses. Finally, concurrent and predictive validity were demonstrated by data from correlational analysis between the test and other related measures of current and future status, respectively.

By the time of the publication of the 1966 Standards for Educational and Psychological Tests, the four types of validity had been condensed into a tripartite framework including content validity, criterion-related validity

(subsuming predictive and concurrent), and construct validity (as well as reliability), that has persisted through the publication of the 1974 and 1985 Standards for Educational and Psychological Testing. The most recent set of standards maintains the three-pronged framework, but also reflects changing conceptions of validity by moving toward the broader notion assumed today.

Contemporary validity frameworks, like those designed for the evaluation of performance-based assessments (e.g., Haertel, 1991; Linn, Baker, & Dunbar, 1991; Frederiksen & Collins, 1989) have become broader and more inclusive (e.g., consequential validity criteria). Nevertheless, they continue to emphasize technical validity criteria. The need for human judgment in the evaluation of complex performances and the fact that the primary purpose is to generalize these assessments to broader contexts frequently using information from only one assessment tool (i.e., high stakes), encourages a focus on technical validity criteria such as reliability (e.g., agreement between judges), generalizability (e.g., transfer across time, task and situation), and standardization (e.g., controlled testing conditions). Thus, the technical lens continues to be important in validity research on performance-based assessments.

Other researchers have suggested the use of methodological approaches coming out of interpretative traditions, including prolonged engagement, multiple sources of evidence, and reactions from colleagues, for expanding the models for studying validity (e.g., Moss, 1992; Johnston, 1989). Moss

(1994), for example, drew on the interpretative, hermeneutic tradition to create an alternative model to interrater reliability for warranting interpretations. From this perspective raters should be asked to discuss and negotiate differences in interpretation in an attempt to come to some consensus, in contrast to providing independent ratings (Moss, 1994).

Delandshere and Petrosky (1994) applied a methodology consistent with Moss' (1994) model. They grounded interpretations of teacher performance and produced consistent judgments through a consensus-building procedure, in contrast to psychometric standardization of tasks, procedures, and scoring. This methodological approach involved the development of a shared understanding of critical performance dimensions between professional judges, the triangulation of multiple converging evidence (e.g., artifacts, responses to questions), a professional interpretation in the form of a written interpretative summary, and confirmation or disconfirmation by a second professional. While these two examples represent a change in the conception of what counts as evidence for validity claims (i.e., evidence of consensus versus independent agreement), their focus continues to be a technical aspect of assessment, namely, reliable interpretation and scoring.

In general, the technical lens on assessment value has served the educational measurement community well in evaluating constrained-response standardized tests, which became increasingly popular in education

during the 1930s through the 1950s (Hallam, 1995). This lens helped test developers establish the soundness of tests used during the first half of this century when there was a concern about the lack of consistency in informal teacher judgments (Hallam, 1995), a need to assess mastery of discrete, rules and skills which were the typical focus of curricula of the day (Shepard, 1989; Langer, 1990), and mounting pressure for an efficient, cost effective way to assess large numbers of individuals for accountability purposes (Stiggins, 1991). Moreover, contemporary validity researchers including those designing performance-based assessments (e.g., Haertel, 1991; Linn, Baker & Dunbar, 1991) and those exploring the value of assessment using alternative methodological (e.g., Moss, 1994) continue to believe that these technical aspects of validity are important.

The limitations of the technical lens were highlighted by its neglect of emerging issues in psychological measurement (i.e., role of theory in assessment). A growing concern for the role of theory in psychology and education led to a broadening of the assessment validity concept and the addition of the theoretical lens for judging the value of assessments.

Theoretical lens: valuing assessments as tools of theory

A broadening of the validity concept was initiated through the introduction of the construct validity concept in the 1955 Technical Recommendations for Achievement Tests (AERA & NCMUE) and the 1954 Technical Recommendations for Psychological Tests and Diagnostic

Techniques (APA). In the 1955 publication, the construct validity of educational tests was characterized in terms of discriminating power (ability to discriminate between students in predictable ways) in conjunction with content validity evidence. The 1955 Technical Recommendations also reported the importance of factorial studies, and external correlational data in the establishment of a test's construct validity. Finally, the recommendations suggested the need to outline the theory underlying the test and present data that supported the theory.

Cronbach and Meehl (1955), following their participation in developing the recommendations, published an article describing specific methods for the establishment of construct validity. In their article, they suggested that the notion of construct validity should be used "to specify how one is to defend a proposed interpretation of a test" (p. 282). While they argued that construct validation might be important to investigate for any type of psychological test (i.e., achievement, aptitude, interest), they recommended that it was most relevant to tests in which test behavior or its relationship to a criterion measure were not of interest, but in which a theoretical *construct* representing an underlying trait and explaining test behavior was the focus of study.

Since the introduction of the notion of construct validity, it has assumed an increasingly central role in the assessment validity literature. Loevinger, as early as 1957, argued convincingly and at length (nearly 60

pages) that, "construct validity is the whole of validity" (p. 636). This point was echoed later by Anastasi (1961) who described construct validity as "a comprehensive concept, which includes the other three types" (p. 150). Anastasi (1961) extended the argument further by suggesting two contributions that the concept of construct validity could make to psychological testing. She argued that the construct validity concept not only brought attention to the importance of grounding test construction in explicit theoretical foundations, but it precipitated the search for novel ways of collecting validation evidence (Anastasi, 1961).

These discussions also foreshadowed attempts to integrate the validity concept around construct validity (e.g., Moss, 1992; Messick, 1989b; Cronbach, 1988). Cronbach (1971) extended the argument made by Loevinger (1957) and Anastasi (1961) that all forms of measurement, even educational measurement, needed to be validated in terms of construct validity. In this context he argued that, "whenever one classifies situations, person, or responses, he uses constructs" (p. 462). In other words, even subject-matter learning has associated theoretical constructs (Messick, 1975). By the early 1980s, subject-matter research (e.g., reading) grounded in the cognitive paradigm further supported the validation of educational achievement tests in terms of their underlying theoretical constructs. For example, Curtis and Glaser (1983) recommended that tests of reading achievement be grounded in reading theory in order to provide for meaningful interpretations of test

scores. While highlighting the potential role of theory in test develop, Curtis and Glaser (1983) continued to support technical, psychometric standards, stating that the goal of test development should be, "to integrate better the two worlds of psychometrics and experimental psychology" (p. 143)."

The focus on construct validity expanded the range of methods for studying the validity of measurements. Cronbach (1971) and Messick (1975), for example, suggested that concurrent and content validation procedures were limited and that the most efficient way to address construct validation was through the collection of what Campbell & Fiske, (1959) and Campbell (1960) described years earlier as convergent and discriminant validity evidence--evidence which suggested that a construct was "like" some constructs that it ought to be related to and "unlike" other constructs it ought not be related to, from a theoretical perspective. Messick (1975) went on to declare that the search for rival hypotheses was the hallmark of construct validation.

Messick (1989b) later expanded the notion of construct validation by proposing a unified model for representing validity, using construct validity as the central concept. He argued that validity is, "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment" (p. 5). Messick's (1989b) construct validity-based framework included two interconnected facets to

create a broad validity concept. The framework's progressive matrix crossed a *source of justification of testing* (i.e., evidence versus consequences) facet, with a *function or outcomes of testing* (i.e., interpretation versus use) facet, and included construct validity in every matrix cell (see Messick, 1989b for further description of this framework).

Messick (1989b) argued further that, "validation is scientific inquiry into score meaning, that score-based inferences are hypotheses and that validation of such inferences is hypothesis testing " (p. 64). For Messick (1989b), validity inquiry was not simply a problem of evaluating tests, it was one of developing and evaluating scientific hypotheses. Thus, with the proposal of this validity framework, the verification of the scientific constructs underlying assessments was placed "center stage" as the explicit, centralizing force in validity inquiry (Moss, 1992) and the primary basis for judging the value of assessments.

Validity frameworks specific to performance-based assessments (e.g., Haertel, 1991; Linn, et al., 1991; Frederiksen & Collins 1989) also include specific criteria which stress the value of assessments as tools of theory. Educationally-oriented, performance-based assessments were conceptualized to better reflect current theory-grounded assumptions underlying teaching, and learning as well as associated curriculum tasks and instructional approaches. As a consequence, performance-based validity frameworks include refined criteria (e.g., representativeness, coverage) for addressing the

value of assessment content and performances, an essential piece of construct validation.

For example, Linn et al. (1991) included at least four (out of eight) validity criteria that explicitly reflected a concern for assessment content and tasks including : (1) content quality, (2) content coverage, (3) meaningfulness, and (4) cognitive complexity. While Linn et al.'s (1991) content quality criterion was similar to what has traditionally been labeled content validity, namely, that content must represent best current understanding of the field as indicated by subject matter experts, content coverage addressed process as well as content representativeness. Meaningfulness and cognitive complexity criteria, however, moved beyond traditional content validity concerns and the opinion of content experts to the evaluation of assessment tasks and of student performances in terms of curriculum and instruction. They suggested that the meaningfulness and cognitive complexity of assessments be evaluated in terms of assessment-task and student-responses analyses (e.g., how do students interpret questions?). Frederiksen & Collins' (1989) framework also included three validity criteria that explicitly reflected a concern for assessment content and tasks: (1) Scope, (2) Directness, and (3) Transparency. The scope criteria was similar to content coverage. Directness and transparency, however, moved beyond this notion. These concepts addressed explicit evaluation of curriculum-specific performances with standards of quality that were made explicit to test takers.

From the perspective of construct validity, the value of assessment is judged in terms of its ability to provide information about theory (e.g., psychological, curriculum). The focus on testing theoretically-grounded hypotheses recognizes the role of assumptions, theories, and constructs in the development and use of assessments, making explicit their role in interpretation. Valuing assessments as tools of theory also attempts to remedy the misuse and negative impact of assessments by making assessment constructs (and interpretations) clearer and more relevant to test taker and test users. Nevertheless, this focus places a premium on the "formal," theory-grounded interpretation of constructs, privileging the interpretations of test developers. While the construct validity notion helps broaden conceptualization and research on assessment validity, this work represents an expansion of the psychometric, scientifically-grounded approach to the evaluation of assessments reflected in the technical lens. The understanding and perspectives of test developers set the agenda for evaluating assessments, and the psychometric approach remains the primary method for collecting validity evidence and establishing an assessment's value. Thus, from the perspective of the theoretical lens, the individuals who actually use assessments continued to be left out of the evaluation of the assessments.

<u>Consequential lens: valuing assessments in terms of their impact on society</u>

Messick (1975) was the first to suggest evaluating tests in terms of their impact on society. He suggested that there were two questions to ponder

when considering whether a test ought to be used for a specific purpose: First, "is the test any good as a measure of the characteristic it is interpreted to assess?" Messick (1975) believed that this first question was a technical and scientific one (represented by the technical and theoretical lenses discussed in the last two sections and reflected by the bulk of validity work up to that point in time). Messick's (1975) second question was, "should the test be used for the proposed purpose?" Messick considered this second question to be an ethical one which required an evaluation of the potential consequences of testing. Cronbach (1988) supported Messick's consequential perspective in the evaluation of tests, arguing that those validating tests had an obligation to review and guard against adverse consequence of assessment practices. While neither Messick or Cronbach suggested the relative weight that each question should be given when judging the value of assessments, the recommendation that consequences be considered was a fundamental shift from issues related to the technical and theoretical aspects of assessments themselves to the specific contexts of their use.

With the introduction of this consequential perspective and a concern for the apparent negative consequences of high-stakes testing for both teaching and learning, came empirical work examining the consequences of these tests on the educational system. As a result of this work, testing has been implicated in lowering student motivation for learning (e.g., Paris, Lawton, Turner and Roth, 1991), narrowing curriculum and instruction

(Shepard, 1993), and negatively impacting the attitudes of teachers (Smith, 1991). Paris et al. (1991) in their review on the development of student self–perception concluded that, "findings revealed a cumulative, negative impact [of testing] on students that can be summarized in three general trends: growing disillusionment about tests, decreasing motivation to give genuine effort, and increasing use of inappropriate strategies" (p. 14). Smith (1991), in her qualitative study of the effects of external standardized testing on teachers, found that these tests not only narrowed curriculum offerings and time available for instruction, but resulted in an overwhelmingly negative attitude toward this form of testing on the part of teachers.

Findings from consequential validity studies like these contributed to a growing awareness of the negative impact of externally-mandated, standardized tests used for high-stakes purposes (i.e., accountability). These studies also highlighted the weaknesses of the standardized testing technology for representing and encouraging valued curricular goals and performances. The ambition to create assessments that fair better when evaluated in terms of consequential validity criteria has prompted the search for assessment alternatives (e.g., performance-based assessment). Because they reflect relevant content and meaningful tasks, performance-based assessments have been increasingly endorsed for use in education to remedy the negative impact of standardized tests (Baker, O'Neil, & Linn, 1993) and

encourage desired systemic effects on teaching and learning (Frederiksen & Collins, 1989).

Validity frameworks designed specifically for the evaluation of performance-based assessments (e.g., Linn et al., 1991; Frederiksen & Collins, 1989) include explicit criteria for addressing assessment consequences and attempt to balance both consequential and technical considerations (Moss, 1992). Because of findings which sugges the negative psychological and instructional impact of standardized tests that do not match curricula, performance-based assessment validity frameworks emphasize the selection of relevant content, and meaningful performances toward the improvement of assessment impact.

Researchers have only begun to explore the consequential validity of performance-based assessments. Stewart, Paradis, & Aegerter (1992), for example, examined the ways in which portfolio implementation empowered teachers. These researchers employed a school-level case study methodology, as they held weekly seminars with teachers to discuss portfolios and their implementation. Drawing on fieldnotes from meetings, interviews with teachers, and audiotapes of classroom instruction, these researchers explored the attitudes, understandings, and impact of portfolio assessment on teachers. While Stewart et al. (1992) examined the impact of teacher-initiated portfolios, Mosenthal, Lipson, Mekkelsen, Daniels, & Jiron (1996) explored the consequences of the large-scale Vermont Assessment Program portfolio

(writing component) mandate on the classroom instruction and assessment of fifth-grade students.

From the perspective of consequential validity, the value of assessments are judged in terms of their impact on the educational system (e.g., teachers and teaching, students and their learning). Valuing assessments in terms of their consequences attempts to remedy negative impact and encourage the positive outcomes that are the primary goals of education. Validity frameworks that include consequential validity (e.g., Linn et al., 1991; Frederiksen & Collins, 1989) have focused on curricular-assessment alignment and positive instructional impact (and technical aspects of validity like generalizability). A few studies have examined the consequences of assessment from a personalized perspective (e.g., Paris et al., 1991). Assessment developers are beginning to understand and anticipate potential consequences of assessment, realizing that the full impact of assessment requires a lengthy period of implementation and evaluation. In fact, it may not always be possible to identify the direct impact of assessments (e.g., Mosenthal et al., 1996), particularly those assessments that are used but not of much consequence (a situation which may arise when an assessment program is in place and multiple indicators are available). Finally, the consequential lens tends to represent society as a relatively passive receiver of assessment information rather than an active participant in assessment development and implementation.

The addition of the consequential lens provides a window on the value of assessments from the perspective of society. Consequential validation allows us to consider the social, value-laden aspects of assessment, addressing the role of social values in assessment and the impact of assessment on the lives of school personnel (e.g., principal, teachers), and students. Nevertheless, the social impact of assessment is only one aspect of value from the point of view of society and only one approach to building a rationale for the inclusion of tools in assessment programs. The approach to social value assumed in the present study (i.e., social validity) encourages society to actively participate in the development and implementation of assessments by considering the uses consumers make of assessment information. This approach also examines the value consumers attach to assessment information and the tools used to generate it.

Theoretical framework: Social validity

In the previous sections, I drew on two bodies of research concerning educational assessment to argue that there is a great deal of assessment taking place in our schools and we require an approach to determine the value of expanding assessment programs which is beyond the scope of current validity research. In this section, I introduce and examine the social validity construct. I describe social validity in terms of it's origin, and focus. I also explore its potential contribution to the evaluation of assessment programs.

I discovered the obscure social validity construct in the unlikely

literature of applied behavior analysis. The construct of social validity was proposed by Wolf in 1978 as a lens for examining the value of educational intervention programs in terms of their goals, procedures, and outcomes. In his seminal paper introducing the concept, Wolf (1978) made the case for what he called "subjective" measurement (e.g., measurement of opinions, feelings, beliefs), in a field priding itself on objective, behaviorally-oriented measurement. In the process of making his case, he related a story about how he, while helping to create the Journal of Applied Behavior Analysis, had committed the journal to the subjective goal of "publishing applications of the analysis of behavior to problems of *social importance*" (emphasis added; Wolf, 1978; p. 203).

Behavior analysts rejected introspective psychology and the study of theoretical constructs. They embraced the behaviorism of John Watson and B.F. Skinner and the study of observable, operationalized, and quantifiable behavior. From this perspective constructs were equivalent to objective, measurable operations (Cherryholmes, 1988). Nonetheless, Wolf (1978) in "a moment of haste" (p. 213) had committed his journal to a purpose that was clearly subjective. Wolf (1978) defended the purpose of social importance, both for his work and his journal, stating that, "behavioral analysis needs to be a responsive consumer-oriented applied social science" (p. 213) in order to achieve its goals. This purpose was embodied in the social validity construct he advanced.

Wolf (1978) introduced the construct of social validity to raise awareness among his colleagues for the need to consider the values of consumers in the design of intervention programs. He argued that greater consideration of consumer needs would increase the likelihood that program consumers would accept the intervention programs that behavior analysts developed. He also suggested methods for querying consumers about program dimensions. In the following sections I describe the components of social validity inquiry, as proposed by Wolf (1978) and other behavior analysts, including its goals, target consumers, program dimensions, and methods of data collection. I also discuss ways in which this framework informed my dissertation research.

Goals

Social validity researchers (e.g., Schwartz & Baer, 1991; Wolf, 1978) had a pragmatic motivation for conceptualizing this form of validity. These researchers were behavior analysts and developers of intervention programs. Because they believed that social validity data could be used to plan, implement and evaluate their programs in a way that would encourage consumer use, they investigated the values of potential program consumers. Schwartz and Baer (1991), for example, argued that in order for program developers to anticipate rejection of a program, it was necessary to query potential consumers about program acceptability.

Like behavior analysis, the science of assessment has also become an

applied technology as it is implemented in the context of schools (Schwandt, 1989). Assessment data that are collected as part of a school assessment program are used by a range of consumer groups to make sense of and make decisions about the progress of students and schools (Farr, 1992). Given this fact, I believed the social validity data could inform the design of widely-valued assessment tools and broader assessment programs.

Program dimensions

Wolf (1978) laid out an approach to social validity inquiry that involved the evaluation of critical program dimensions by program consumers. Dimensions which Wolf (1978) suggested were deserving of study included program goals, procedures and outcomes. Social validity research involved researchers asking consumers if program goals were important, procedures were acceptable, and if they were satisfied with program results.

While the specific dimensions of assessment programs differ from those of intervention programs, this approach provided a useful framework for identifying and exploring critical aspects of assessment programs. Assessment program dimensions that I believe are important to evaluate using social validity methods include the uses consumers made of assessment information (the goals of the program), the assessment tools used to collect and organize assessment information (procedures) and the consequences of assessment use on the educational system (outcomes).

Consumer groups

The primary strength of social validity research is that fact that it assumes that consumers, rather than developers, are the best judges of their own program needs, preferences, and satisfaction (Wolf, 1978). Social validity researchers plan and evaluate programs through the analysis of feedback elicited from program consumers. Thus, one of the challenges of social validity research is identifying relevant consumers, those individuals whose acceptance of a program is critical to its viability.

Schwartz and Baer (1991), for example, categorized consumers as direct and indirect. Despite the fact that indirect consumers were described as individuals that may be affected by a program, they are not its primary recipients (e.g., the public). Direct consumers, on the other hand, were the primary recipients of a program and their use and acceptance of the program was critical for its continued viability (e.g., students, teachers). To facilitate the ongoing use of a program, Schwartz and Baer (1991) argued that the first priority of social validity study was to understand the values of direct consumers.

Direct consumers of assessment programs are the primary recipients of assessment information such as the school administrator, teacher, parents, and students. While state and district policy makers may also receive assessment information, their active role in the selection and implementation of large-scale assessments ensures the recognition of their

values. Consequently, in the present study I was interested in the values of the direct consumers (i.e., school administrator, teacher, parents, and students) who have less power to affect change in the educational system, but whose support of an assessment program is critical to its success.

Data sources

Subjective data from interviews and written surveys are the hallmark of social validity research (Schwartz & Baer, 1991), a feature shared with qualitative research traditions (e.g., Bogden & Biklen, 1992; Strauss & Corbin, 1990). A few behavior-oriented social validity researchers (e.g., Hawkins, 1991; Winett, Moore, & Anderson, 1991), however, have advocated that more objective and verifiable forms of data be collected in place of or in conjunction with consumer opinion data. Some researchers condemn the use of subjective data (e.g., Hawkins, 1991) altogether. Other researchers (e.g., Winett et al., 1991) advocate the use of epidemiological/normative data, in addition to subjective social marketing data, as a "basis for defining verifiable importance [of program goals] and for prioritizing program problems" (p. 219). Despite debate in the behavior analysis literature, most social validity researchers have endorsed the collection of interview and survey data toward the understanding of the opinions and values of consumers.

In sum, social validity researchers advocate the use of survey and interview methods to understand the perspectives of consumers for the purpose of improving intervention programs. Another important strength

of the social validity lens is that it provides a framework for exploring the social value of programs within the contexts in which they are used. Social validity research also encourages consumer participation in program design and use. Thus, the social validity construct enabled me to take into account the specific assessment uses and needs of consumers when evaluating and (re)designing a literacy assessment program.

While the lens of social validity takes into account the value of assessments from the perspective of the consumers who use them, the construct, as defined by Wolf (1978) and others (e.g., Schwartz & Baer, 1991), is limited to the measurement of consumer satisfaction. In this sense, it might better be described as consumer validity. In the present study I expanded the social validity construct to emphasize the discourse of assessment users and their understandings of educational assessment in context (Cherryholmes, 1988). In other words, this study addressed epistemologically-distinct questions; it focused on the phenomenological perspective of assessment users rather than the viability of any given assessment tool or program. Through the application of this expanded social validity construct, I hoped to understand the values and empower the voices of assessment users who had not traditionally had a role in assessment development and validation.

Concluding Comments

I draw on the literature reviewed in this chapter for the design, construction, and interpretation of my dissertation project. The value of

assessment tools and information from the perspective of those who use them is the focus of my work. The review of literature on assessment in education allowed me to contextualize Highmeadow's literacy assessment program (which I describe in Chapter 4) in the broader historical trends of expanding educational assessment. Moreover, it helped me to identify consumers who have historically cared about and used educational assessment tools and information. This insight impacted my decision to consider the use and value perspectives of school administrator, teacher, students, and parents as the primary consumers of assessment information.

My examination of assessment validity helped me to understand that, while the lens of social consequences has begun to recognize the perspective of society in judging the value of assessments, assessment value has typically been conceptualized through the technical and theoretical lens of traditional psychometrics. In general, this insight sensitized me to the need to explore the value of assessment tools and information from the perspective of society. In particular, it led me to study the uses of assessments by school principal, teacher, students and parents with an eye toward understanding how these consumer groups value diverse tools and information that make up a school literacy assessment program. Both the literature on social validity (which provided a useful construct for conceptualizing and studying the social value of assessments), and my close examination of assessment consumer-use patterns and values served as a foundation for the evaluation

of the assessment program that I present in Chapter 5. In Chapter 6, I examine a performance-based assessment in terms of its potential for meeting the unaddressed needs of assessment consumers

Research questions

Specifically, this study addressed the following three sets of research questions:

(1) What tools made up Highmeadow's literacy assessment program, and what information was available to assessment consumers?

(2) How did consumers use available assessment tools and information, and what dimensions of assessments (and associated information) impacted how they were used and valued by assessment consumers?

(3) What assessment gaps, defined in terms of consumer reported valued dimensions, were present in Highmeadow's literacy assessment program, and what is the potential value of the performance-based for filling those assessment gaps?

CHAPTER THREE

METHOD

The development and implementation of the performance-based assessment, which was designed to provide information about students in a literature-based classroom, took place within the context of a school-wide literacy assessment program. I explored the potential role of alternative assessments in this context. Specifically, I was interested in whether the performance-based assessment would be a valued source of assessment information and should be included as a component of the established literacy assessment system.

After analyzing the literacy assessment program in terms of tools and available information, I generated categories of information use, and value dimensions associated with the use of assessment information across consumers (e.g., authority, standardization). I also identified the extent to which necessary information was not provided to consumers from the established program. Finally, I evaluated the performance-based assessment in terms of its value for addressing the information needs of assessment consumers.

School context

Because this research involved a case study (Merriam, 1988) of a school-based literacy assessment program, an understanding of the school context was critical to interpreting the research findings. The target school and

classroom were located in a large, midwestern city. It was a *School of Choice,* where those attending had requested the school, and were selected by lottery from a large set of applicants. The teaching staff's practices were innovative; they were involved in many reform efforts (e.g., school-wide, portfolio assessment implementation), and in demand by parents and students (i.e., percent of students in lottery who get to attend is low). Overall, the teaching staff and administration were highly motivated to improve instructional practices and enhance student growth.

The target school's drive for improvement and support of innovation made it a highly-appropriate site for the present study. Because the performance-based assessment that was implemented was difficult and time-consuming to put into practice, it required school commitment (Valencia, 1993). My study necessitated a setting where such commitment was part of the system. The administration and staff were committed to alternative curriculum, instruction and assessment as evidenced by the presence of alternative-practice goals in its school improvement plan. While the school was committed to innovation in assessment, its assessment program was in transition and expanding (typical of many schools today). The diverse set of assessment tools (e.g., standardized tests, classroom-based assessments) provided a fitting context for the exploration of use patterns for a variety of different forms of assessment. Due to the school's status of expanding assessment, it was an ideal candidate for an assessment program

evaluation.

Participants

The participants included one fifth grade-level teacher, June[1], her 26 students, their parents, and the school's principal, Joan. June, a 30-year-old woman, had over five years teaching experience at the time of this study, all of it at the fifth-grade level and all in the focus school and classroom. She received a Literacy Master's degree from a large, local university in August of 1993.

During the time when she was pursuing her Master's degree, she became increasingly interested in assessment issues. The alternative assessment reform effort in her school sparked initial interest. As a part of the Master's program, June was enrolled in a classroom literacy assessment course which I taught. In this class, June was required to develop a plan for implementing literacy portfolios in her classroom. Following the course, June implemented the portfolio plan (which targeted her Book Club reading program) in her classroom and presented the results at local and national teacher conferences. In addition, June volunteered to participate in a large-scale assessment project of which the present study was a part. This large-scale assessment project involved the study of June's recently initiated (i.e., one-academic year) classroom-based, portfolio assessment system and the

---

1 Pseudonyms have been assigned to both the classroom teacher and school administrator to preserve anonymity.

performance-based assessment which is the focus of the present study.

Students included 14 girls and 12 boys from a predominately white, upper-middle class, suburban community. Six focus students representing a range of literacy-ability profiles (i.e., high, average, low) as judged by June and myself, were also selected for closer study. Finally, the principal, Joan, a 43-year-old woman, was active in her professional community (e.g., presenting at many local conferences, working on a doctoral degree in educational administration) and involved in the day-to-day instructional practices of the teachers in her school building (e.g., making frequent visits to classrooms). She was motivated to provide the students at her school with a strong educational experience and actively supported teachers' efforts to improve their instruction by recognizing innovative teaching practices and professional development (e.g., encouraging teachers to present at and attend professional conferences). She also introduced new educational initiatives into her school, including a school-wide alternative assessment reform effort which contributed to June's interest in alternative assessment and the present effort to evaluate the established literacy assessment program.

Instructional context: Book Club literature-based reading program

June had been implementing a literature-based reading program called Book Club (see McMahon, Raphael, Goatley & Pardo, 1997; Raphael, Pardo, Highfield, & McMahon, 1997) in her fifth-grade classroom for two years, at the time of this study. While June's literacy curriculum included a process

writing component, Book Club served as the centerpiece of her literacy program and the target of her portfolio assessment system. June also taught social studies and attempted to integrate relevant subject matter (e.g., students read and discussed historical fiction and drew on informational texts encountered during social studies) into her Book Club instruction. The Book Club curriculum was grounded in three theoretical perspectives and revolved around four instructional components which were critical in the design of the performance-based assessment.

Theoretical grounding

The three theoretical perspectives that guided the development of the Book Club curriculum included the following: (1) a socio-cultural perspective on learning, (2) reader response literary theory emphasizing personal response and literary analysis, and (3) curricular integration emphasizing the interrelated development of language and literacy (i.e., reading, writing, listening, & speaking), each of which is described in detail below.

Socio-cultural perspective on learning. The performance-based assessment was designed to reflect the social constructivist principles (e.g., Gavelek, 1986; Wertsch, 1985; Vygotsky, 1978) on which the curriculum was grounded. From this learning and instructional perspective, knowledge is socially constructed within the context of collaborative, purposeful activities. Tasks and materials must maintain their holistic and authentic nature while providing students with multiple opportunities to demonstrate, internalize,

and transform their knowledge and understandings. Book Club instantiated these principles through activities such as having students read complete novels, and interact in the public/social domain within the context of whole-class community share and small-group book clubs.

Reader response literary theory. The Book Club curriculum embodies a reader response orientation to the reading process. This orientation emphasizes the transactional nature of reading (e.g., Rosenblatt, 1991; Langer, 1990), where the reader plays a central role in the process of constructing meaning, responding both aesthetically and efferently as their interpretations unfold. Book club instantiates these principles through the direct instruction of both text-oriented (e.g., prediction, summary) and reader-oriented (e.g., evaluation, self-in-situation) responses, while emphasizing the evolutionary, multidimensional, and intertextual nature of interpretation.

Curricular integration. The Book Club program was designed to reflect a belief in the interrelated development of language and literacy (i.e., reading, writing, listening and speaking). Because knowledge is assumed to be acquired through social interaction, and the primary means of such interaction is through language, language plays a central role in learning (Wertsch, 1985; Vygotsky, 1978). In this way, language, in both oral and written forms, becomes a tool of thought and mediates all learning. Not only do oral and written language mediate learning, they are interactive language processes which support the development of each other as they both

contribute to new forms of thought and learning (Wells & Chang-Wells, 1992). These principles are instantiated in the Book Club program through student response in multiple modes. During instruction, students read extended texts, speak and listen in large- and small-group discussions, and write in response logs. These theoretical perspectives shaped the contexts and tasks defining both Book Club instruction and the resulting performance-based assessment. The Book Club curriculum is described in the following section.

Instructional components

The Book Club curriculum includes four instructional components: (1) reading, (2) writing, (3) small-group book club discussion, and (4) community share, a whole-class setting for discussion and instruction. The hub of the literature-based reading program is the small, student-led discussion group. In these groups, students talk about topics and issues that they find interesting after reading trade books. The reading component focuses on building fluency, increasing reading vocabulary, acquiring and using comprehension strategies, and learning to recognize and understand various genres and engage in aesthetic and personal response while reading high-interest, trade books.

The writing component involves writing before, during and after reading to facilitate discussion of text, encourage students to adopt relevant stances (Bisesi, 1993; Langer, 1990) and promote the synthesis of ideas within

and across similar texts (e.g., genre, author, theme). Community share involves the teacher meeting with the class as a whole and helping the students prepare for their small-group discussions or facilitating the sharing and debating of ideas. Finally, instruction involves the teacher directly helping students to improve their journal responses and student-led discussions.

Curriculum performance dimensions

The Book Club curriculum was developed around four literacy-performance dimensions that were emphasized in instruction and targeted by the performance-based assessment. This dimensional framework includes: (1) *Language conventions* (e.g., writes conventionally, uses appropriate language choices), (2) *Comprehension* (e.g., makes predictions, clarifies understandings of text, makes intertextual connections), (3) *Response to literature* including both **personal response** (e.g., shares own experiences, puts self in situation of characters), **critical literacy** (e.g., uses evidence from text/personal experience to support ideas/opinions, asserts personal "voice"), and **creative literacy** (e.g., "what if") and (4) *Literary elements* (e.g., identifies different genres and author's craft, understands point of view).

Performance-based assessment

The performance-based assessment was developed by June and myself, in concert with a Book Club curriculum developer, and a second Book Club teacher, Sally. The performance-based assessment was created to be used by

teachers implementing the Book Club, literature-based reading program (Bisesi & Raphael, 1997). In developing the assessment, we hoped to find a compromise between formal, standardized tests that did not tap the curriculum-related goals we cared about and the informal, often difficult to interpret information derived from students' year-long portfolios. Thus, as performance-based assessment designers, we hoped to achieve the following three goals: (1) to create a valid assessment of Book Club-related literacy growth and achievement and curriculum effectiveness, (2) to provide useful information about curriculum-related literacy performance to relevant assessment consumers, and (3) to supplement/compliment information obtained from forms of assessment already being implemented.

Design and development

The performance-based assessment was developed within the context of monthly assessment group meetings (taking place from August 1993-August 1994). Early in assessment design, the group read widely on the topic of performance-based assessment. As we read, we noticed that performance-based assessment developers (e.g. Abruscato, 1993; Stiggins, 1987) suggested that these assessments consist of a standard set of activities that created the same measures of students' literacy performance and progress across contexts (standardization), a feature we believed would help us achieve our goal of evaluating curriculum effectiveness.

We also came to the conclusion that we could best achieve our goals

for the performance-based assessment by focusing on student performance of tasks and activities that were of direct interest to us, "valued in their own right" (Linn, Baker & Dunbar, 1991; p. 15). Thus, we decided that we should look to the Book Club curriculum itself to select our tasks and materials. We believed that a performance-based assessment with these features would be most likely to compliment other sources of assessment information and provide curriculum-related achievement information that might be useful to relevant assessment consumers (evaluating this particular goal was the focus of the present study).

Tasks. Like other performance-based assessments such as NAEP (National Center for Education Statistics, 1994), we structured the assessment around an integrated instructional unit. However, our assessment was designed specifically with the four Book Club instructional components in mind. The performance-based assessment was created to provide information about student performance on four instructional activities/tasks: (1) reading portions of a text, (2) responding in writing to the text that had been read, (3) participating in small-group (i.e., 4-6 students) discussions about the text, (4) sharing with the class ideas that had been discussed in small groups.

Artifacts. These four activities generated several samples of performance, called "artifacts." The primary artifacts targeted for collection during the six-day, performance-based assessment cycle included: (1)

audiotaped recordings of student oral reading, (2) written journal-entry responses, (3) audiotaped recordings of student discourse during small- and large-group discussions, and (4) student written self-evaluations of their book club performance and their journal-entry writing.

Texts. We selected three different text genres (i.e., informational text, short story, and novel) to be used as part of the performance-based assessment. These text types were chosen because they paralleled the reading tasks that students experienced within Book Club and the kinds of reading performances in which students were expected to succeed according to district and state guidelines. The informational selection represented the content-area reading that was part of their program. Trade books, usually in the form of novels, were the primary texts used during Book Club. Students read novels ranging from Hatchet (Paulsen, 1988) an adventure story, to The Upstairs Room (Reiss, 1972) a piece of historical fiction at a Jewish family during World War II. Selecting chapters from the middle of the students' novels provided a context in which they had developed some background knowledge, had worked together in their book clubs for at least a week, and were at a point of reflecting upon events in the novel. Finally, the short stories were illustrative of some of the picture books used within instructional units, such as Sadako and the Thousand Paper Cranes (Coerr, 1977).

In addition to their curricular validity, these texts provided interesting

comparisons from a research perspective. For example, we wondered if both events around the narrative texts (i.e., the short story, the novel chapters) were necessary or if similar information would be gained from each. If the latter, then the performance-based assessment might be considered as informative with simply two of the two-day events. We also wondered if students would respond differently to the informational and narrative (i.e., novel, short story) texts. On the state-mandated reading test (i.e., Michigan Educational Assessment Program), students had experienced much greater difficulty with informational texts than narratives, and this had become a concern among the administration and teaching staff at Highmeadow.

Pilot Administration. During the pilot study, we collected artifacts for the four tasks including audiotaped recordings of both oral reading and book club discussions, written-journal entries, and written self-evaluations. Written-journal entries were collected from each student daily, since we felt their ability to express their personal response to literature was a critical goal for Book Club and collecting such samples was not difficult. Because of a limited amount of audio-taping equipment, we taped each book club once per two-day cycle, taping half the book clubs on the first, and the other half on the second day.

The performance-based assessment included students' activities and products (e.g., journal-entry samples, discussion recordings, oral reading samples) from three standard, two-day Book Club "events." One event

focused on an informational article, the second was based upon two middle chapters of the novel students read as part of their Book Club program, and the third used a short story. All texts related to the unit theme within the classroom (i.e., World War II). All participating students read the selections, created a written-journal entry, engaged in a book-club discussion, and participated in a whole-class community share which standardized the activities. The resulting artifacts served as a basis for analysis of strategy use and literacy performance.

Developing scoring criteria. Working closely with June and Sally, we began by considering the goals of the performance-based assessment, emphasizing that we were most interested in students' oral and written response to the texts they read. Thus, our scoring efforts concentrated on the students' written-journal entries and their book club discussions. In designing scoring rubrics, we consciously decided to use a 3-point, rather than 5-point scale, since the latter was associated with typical grading patterns (e.g., A, B, C, D, F). Thus, our journal-entry and discussion scoring rubrics consisted of three levels of performance each. We also decided to use a holistic rating scale that covered several "dimensions" or "criteria," since others (e.g., Freedman, 1979, 1993) have found that holistic scores reflect how well students develop and organize ideas while taking an entire artifact into account.

To define each performance level, or interpretative category (Moss,

1996), we drew on the curriculum-performance framework dimensions in a deliberate attempt to match instructional and assessment goals and provide correction for the malalignment problem evident with other forms of assessment (see Bisesi & Raphael, 1997). Specific scoring criteria defining each level of performance were selected to help us distinguish among students' performances and with sensitivity to both informational- and narrative-text responses (Bisesi, 1996).

For example, the highest level for a written-journal response, a "3," was assigned to student entries that focused on major themes, include evidence from the text to support their position, explored different responses invited by the text and linked them together in relevant ways, had an apparent purpose for their writing, had a focused and coherent response, and had a date on the entry. While a "3" response may not have addressed all these criteria equally well, together they provide an image of what a level 3 response should have. In contrast, a level "1" response was superficial, including little reference to the text, and no clear purpose. These responses were often limited to a string of trivial details with a lack of coherence. Thus, our rubrics had performance levels with explicit criteria that lead to a score. Table 2 details the performance criteria for both journal-entry and book-club discussion rubrics.

Table 2-Scoring Rubrics for Journal Entries & Book Club Discussions

Scores	Journal Entries	Book Club Discussions
3	•Focuses on major themes, issues, questions or characters. •Effectively uses evidence from text and/or personal experience to support ideas •Produces multiple, related & well-developed responses •Writes for a clear purpose •Generates a well-focused, connected and coherent response •Dates entry	•Focuses on major themes, issues, questions or characters. •Effectively uses evidence from text, content area and/or personal experience to support ideas •Appropriately introduces new ideas •Builds/expands on others ideas •Respects others ideas •Talks for a clear purpose •Appropriately supports less active members of the group
2	•Focuses on secondary themes, issues, questions or characters OR lacks detailed discussion of major themes. •Uses little evidence from text and/or personal experience to support ideas OR use of evidence is less than effective •Demonstrates some sense of purpose for writing •Generates a somewhat focused, connected and coherent response	•Focuses on secondary themes, issues, questions or characters OR lacks detailed discussion of major themes. •Uses little evidence from text and/or personal experience to support ideas OR use of evidence is less than effective •Demonstrates some sense of purpose for speaking •Builds some on others ideas but may resort to round robin turn taking •Demonstrates some respect for others ideas •Less than effective at introducing new ideas
1	•Superficial response with minimal reference to the text or personal experiences •A string of trivial textual details •Demonstrates no clear purposes for writing •Generates an unfocused, unconnected and incoherent response •Does not date entry	•Superficial response with minimal reference to the text or personal experiences •Talks about trivial textual details or irrelevant personal experiences •Perseverates on ideas--does not build on them •Does not introduce new ideas •Demonstrates no clear purposes for speaking •Speaks very infrequently •Raises hand before speaking and/or resorts to round robin turn taking

The rubrics and scoring system provided a means for evaluating students' performances in Book-Club related response activities using a standard metric (see APPENDIX A for sample journal entries scored at the three levels). Interrater and intrarater (over a one-year interval) agreements were found to exceed 85% for journal entries. Intrarater agreement was 87% for discussions.

Data sources and collection procedures

Data sources and collection procedures were consistent with social validity methodology (e.g., Wolf, 1978) and phenomenologically-oriented, qualitative research approaches (e.g., Bogden & Biklen, 1992; Cherryholmes, 1988). Primary data sources for this study included: (1) written-survey responses from the school principal, the classroom teacher, her students and their parents collected in the fall of 1994, (2) written-survey responses from the students and their parents collected in the spring of 1995, (3) transcripts of interviews with six focus students collected in the fall of 1994, and 4) transcripts of interviews with the school principal, classroom teacher, and the six focus students collected in the spring of 1995.

Supporting data sources included: (1) performance-based assessment artifacts and scores for both fall 1994 and spring 1995 administrations; (2) weekly fieldnotes documenting school-based activities including instructional practices, student performance, performance-based assessment administrations, and parent-teacher conferences; (3) classroom-based portfolio

assessment artifacts from six focus students collected twice, during the fall of 1994 and spring of 1995; and (4) other assessment-related tools and documents (e.g., testing manuals, testing schedules, checklists, newsletters). Collection of these data took place throughout the 1994-95 academic year within the timeframe detailed in Table 3.

Table 3-Timeline for data collection

TIMELINE	DATA COLLECTED
Mid-September 1994	•Collected fall-survey information from the school principal, the classroom teacher, the students, and their parents. •Conducted fall performance-based assessment
October - November 1994	•Conducted fall interviews with and collected examples of classroom-based portfolio artifacts from six focus students •Conducted monthly classroom observations •Conducted parent-teacher conference observations
January -March 1995	•Conducted spring interviews with and collected examples of classroom-based portfolio artifacts from six focus students •Conducted interviews with principal and teacher •Conducted monthly classroom observations •Conducted parent-teacher conference observations
May 1995	•Collected spring surveys from students, and parents. •Conducted spring performance-based assessment

Surveys

I designed the Fall 1994 surveys to tap assessment consumers' knowledge and attitudes about literacy and the literacy assessment program at the target school. The Spring 1995 surveys were designed to tap students' and parents' attitudes toward the information they received from the performance-based assessment. Both surveys included a combination of

limited-response (i.e., yes-no) and open-ended questions (see Appendices B-G to review survey questions) as suggested by Wolf (1978), to provide respondents with direction in response while offering the greatest latitude to qualify their answers. Survey response rate was 100% for both students and parents in fall and spring. I collected surveys from June and Joan in the fall only.

Interviews

In early spring, I interviewed June about her goals for her students in terms of their literacy development, her instructional focus, her beliefs about literacy instruction, her literacy assessment uses and needs, and her attitude toward the performance-based assessment. I also interviewed Joan at that time about her assessment uses and needs, and her attitude toward the performance-based assessment. Finally, I conducted interviews with six focus students in both fall and spring. The fall, student-interview protocol included questions regarding their knowledge about and attitude toward literacy and literacy assessment. The spring, student-interview protocol included questions about their understanding and attitude toward the performance-based assessment. I designed interview questions to parallel those making up the surveys to provide comparable data from multiple sources. Interview protocols are included in Appendices I-L for review. All interviews were tape recorded and professionally transcribed. I also edited all interview transcripts.

Observations

Throughout the 1994-1995 school year, I conducted weekly classroom observations of literacy instruction periods and documented my observations in the form of written fieldnotes. My fieldnotes included documentation of instructional practices, and student learning, as well as teacher and student uses of assessment information. I also observed parent-teacher conferences and recorded assessment information use patterns in this context.

Classroom-based portfolio artifacts

June collected and evaluated artifacts for all students as part of her portfolio assessment over the course of the 1994-1995 school year. A sweep (Valencia, 1993) of portfolio contents was made in both the fall and spring for the six focus students. Collected artifacts were photocopied and originals returned to the classroom portfolio. A detailed description of the artifacts collected is included in Chapter 4 in the section on the portfolio assessment system.

Performance-based assessment

Performance-based assessment data collection took place across two, four-day administrations during the fall of 1994 and spring of 1995. During the fall event, students participated in a unit on Canada, reading the novel Hatchet (Paulsen, 1988) about a boy who survives a plane crash in the Canadian wilderness, and two informational articles on pollution policy between the United States and Canada (Sizemore, 1988; Gloucester Press,

1987).

During the spring event, students engaged in a unit on World War II, reading two informational articles, one a chapter (i.e., "Aggression on the March") from the textbook, The Day Pearl Harbor was Bombed: A Photo History of World War II (Sullivan, 1993), and the other written by June for the purpose of this assessment. The spring event also included chapters from one of the following two novels: Devil's Arithmetic (Yolen, 1990), or Number the Stars (Lowry, 1989).[2]

After reading the selection for the day, students spent 10-15 minutes writing in their response journals, prior to participating in their small-group discussions. Because June had students respond in their journals with and without prompts during instruction, we collected journal entries under both conditions. The use of teacher prompts was counterbalanced so that each student responded to a prompt (e.g., "What trends or ideas do you notice that all three axis powers display?") on one of the two days of each event, with open response on the other day. Counterbalancing was designed to determine if teacher prompting used as part of Book Club instruction resulted in better student performance. I audiotaped student-led discussions on one day for each text genre during each four-day cycle.

Data analysis procedures

[2] I did not implement a short story text because it had not provided any additional insight into student response when included as part of the pilot administration.

My research addressed the social validity (Wolf, 1978) of the literacy assessment program and the performance-based assessment. The questions I raised for study concerned the uses assessment consumers made of assessment information, the value they attached to particular dimensions and properties of assessments, the gaps they perceived in the established literacy assessment program, and the value of the performance-based assessment for addressing their assessment needs.

My approach to data analysis was based on that suggested by Glaser and Strauss (1967) and Strauss and Corbin (1990) for the generation of grounded theory. Through the application of the constant comparative method of analysis (Glaser and Strauss, 1967), I engaged in continuous coding and sorting of my data to classify assessment-consumer uses, identify assessment dimensions and properties that consumers valued, and generate an integrated theory of assessment-consumer value. I then used this framework to identify gaps in information available from the established assessment program (i.e., needs). Finally, I explored the value of the performance-based assessment for addressing assessment information needs.

Assessment program tools and information

To answer the question regarding the literacy assessment program's tools and information, I read and reread interview and fall-survey responses to generate a list of assessment tools that consumers stated were administered to or collected from students. I also characterized available assessment

information (e.g., frequency, form). I conducted this analysis to determine if consumer groups failing to use information did so because it was not available or because they did not find it valuable for their desired uses. I triangulated these findings with my direct observations of assessment-tool administration over the course of the year documented in fieldnotes, and with the published school-testing schedule. I also looked at the tools themselves (e.g., standardized test booklets, classroom assignments) and supporting documentation (e.g., test manuals, descriptions of classroom assignments, parent-teacher conference interactions documented in observational fieldnotes) to better understand each tool and associated information.

Assessment uses and dimensions of value

Through further comparative coding and analysis of interview and survey data across consumer groups, I identified patterns of assessment-consumer information use, namely, how each group of assessment consumers stated that they used available assessment information. Again, I triangulated these data with my own observations of assessment-information use (e.g., teacher sharing results of assessment with parents at conference time) documented in fieldnotes. I then generated properties and dimensions associated with consumer use and valuing of assessment information.

The value of the performance-based assessment

To address the research question about performance-based assessment

information value, I analyzed survey and interview data to identify the properties of consumer-stated assessment needs. I then examined the performance-based assessment in terms of its potential for addressing these needs, by evaluating the properties of the assessment in terms of the value statements of the consumers.

In the following analysis chapters, 4 though 6, I describe Highmeadow's literacy assessment program, analyze patterns of assessment use both across and within consumer groups identifying the dimensions and properties of assessment tools valued by assessment consumers, and explore the value of a performance-based assessment in terms of its potential for meeting the assessment needs of consumers.

CHAPTER FOUR

THE ASSESSMENT PROGRAM

Educational assessment programs consist of various tools that are implemented to collect information about student performance. This information is then reported to interested assessment consumers. To understand the value that assessment consumers, including Joan (school principal), June (5th-grade teacher), her students, and their parents, attributed to information received from Highmeadow's literacy assessment program, I first identified the assessment tools constituting the program. I then characterized the information that was available to consumers from these assessments. Thus, I organized this chapter around the following two research questions: (1) What tools made up the literacy assessment program? and (2) What information was available to assessment consumers?

To answer these questions, I read and reread interview and fall-survey responses to generate a list of the assessment tools that were regularly administered to or collected from students. I also characterized the assessment information (e.g., individual scores, group scores, narrative description) from each tool that were available to each group of assessment consumers. I triangulated findings across consumer group data, with my own observations of assessment-tool administration over the course of the year documented in fieldnotes, and with the published school-testing schedule.

I conducted these analyses to provide a context for understanding

consumer-assessment use and to highlight Highmeadow's expanding assessment program as reflective of the trend in many schools. These analyses also provided insight on assessment-information availability, allowing me to identify factors impacting the use of assessment information (e.g., Did consumer groups fail to use information because it was not useful or because it was not available?) which I describe in Chapter 5.

In addition, I looked at the tools themselves (e.g., standardized test booklets, classroom assignments) and any supporting documentation (e.g., test manuals, descriptions of classroom assignments, parent-teacher conference interactions documented in observational fieldnotes), to better understand the characteristics of each tool. This analysis also helped me to determine the kinds of supplementary documentation and explanation that were available to each assessment-consumer group. Thus, in this chapter, I describe the literacy assessment program including the assessment tools regularly administered to and collected from students and the characteristics of resulting information made available to each group of assessment consumers (e.g., school administrator, classroom teacher, students, parents).

Highmeadow's literacy assessment program

Highmeadow's literacy assessment program was made up of a diverse set of externally-mandated (from outside the classroom) standardized tests, and classroom-based assessment tools and artifacts (e.g., portfolio). While assessment information was made available to each consumer group, the

source (i.e., assessment tools) and character (e.g., test scores, narrative descriptions) of information differed across groups. In delineating the assessment program, I first describe the assessment tools that defined the program and the approximate dates of their initial implementation. I then outline the characteristics of the assessment information available to each group of assessment consumers over the course of the 1994-95 school year.

Assessment tools and artifacts

Figure 3 illustrates Highmeadow's literacy assessment program within the contexts of history and the educational system. The ten assessment tools defining the program are listed inside the four circles. The approximate dates when assessment tools were first implemented are listed on the right. Each circle represents the level of the educational system on which the tool's implementation was mandated: State, district, school, and classroom levels.

It is interesting to note that the bulk of the current program has evolved since 1990. The only tool implemented earlier was the state-mandated reading Michigan Educational Assessment Program test. The early 1990s initiated a period of curriculum revision for Highmeadow's district which explains the fact that several new assessment tools were added to the program on the district level at that time. For example, a commercial basal reading test (e.g., Silver, Burdett, & Ginn, 1993), the Comprehensive Test of Basic Skills (CTB/McGraw-Hill, 1989) which is an achievement test, and the Cognitive Abilities Test (Thorndike & Hagen, 1986) which is an aptitude test,

were all implemented by the district around 1990. Revision of the district-wide report cards also began about that time.

On the school level, the Botel Reading Inventory (Botel, 1970) was implemented to help identify students for the at-risk program that was put into place during curricular restructuring. And finally at the classroom level, 1990 was when June started teaching fifth-grade at Highmeadow, evaluating students and holding parent-teacher conferences. As Figure 3 clearly illustrates, different types of assessment tools were mandated simultaneously on multiple levels of the system with little consideration or evaluation of the program as a whole.

Figure 3 also demonstrates evidence of the historical trend in educational assessment toward a dual-system assessment program. The literacy assessment program at Highmeadow included both externally-mandated (outside the classroom) standardized tests/assessment tools administered the same way to all students, and curriculum-specific assessment taking place in classrooms. The standardized assessment system included seven state-, district- and/or school-mandated standardized tests or assessment tools. The classroom-based assessment system consisted of three classroom-oriented, teacher implemented assessment tools/artifacts.

Table 4 provides an overview of the tools and artifacts constituting Highmeadow's literacy assessment program including a brief description of each tool or artifact, and a schedule of administration and/or collection.

Figure 3-Highmeadow's literacy assessment program

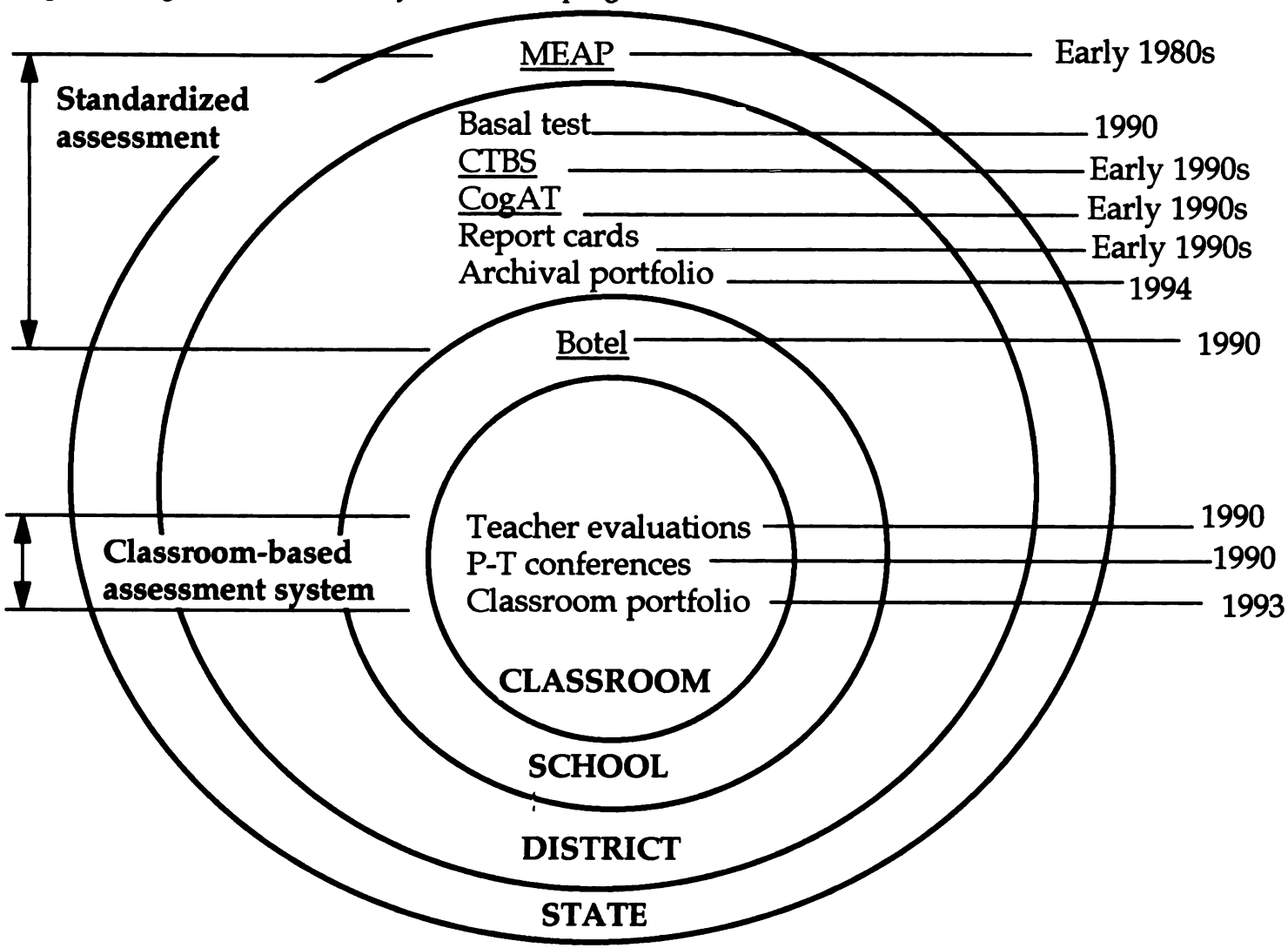


Table 4-Overview of Highmeadow's literacy assessment tools & artifacts

Assessment tool	Description	Collection schedule
Standardized assessment system		
Michigan Educational Assessment Program (MEAP)	Standardized, criterion-referenced, multiple-choice test of reading achievement	Fall of 4th-grade
Botel Reading Inventory	Standardized, multiple-choice tests of word recognition and word opposites	Fall 1st through 5th-grade
Comprehensive Test of Basic Skills (CTBS)	Standardized, norm-referenced, 4-choice multiple response item test of basic skills in reading, language, math, & study skills	Spring 3rd & 5th-grade
Cognitive Abilities Test (CogAT)	Standardized, norm-referenced, 3- or 5-choice multiple response item test of ability to work with verbal, quantitative, & geometric symbols	Spring 3rd & 5th-grade
Basal Test	Standardized, criterion-referenced, 4-choice, multiple-response item test of reading vocabulary & comprehension	Spring 2nd through 5th-grade
Reading/writing archival portfolio	Standardized set of literature response & writing process artifacts	Spring K through 5th grade
Report Card	Standardized, 5-point rating scale covering all areas of the school curriculum including the literacy-related areas of listening, speaking, reading, & writing	Quarterly 2nd through 5th grade
Classroom-based assessment system		
Parent- Teacher Conference	Teacher interpretations of student work, performance, and progress communicated to parents in the form of face-to-face interactions	Quarterly 2nd through 5th grade
Portfolios & other classroom /homework artifacts	Parent, student, &/or teacher interpretations of classroom artifacts (e.g., reading portfolio, process writing folder, Thursday folder) or homework performance (e.g., reading a chapter from a tradebook).	Ongoing in June's classroom
Teacher evaluations & other informal communication	Teacher interpretations of classroom activity & student performance recorded in written notes (e.g., "Let's communicate"), checklists, and newsletters. Phone calls and other informal, face-to-face communications also took place.	Ongoing in June's classroom

Standardized assessment system. Seven state-, districts-and/or school-mandated, standardized (administered and scored under the same conditions across students or evaluated on a common set of performance artifacts or outcomes) tests listed in Table 4 made up the literacy, standardized-testing system at Highmeadow. An examination of Table 4 suggests that the target fifth-grade classroom provided a choice context for this case study because six of the seven standardized assessments constituting the system were administered to or collected from Highmeadow fifth-grade students across the 1995-96 school year. Thus, targeting a fifth-grade classroom at Highmeadow maximized my ability to explore the use of standardized test information by assessment consumers.

The standardized assessment system at Highmeadow included five standardized tests: (1) the reading Michigan Educational Assessment Program (MEAP), (2) the Botel Reading Inventory (Botel, 1970), (3) the Comprehensive Tests of Basic Skills-Fourth Edition (CTB/McGraw-Hill, 1989), (4) the Cognitive Abilities Test-Form 4 (Thorndike & Hagen, 1986), (5) a basal test. It also included a reading/writing archival portfolio, and a school-wide report card.

The MEAP is a state-mandated assessment program. The reading MEAP was designed to assess students' ability to construct meaning during reading from story (34 items) and informational (34 items) texts. Students are assigned criterion-referenced scores (ranging from 0-350) and placed into one

of three achievement categories: (a) LOW--below 299 on both informational and story genres, (b) MODERATE--below 299 in either informational and story genres and above 299 in the other, (c) SATISFACTORY--above 299 in both informational and story genres.

The reading MEAP was initially developed and implemented during the early 1980s and is currently administered to all students in the public schools of Michigan during the fall of their fourth (seventh & tenth) grade year(s). It was originally designed to reflect the "new definition" of reading established by the state and provide criterion-referenced information to teachers for the purpose of improving their instruction. Though not its original intent, information from the MEAP is now published in newspapers across the state, providing statewide accountability information to policy makers, and the public.

The Botel Reading Inventory (Botel, 1970) consists of four tests including the word recognition and word opposites tests. The word recognition test assesses oral reading fluency using eight, graded (i.e., PP, P 1^2, 2^1, 2^2, 3^1, 3^2, 4+), 20-word lists. The word opposites test uses 10-word lists representing samples of reading material from 10 levels (i.e., 1, 2^1, 2^2, 3^1, 3^2, 4, 5, 6, 7-8, 9-12) and provides an estimate of reading comprehension. The school-wide (2nd through 5th grade) administration of this inventory during the fall of each year was mandated by school administrators in 1990. The Botel was implemented to provide school administrators with annual

information on student literacy performance for screening purposes. The two mandated tests were administered to the target fifth-grade students during the fall of 1994-95 school year.

The Comprehensive Tests of Basic Skills-Fourth Edition (CTB/McGraw-Hill, 1989) is a standardized achievement test which was created to test basic reading, math, and study skills common to many elementary school curricula. Literacy-related subtests administered to both third- and fifth-grade students during the spring of each school year included: (a) Reading Vocabulary (40 items), and Reading Comprehension (50 items) which were combined to come up with a Total Reading score (raw score range 0-90); (b) and Language Mechanics (36 items), and Language Expression (48 items) which were combined for a Total Language score (raw score range 0-84). Raw test scores, proportion correct, percentile ranks, and standard age scores (Mean=100) are made available by the district to school administrators and classroom teachers. The administration of this battery was mandated by the district in the early 1990s to provide information on program effectiveness and demonstrate accountability. The battery was administered to the target fifth-grade students during March of the 1994-95 school year.

The Cognitive Abilities Test-Form 4 (Thorndike & Hagen, 1986) is a standardized aptitude test designed to identify both the level and pattern of students' abilities to work with three basic types of symbols: verbal, quantitative, and geometric. The three subtests of the battery are designed to

assess general, cognitive ability and style in order to predict future school success (it is a surrogate IQ test). The verbal subtests targeted literacy-related abilities including verbal categories, analogies, and vocabulary. These subtests consisted of 75, five-choice multiple response items. Raw test scores, proportion correct, percentile ranks, and grade equivalent scores are made available by the district. The administration of CogAT to third- and fifth-grade students was mandated by the district in the early 1990s for the purpose of student screening. The CogAT was administered to the target fifth-grade students during early March of the 1994-95 school year.

The administration of basal reading test to all students in the district was mandated by the district in 1990. At that time, the district's reading curriculum was being revised to better reflect Michigan's core curriculum and the district required information on the reading performance of all students in the district as a means for evaluating curriculum effectiveness. The criterion-referenced basal test was given to all students in the district in the spring of each school year. The specific test administered to the participating fifth-graders in the spring of 1995, Dream Chasers Skill Progress Test, was an end-of-the-fifth, grade-level test selected from those published as part of a basal reading program (i.e., Silver, Burdett & Ginn, 1993). The test consisted of 10, four-choice items addressing reading vocabulary and 25, four-choice items addressing reading comprehension. Raw test scores and proportion correct were made available by the district to the school administrator and

teachers.

The reading/writing archival portfolio was mandated by the district in 1994 as a tool for curriculum monitoring and student screening. The portfolio contents were standardized, requiring teachers to collect two representative samples of student literature response and process writing at the end of each year. Artifact choices were standardized by grade level. Fifth-grade teachers were asked to select and collect two of the following artifacts for each student: (a) journals/logs, (b) short fiction, (c) compositions, (d) letters, (e) poetry, (f) reports, and (g) summaries. Portfolio artifacts were housed in Joan's office and made available to teachers and parents.

The school-wide report card had been under revision for several years. The current report card had been implemented for a two-year period and consisted of a list of target performance outcomes in the areas of attitudes & behaviors, communicative skills, mathematics, science, health, and social studies. The literacy-related portion of the report card included a list of 15 communication skills in the four areas of **listening**, (i.e., demonstrates appropriate listening behaviors, demonstrates auditory comprehension), **speaking** (i.e., communicates ideas effectively), **reading** (i.e., understands vocabulary, applies a variety of reading strategies, constructs meaning from literature, constructs meaning from informational text, identifies various literary forms, identifies story elements, reads fluently, participates in independent reading) and **writing** (i.e., applies the writing process, expresses

ideas in a variety of literary forms, writes legibly, spells accurately).

Students were evaluated by the teacher on these 15 skills at four points during the school year, twice each semester. Teachers graded students using a five-point rating scale (i.e., 5=performing beyond grade level, 4= high achievement at grade level, 3=demonstrating grade level skills, 2=developing grade level skills, 1=performing below grade level). While this scale guided teacher evaluations of students' curricular-related progress, it did not offer explicit criteria for "grade level" skills and performance. Thus, it was left up to individual teachers to interpret and evaluate student performance.

Classroom-based literacy assessment system. The literacy assessment program also included a classroom-based assessment system. Table 4 lists the three types of literacy assessments making up this system. Because June was an innovative teacher experimenting with new literacy curricula (e.g., Book Club) and a variety of classroom-based alternative assessments (e.g., portfolios), this classroom provided an ideal context for exploring the use and value of classroom-based assessment information by assessment consumers. The classroom-based literacy assessment system included three types of assessment tools and artifacts implemented over the course of the 1994-95 school year including: (1) parent-teacher conferences, (2) student-generated artifacts, and (3) written teacher evaluations.

All teachers at Highmeadow, including June, conducted conferences with parents four times during the school year when report cards were

completed. Conferences consisted of 10-minute, face-to-face interactions between June and her students' parents. During each conference with a parent, June shared a narrative description of the child's progress in language arts (and social studies) including classroom artifacts, and her interpretations of the child's work in the form of checklists and written feedback. June also discussed marks on the child's report card.

June reported having students collect various classroom artifacts in a Book Club portfolio. This portfolio included artifacts and tools reflecting all aspects of the Book Club reading curriculum. Artifacts included daily journal entries, and biweekly discussion and journal self-evaluations. Evaluative tools included in the portfolio were weekly written feedback to students, and biweekly discussion and journal-entry checklists. Additional reading artifacts/tools included twice-yearly, transcribed, oral-reading samples with miscue analysis. Think alouds and reading-strategy surveys were collected to monitor the students having difficulty reading.[3]

June also had students keep process-writing folders where they collected writing drafts. "Have-a-go" spelling lists (Routman, 1991) were also placed in the folders. These lists included misspelled words drawn directly from process-writing drafts on which students took tests and saved for later use as a spelling-dictionary tool. The "Thursday folder" included classroom

---

[3]The fifth-grade students in June's classroom were all average or above average achievers. While one or two students in her classroom demonstrated more difficulty reading than the others, they were not classified as special-needs students.

artifacts and homework assignments that went home weekly to parents. Artifacts in the Thursday folder were sometimes accompanied by teacher evaluations (e.g., journal checklist, written feedback), but not universally.

June reported to students and parents in the form of written evaluations and other informal communications. June wrote biweekly progress notes to her students or their parents. Other communications took the form of monthly newsletters to parents that highlighted instructional activities taking place in the classroom during the month. Finally, phone calls to parents and other face-to-face interactions between June, her students' parents, and Joan (i.e., collaborative consultations) took place.

In summary, this analysis illustrates the dual systems of externally-mandated standardized assessments, and classroom-based assessments making up the literacy assessment program at Highmeadow. This analysis also provides a sense of the literacy program's expanding character over the course of its history and the need to determine the role that constituent assessment tools are serving in the lives of assessment consumers. Thus, in the next section I identify the assessment information available to each group of assessment consumers to supply a context for understanding information-use patterns and dimensions of value discussed in Chapter 5.

Literacy assessment information available to consumers

Highmeadow's literacy assessment program included a diverse set of tools which resulted in a variety of available information. As suggested by

Table 5, assessment information was available to all consumer groups. Nevertheless, as indicated by the blackened cells of the table, information from all assessment sources (i.e., assessment tools) was not available to all consumer groups. Furthermore, the form of available assessment information (e.g., test scores, narrative descriptions), as suggested by the descriptions of information in the table cells, differed across consumer groups. In this section I describe the sources and forms of assessment information available to each of the four groups of assessment consumers over the course of the 1994-95 school year.

Table 5-Assessment information available to each consumer group

Tools	Administrator	Teacher	Parents	Students
MEAP	Test scores	Test scores	Test scores	Test scores
BOTEL	Test scores			
CTBS	Test scores	Test scores	Test scores	
CogAT	Test scores	Test scores	Test scores	
Basal Test	Test scores	Test scores		
Reading/writing archival portfolio	Written artifacts	Written artifacts	Written artifacts	
Report Card	Individual grade reports for students in school	Individual grade reports for students in classroom	Child's grade report	Own grade report
Parent-Teacher Conference		Face-to-face conversation	Face-to-face conversation	
Portfolios & other classroom /homework artifacts or behaviors		Performance artifacts (with teacher or student interpretation)	Performance artifacts (with teacher, student or parent interpretation)	Performance artifact (with teacher, self, or parent interpretation)
Teacher written evaluations & informal communications	Written notes or face-to-face conversation	Written notes or face-to-face conversation	Written notes or face-to-face conversation	Written notes or face-to-face conversation

School administrator. Joan reported receiving assessment information from tools that were part of the standardized assessment system. While classroom-based assessment information could have been obtained directly from June, it was not readily available for Joan's use. Most standardized assessment information available to Joan took the form of test scores. Information from the MEAP, CTBS, CogAT, and basal test was made available to Joan as score reports. These reports included desegregated test scores (e.g., individual student scores) as well as aggregated scores on the level of the school, and other subgroups (e.g., grade-level, classroom, gender, ethnicity), as suggested by Joan's comments made during her interview:

> "Well, other than individual students...[I] look at the total, the total so [I] can see...we're movin' up. And then there's that whole equity issue, you know, some groups within. But you can easily look at their scores if they're all listed out like that, to see how are females compared to males. How are different ethnic or gender groups... And the desegregation of data... Sometimes you can't find a pattern in the general population but you can find it in the subgroups of the population."

Norm-referenced scores (e.g., percentile rank, grade equivalent/standard age score) were available to Joan from the CTBS, and CogAT, while the basal test and the reading MEAP provided norm-referenced (e.g., scale scores) and criterion-referenced information (e.g., raw score, proportion correct,

achievement categories). Finally, Joan had information from other standardized assessment tools including individual-students ratings (determined by the classroom teacher) from quarterly report cards and archival portfolio artifacts (selected by the classroom teacher) spanning several years.

Classroom teacher. June, as classroom teacher, had the broadest range of assessment information available to her. June reported having access to individual scores from all standardized tests (except the Botel) administered to students in her classroom. June, like Joan, reported receiving these scores in the form of written reports. These reports included desegregated test scores (e.g., individual-student scores) as well as aggregated scores for her classroom. June also reported filling out language arts report cards quarterly and collecting archival portfolio artifacts at the end of each year.

Additionally, June reported having information available from a classroom-based literacy assessment system that she had implemented. This system included a Book Club reading portfolio as well as a process-writing folder. June had students collect process-writing drafts in their writing folder along with "have-a-go" spelling lists (Routman, 1991) generated from misspelled words in process-writing pieces, and process-writing feedback checklists received during biweekly teacher-student conferences.

The Book Club reading portfolios consisted of the artifacts and evaluation tools listed in Table 6.

Table 6-June's reading portfolio

Portfolio Artifacts	Portfolio evaluation tools	Schedule of evaluation
Daily Journal entries	Narrative evaluation/ feedback	Weekly
	Journal checklist	Biweekly
Daily Discussions	Narrative evaluation/feedback based on anecdotal notes	Weekly
	Discussion checklist	Biweekly
Biweekly self-assessments	Narrative feedback	Biweekly

June evaluated Book Club related student artifacts (including journal entries and book club discussion) every one or two weeks using checklists she created and anecdotal notes she recorded during classroom observations. Students generated a written evaluation of their own performance on a biweekly basis which June also collected. Finally, June collected two additional classroom-based literacy assessment artifacts including oral-reading samples on all students twice a year (fall and spring), and think alouds from at-risk students in her class.

Parents. Parents also had a range of literacy-assessment information available to them from both the standardized assessment and classroom-based assessment system. Joan and June both reported that individual scores on all standardized tests were recorded in individual students' CA-60s (i.e., a cumulative record file located in Joan's office) and were available to students' parents upon request. June also reported sharing individual-student test scores (i.e., percent correct and percentile rank) from the CTBS and CogAT

with parents during spring parent-teacher conferences (a claim supported by fieldnotes data).

All parents reported receiving MEAP scores. Joan indicated that reports including individual-student MEAP scores and achievement categories (e.g., low, satisfactory) were sent home to parents from the school with third-quarter report cards. Seven parents (out of 26) also reported obtaining MEAP-score information from public documents including the local newspaper (e.g., "MEAP scores are generally also listed in the local newspapers for all nearby districts and a sampling of others statewide") and school reports/newsletters (e.g., "Annual School report I received in the mail"). MEAP information from public documents such as these took the form of aggregated, school-level scores as suggested by parent report (e.g., "No information is received on classmates individually.") and review of public documents (e.g., school report, newspaper). Finally, parents reported receiving student report cards on a quarterly basis.

Parents also reported receiving information directly from classroom-based assessments on an ongoing basis, as suggested by Table 7.

Table 7-Parent reported classroom-based information sources and schedule

Classroom-based assessment information sources	Frequency of receiving information
Parent-teacher conference	Quarterly
Student performance artifacts (e.g., Thursday folder)	Weekly-quarterly
Teacher evaluation and communications	Periodically-bimonthly

Table 7 lists the classroom-based information that parents reported having available and the schedule of availability. While student-performance artifacts (e.g., journal entries) were sent home weekly via the "Thursday folder" to parents for review, they were also made available (along with report cards) to parents as a basis for discussion during parent-teacher conferences. Parents also reported obtaining classroom-based assessment information by talking with their children about artifacts completed in the classroom and homework assignments (e.g., reading a chapter from a tradebook). Finally, parents received classroom-based assessment information from teacher evaluations and communications including bimonthly newsletters, monthly written-progress notes sent home and periodic phone calls made to parents.

Students. Students had the narrowest range of assessment information available to them. They received scores from only one of the standardized tests (i.e., MEAP). Student also had access to report cards, and information from a variety of classroom-based assessment artifacts and tools. While most students (21/26) reported receiving individual scores through the mail (e.g. "A paper was sent in the mail with your grades," "I got a certificate in the mail," "You get a letter in the mail if your score was high."), five students reported receiving their scores directly from their parents (e.g., "My parents told me"). Only one student, however, reported having access to school-level MEAP scores ("The newspaper tells how the whole school did").

In additional, students reported receiving a range of information from classroom-based assessments. This information included student-generated artifacts (e.g., journal entries, discussion transcripts) and teacher evaluation of those artifacts (e.g., checklists, written narrative feedback, conferences). Students also reported obtaining information through self- and parent-evaluation of classroom artifacts.

Summary

Although assessment information was made available to each consumer group, the source (i.e., assessment tools) and type (e.g., test scores, narrative descriptions) of information differed across groups. While Joan, as school administrator, received a range of standardized assessment information (e.g., individual and aggregated test scores, report cards), she had access to very little classroom-based assessment information. In contrast, the classroom teacher, June, had access to a broad range of standardized and classroom-based assessment information. Parents also had access to both standardized and classroom-based assessment information. Information availability was narrowest for students. Analysis of the types of assessment information available to assessment consumers helps to distinguish gaps in information availability (that are easily corrected by better reporting practices) from assessment information that was unobtainable from the assessment program. This analysis also provided insight into the identification of factors impacting the use and value of assessments described next in Chapter 5.

If literacy assessments are used by a broad range of consumers for a variety of different purposes (Farr, 1992), this may necessitate the design of assessment programs that include several assessment tools, like the program at Highmeadow. While complex program designs may be warranted, program planning is often additive rather than integrative. By additive I mean that assessment tools are independently *added* to an assessment program by policy-makers, administrators (frequently norm-referenced standardized tests) and teachers (most often classroom-based assessments) without considering what the established program has to offer. A more integrative planning model would address the value of the program as a whole (including all of its constituent tools) prior to the addition of new tools.

The gradual, mindless, and piece-meal accumulation of assessment tools making up Highmeadow's program reflects the additive model of assessment implementation design. The identification of unnecessary redundancy in and consumer valuing of available information would be more likely to result in an integrated assessment program that addresses the desired uses and values of consumers while minimizing the amount of time students spend on assessment tasks. Thus, in Chapter 5, I evaluate assessment tools (and associated information) making up the Highmeadow literacy assessment program through an analysis of consumer assessment use and value.

CHAPTER FIVE

ASSESSMENT PROGRAM VALUE

The usefulness of an assessment is a value judgment (Messick, 1989a) influenced by our desired uses and our beliefs about what makes assessment tools and information meaningful. Assessment-consumer groups with different needs (Farr, 1992) may use and value different kinds of assessment tools and information. If the diverse consumer groups at Highmeadow use and value different kinds of assessment information, this fact would help account for the broad and expanding literacy assessment program described in Chapter 4.

Shifts in literacy theory (e.g., Rosenblatt, 1991) and classroom instructional practices (like those associated with the Book Club curriculum implemented in June's classroom) may also contribute to this expansion in assessment. Assessments which were implemented prior to recent curriculum and instructional changes (e.g., basal test) may not provide useful information to current assessment consumers. Moreover, tools implemented at different levels and times by different policy makers may create unnecessary redundancies in the program (and result in the dual-system assessment program outlined in Chapter 4).

Assuming we desire to keep the assessment of students to a minimum, we must decide if there are tools that can reasonably be excluded from assessment programs. But how do we justify our decisions, if all the

assessment tools have been implemented to provide valued information of some kind? The key to deciding which assessment tools should be excluded lies in identifying and eliminating tools that are unused, those that provide unnecessarily redundant information, and those that address less important consumer uses.

In this chapter I evaluate the assessment tools constituting Highmeadow's literacy assessment program by examining patterns of assessment use by consumer groups. In my analysis, I first characterized patterns of assessment use both across and within assessment consumer groups that were suggestive of assessment value. I then identified properties and dimensions of assessments that accounted for their value by assessment consumers. Thus, this chapter is organized around the following two research questions: (1) How did assessment consumers use available assessment tools and information? and (2) what dimensions of assessments (and associated information) impacted how they were used and valued by assessment consumers?

Assessment use by consumers

To evaluate the assessments making up Highmeadow's literacy program which was described in Chapter 4, I explored patterns of assessment use by consumers. The ways in which assessment consumers use assessment tools and information is one indication of the value they attribute to them. Thus, my analysis of these patterns provided evidence of value. In this

section I examine two patterns of use: (1) use of assessment tools across consumer groups, and (2) specific uses each consumer group made of assessment tools.

Use of assessment tools across consumer groups

To understand patterns of use across assessment tools and consumers, I first crossed-referenced assessment tools with consumer groups to reveal the number of consumers which made use of each tool (see Table 8). In the table cells following each consumer column heading, I have indicated the tool's status of use by the consumer. A blackened cell indicates that information from the tool was used by the consumer. A dotted cell indicates that the information was available, but not *used* by the consumer. A slashed cell indicates that information from the tool was *not available.* Data to support use status was drawn from consumer interviews and surveys as well as from fieldnotes documenting my direct observations (e.g., I observed June use Book Club portfolios to report to parents during parent-teacher conferences).

Table 8-Profile of assessment tool use across consumers

SYSTEM	TOOL	CONSUMER				
		Administrator	Teacher	Parents	Students	TOTAL
Standardized assessment system	MEAP	■	⋯	■	■	3
	BOTEL	■	▨	▨	▨	1
	CTBS	■	⋯	⋯	▨	1
	CogAT	■	⋯	⋯	▨	1
	Basal Test	⋯	⋯	▨	▨	0
	Reading/writing archival portfolio	⋯	⋯	▨	▨	0
	Report Card	■	■	■	■	4
Classroom-based literacy assessment system	Parent-Teacher Conference	▨	■	■	▨	2
	Portfolios & other classroom/homework artifacts or behaviors	▨	■	■	■	3
	Teacher written evaluations & informal communications	▨	■	■	■	3

Table 8 illustrates that while most (80%) assessment tools were used by one or more consumer groups, two (20%) were not used at all. The MEAP and report cards were the most widely available (i.e., available to all four consumer groups) and frequently used (i.e., used by three of four consumer groups) sources of assessment information. Portfolios and teacher evaluations were not as widely available (only available to three of the four consumer groups including the teacher, parents and students), yet they were used by all consumer groups to which they were available. Nevertheless, evidence suggests that even if this information had been available to Joan, the school principal, she would not have used it. For example, Joan stated in her interview that she did not seek out this kind of information because she did not "generally need information at this level." Thus, the MEAP, report cards, portfolios and teacher evaluations were used by at least three of the four consumer groups, suggesting a degree of value within the context of the assessment program.

In contrast to these widely used tools, information from the basal test and the school-wide reading/writing archival portfolio did not serve any clear need. June had information from the archival portfolio and basal-test data readily available given the fact that she selected pieces for the portfolio and received basal-test scores in the spring of every school year. Despite availability, June made it clear that she did not find this information of value, as suggested by the following interview turn:

"No I don't do anything with them [CTBS, CogAT & basal test results]. They're stupid, they're just like a basic you know there's adding and there's language and...And it takes a whole two weeks of school time to take those three tests. So it's a waste,...and I've petitioned not to administer the basal test at all this year."

Like June, Joan had information from both the archival portfolio and the basal test available to her. While Joan failed to communicate whether she used the archival portfolio, she reported that (she as well as others) did not use or value basal-test results:

Joan: And um, and so they had that reading curriculum and then they selected an assessment that really didn't match. And so basically it did a couple things. First of all, it didn't tell us everything we wanted to know about whether our kids could do the kinds of things we wanted them to do or not. And it also, I feel, encouraged teachers not to move forward in terms of implementing a new curriculum. Cause if you wanta do well in the basal test, you're gonna hafta teach the basal test.

Tanja: Right. And so how is that test information used? Does anyone use it?

Joan: No! We don't.

This exchange not only reveals Joan's aversion to the basal test but it suggests an additive model of assessment implementation. Policy makers and the teacher implemented new assessment tools in an attempt to keep pace with changes in curriculum and accountability demands. As a result, the assessment program expanded and the value of previously implemented assessment tools was not addressed within the context of the evolving

educational system.

Information from the archival portfolio and basal tests was not reported directly to parents and students. Nevertheless, Joan reported that this information was accessible to students and parents via the CA60s (student cumulative academic files) permanently located in Highmeadow's administrative office. Despite some level of accessibility, parents and student failed to seek out this information, suggesting a lack of interest or need.

While frequency of use across consumer groups provided a clear indication of assessment value (or lack of value) for the extreme cases, the evaluation of tools that were available to and used by only one consumer group was more difficult. For example, information from the remaining standardized tests (i.e., Botel, CTBS, and CogAT) and parent-teacher conferences was only available to and used by specific consumer groups. Joan, the school administrator was the only consumer to report using information from the Botel, CTBS, and CogAT. Availability of information from the Botel was restricted to Joan, potentially accounting for its lack of use by other consumers. Nevertheless, June received information from the CTBS and CogAT and shared this information with parents during parent-teacher conferences. Despite this fact, only one parent reported using information from these tests (i.e., "[I know my child is making progress] Thru [sic] interpretation of the various achievement measures, e.g., CogAT, CTBS, MEAP by teachers at parent teacher conferences."). And June reported

finding information from these tests useless as suggested by the interview turn presented above (e.g. "So it's a waste...").

And as might be expected, parent-teacher conferences were only used by parents and teachers. While four (out of 26) students reported indirectly receiving information from these conferences via their parents (e.g., "parents can get the information by going to confrences [sic] and tell you how your [sic] doing"), students did not have direct access to these conferences. Likewise, Joan, the school administrator was not routinely privy to information from parent-teacher conferences. Nevertheless, Joan reported participating in conferences when decisions about student instructional placement were made.

While the value of tools like the MEAP and report cards is clear from their wide use across consumer groups, findings which indicate a selective use and valuing of assessment tools by specific consumer groups raises the issue of whether the restricted use of assessments (e.g. Botel, CTBS, CogAT, parent-teacher conferences) justifies their continued inclusion in the literacy assessment program. The issue of inclusion is a particular concern when considering the imposition of some of these tools on teacher planning and classroom instructional time as suggested by June's interview turn above (e.g. "takes a whole two weeks of school time..."). In the next section, I explore the specific uses each consumer group made of available assessment tools and information. I conducted this analysis to further evaluate the assessment

tools which were used infrequently across consumer groups (e.g., Botel, parent-teacher conferences) and to provide additional evidence of value for more widely used tools (e.g., MEAP, report cards).

Assessment uses within consumer groups

To gain a better understanding of assessment-use patterns laid out in Table 8, I conducted a domain analysis (Spradley, 1980). First, I reviewed survey and interview data for instances when consumers reported using specific assessment tools and information in particular ways (e.g., parents reported using MEAP scores to evaluate school programs). I then grouped the consumer-stated uses into categories. Results from this domain analysis are presented in Figure 4.

Figure 4-Domain analysis: assessment uses

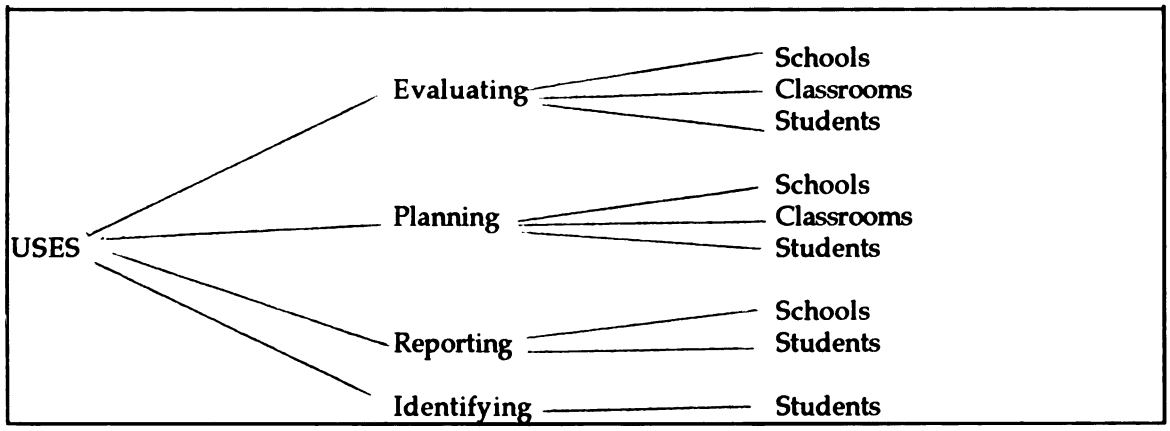


Figure 4 illustrates that assessment consumers used assessment tools in four ways (i.e., evaluating, planning, reporting, identifying) and at three levels of

focus (i.e., school, classroom, student). The four uses included one interpretative and three action-based uses. Evaluating was interpretative. I characterized this use as interpretative because it involved consumer judgment (with or without associated action-based decision making). Instances of this use were similar to what Messick (1989a) has referred to as assessment interpretation (in contrast to what he calls action-based use). For example, parents evaluated school programs, but did not report taking any particular actions based on their evaluation (beyond enrolling in Highmeadow). More often than not, however, instances of evaluation supported the other three uses.

Planning, reporting, and identifying uses all involved identifiable action-oriented decisions based on evaluative judgments. For example, Joan's school program planning resulted in curriculum refinement (i.e., school improvement plan) and staff development design (i.e., arranging inservice sessions for teachers); these changes were based on Joan's evaluation of the quality of the school program. Reporting involved evaluative statements made to policy makers, parents, etc., about school (e.g., MEAP scores reported to the public in the newspaper) and student progress (e.g., report cards sent home to parents). Finally, identifying resulted in the placement of students in special programs (e.g., at-risk, gifted and talented).

In the cell following each consumer column heading in Table 9, I have indicated level(s) of focus of each use characteristic for each consumer group.

Table 9-Uses of assessments by consumers

USE	CONSUMER			
	Administrator	Teacher	Parents	Students
Evaluating	School programs	Curriculum & instruction Student progress	School programs Student progress	School programs Student progress
Planning	School programs	Curriculum & instruction	Support of student progress	Student progress
Reporting	School progress	Student progress		
Identifying	Students			

For example, June's evaluations focused at the levels of the classroom curriculum and instruction (e.g., evaluating the effectiveness of the Book Club curriculum) and the student (e.g., evaluating individual student progress in the area of journal writing), while Joan's evaluations focused at the higher level of school programs. Blackened table cells indicate that the use was not demonstrated by that consumer.

Table 9 suggests that there were instances of consumer groups demonstrating similar assessment uses and levels of focus (e.g., the administrator and parents both evaluated school programs). More often than not, however, consumer groups' differed in either their use pattern or level of focus. This finding partially accounts for the diversity in assessment tools used across assessment groups. Nevertheless, to better understand the assessment-use patterns across groups described in the previous section (and

to generate additional evidence to justify the recommended inclusion or exclusion of particular assessments as part of the literacy assessment program), I analyzed how tools were used by each consumer group (i.e., specific uses). In this analysis, I identified tools that were applied to address multiple uses within consumer groups. I also characterized uses according to their importance to each consumer group and evaluated assessment tools and information in terms of their application toward addressing these uses.

School administrator. Table 10 characterizes the uses Joan made of assessment tools. Joan, as school administrator, engaged in all four identified uses. While her level of focus was most often the school (i.e., three of four uses), she also identified individual students for special programs.

Table 10-Administrator uses of assessment tools

USE	LEVEL	TOOLS
Evaluating	School programs	MEAP, CTBS, CogAT
Planning	School programs	MEAP
Identifying	Individual students	Botel, Report cards
Reporting	School progress	MEAP

This table suggests that Joan highly valued information from the MEAP because she drew on the MEAP for multiple (i.e., three) uses. She used the MEAP to evaluate and plan school programs, as well as to report school progress to policy makers, parents and the public. Joan's use of MEAP information to evaluate and plan school programs is apparent from the

following interview exchange:

Tanja: Okay...what are your school improvement goals related to the MEAP and what other goals do you have and where do they come from?

Joan: I think that the majority of [district] schools use MEAP scores right now to develop their school improvement plans...And so basically our major school improvement goal right now, our student outcome goal is to increase the students' ability to comprehend informational texts.

Tanja: Okay.

Joan: And um, the school, you know, obviously we have informational texts because our MEAP scores are not high. Our overall reading scores for students are high. On narrative texts they were better than informational texts. So basically we just put together a plan and we're gonna be implementing that plan and monitoring it over the next three to five years. Our goal is to obviously to get 100% [passing scores] on the MEAP.

Tanja: What else goes into the school improvement plan?

Joan: I give a presentation on analyzing MEAP to the whole staff. We then plan inservices to help teachers with instructional practices for areas of weak student performance--like informational texts.

This exchange illustrates how Joan used information from the MEAP to analyze student performance and plan school-wide curricular changes (i.e., more curricular focus on informational texts school wide) and staff development. It also demonstrates how she used MEAP scores to evaluate school programs over time (i.e., "monitoring it over the next three to five years"). Finally, Joan's in-depth explanation of how she used the MEAP to evaluate and plan school improvements (and the fact that she spent 22/69 of

her interview turns discussing evaluation and planning uses) suggests the importance Joan placed on these uses. In contrast, Joan failed to discuss the reporting use during her interview and simply listed the people (i.e., policy makers, teachers, parents, public) she reported to and the information she reported (i.e., MEAP scores) in the written survey.

While Joan's reliance on the MEAP is apparent, the value of other assessments to her is less clear. For example, Joan described using information from the CTBS and CogAT for the important purpose of evaluating school programs. The following turn from Joan's interview portrays this use:

> "Well, first of all, though, in our building...we, we haven't used them as much as the MEAP for school improvement. However, we use 'em as another piece. And so we basically just check our kids to see how they're doing. So I, I say we really use 'em...they either confirm or disconfirm our beliefs about kids. That's about it."

A close examination of the language Joan used in this interview turn indicates that she valued the CTBS and CogAT less than the MEAP. Joan stated that she used CTBS and CogAT scores to "just check," and "confirm or disconfirm our beliefs," suggesting that these tests only served a supportive secondary role to information from the MEAP. This conclusion is supported by the fact that 16 of the 22 interview turns devoted to discussion of school program evaluation focused on the MEAP, while the interview turn above

was the only instance Joan addressed the use of the CTBS and CogAT in school program evaluation. Second, use of the CTBS and CogAT was limited to a single function, evaluating. Thus, information from the CTBS and CogAT was unnecessarily redundant in Joan's evaluation of school programs.

While identifying was not a frequently cited use for assessment information (i.e., Joan devoted 2/69 interview turns to discussion of this use), Joan provided in-depth explanations about how she identified students for special programs suggested by the following interview exchange:

> Joan: I start off the beginning of the year with the lower grades cause they're newer to our school and we have an early intervention, you know, strategy... And I've been able to look at report cards and say, you know, when I read this report card, I sorta get the feeling this little cherub's in trouble. And the teacher will go, yeah, now that you mention it...Well, then what kinda trouble are we in here? What do we need to do about it?...We have a lot of monitoring mechanisms in place to identify students that are at risk for anything, for any reason.
>
> Tanja: What are the others?
>
> Joan: We give, this is very old fashioned, but we give a Botel to all our kids, 2nd through 5th grade, which is a words opposites and a word rec just to monitor, you know, a piece of their reading progress. We give that at the beginning of every year and I look at that to see if our kids have grown a year in that particular area. We look at, and then obviously the test scores, we look, I look to see if any students have failed to grow a full year or are scoring way below their peers and are at-risk for experiencing difficulties in reading. If so, I then talk to teachers about my concerns, about what we can do for these at-risk students.

While Joan relied exclusively on information from the MEAP for school-

level uses including evaluating, planning, and reporting, this interview exchange reveals that Joan drew on both Botel scores and report-card marks to identify at-risk students for placement in special programs. While this finding suggests redundancy in the program, it is difficult to determine from this analysis of Joan's assessment tool uses whether she believes this redundancy is desirable.

Classroom teacher. Table 11 characterizes the uses June made of assessment tools. While June did not personally identify students for special programs, she used assessment information to evaluate, plan, and report curriculum and instructional effectiveness and student progress.

Table 11-Teacher uses of assessment tools

USE	LEVEL	TOOL
Evaluating	Student progress	Report cards Parent-teacher conferences Portfolios Teacher evaluations
	Curriculum & instruction	Portfolios Teacher evaluations
Planning	Curriculum & instruction	Portfolios Teacher evaluations
Reporting	Student progress	Report cards Parent-teacher conferences Portfolios Teacher evaluations
Identifying		

Inspection of Table 11 reveals that June drew on the same four tools (i.e.,

parent-teacher conferences, report cards, portfolios, and teacher evaluations) for multiple (two or more) uses. While June's reliance on a single tool for multiple uses suggests her valuing of the tool, her dependence on multiple tools for a single use indicates potentially needless redundancy (as was the case with Joan's use of the CTBS, and CogAT in addition to the MEAP for program evaluation) in the assessment program. A closer examination of June's reliance on multiple tools for a given use, however, suggests that this redundancy was useful and complementary. It is clear June used information from these tools in an integrated fashion, not simply to confirm one another. June's integrative use of information from these tools to evaluate (and guide) student progress, and report progress to parents is suggested by the following interview turn:

> "What I have, then I have my journal checklist and my book club checklist and so then I keep these cards with each of the students names on it and [anecdotal] notes. And then that helps me to fill out the checklist and then if I have comments on their report card and then the parents need data if I say "really not engaged in discussions." Then they say well they're talking a lot at home, then I can say well and I can pull out, when I'm sitting and taking notes this is what I'm seeing in the group. So it really helps to back up what you're thinking. Umm also I collect their journals every week and I write comments and questions and fill out a checklist on that to help them guide them with

umm what they're doing in their umm journals. And then I do a lot of self assessment either just something quickie like you saw today where they take two minutes to write a new journal. What do you think, how do you think, comments when and why and goal setting."

June's response portrays an intricate system for evaluating and reporting student progress to parent. For example, through the use of daily anecdotal note cards (one form of teacher evaluation) documenting student performance during classroom discussions and on journal entries, June filled out biweekly checklists evaluating student progress on specific objectives (e.g., provides evidence from text, expresses opinions, compares story with genres or stories previously read). June then used this cumulative record of performance (in addition to report cards), to report student progress to parents during parent-teacher conference. Thus, June justified the value of using multiple tools to evaluate and communicate individual student progress to parents.

June also used multiple tools to evaluate and plan classroom curriculum and instruction as suggested by the following interview turn:

"So I created this portfolio system to evaluate...my literacy program...I have the checklists and that's easy like to check off, but this is kind of letting me see you know what am I looking for...And that'll affect my instruction cause if I see like something that a lot of kids don't do...I'm seeing that students aren't choosing to respond in their journals in a

certain area. Then I'll kind of pull the whole class together and do you know talk to them about it and give an example. Like maybe read a picture book and then do that type of response orally with the kids and then have them do it as a group."

This interview turn illustrates how June used information from classroom-based artifacts (e.g., journal entries) and her evaluations (e.g., journal checklist) to evaluate and plan classroom curriculum and instruction. It also demonstrates how she analyzed student-performance patterns reflected in her evaluation to target daily literacy objectives and design unique instructional activities. These findings further support the value of this integrated set of complementary tools and justify their combined use, despite apparent redundancy.

Parents. Table 12 characterizes the uses parents made of assessment tools. While parent uses did not include reporting progress or identifying students, 20/26 parents used assessments to evaluate Highmeadow's reading program and all 26 parents used assessment information to evaluate their children's progress. Only four parents, however, used information to plan support of student progress.

Table 12-Parent uses of assessment tools

USE	LEVEL	TOOLS	NUMBER OF PARENTS (n=26)
Evaluating	School programs	MEAP	14
		Report cards	2
		Parent-teacher conferences	3
		Teacher evaluations	1
	Student progress	MEAP	16
		Report cards	16
		Parent-teacher conferences	7
		Portfolios	17
		Teacher evaluations	23
Planning	Support of student progress	MEAP	2
		Report cards	0
		Parent-teacher conferences	0
		Portfolios	1
		Teacher evaluations	1
Reporting			
Identifying			

As suggested by Table 12, the MEAP was the most frequently cited tool (14/26 parents) used to evaluate Highmeadow literacy program (e.g., "It's nice to have a system to compare how various schools are doing [in reading]," "So I can judge the quality of education). Nevertheless, at least one parent acknowledged what she saw as a limitation of the MEAP for this use (i.e, "This is such a small portion of the child's review and there are so many variables at the time of testing. To use this as the only gauge of the school would be short sighted"). Only six parents relied on report cards and classroom-based information sources such as parent-teacher conferences (e.g., "Most of the information about how the school is doing comes from parent-

teacher conferences") to judge how the school was doing and two of these parents used a combination of the MEAP, report cards, and teacher conferences to evaluate Highmeadow's reading program. Thus, information from the MEAP was most widely used and valued by parents for evaluating school programs.

While parents focused on one primary assessment tool for the evaluation of school programs, the MEAP, Table 12 suggests that as a group parents drew on multiple tools for the evaluation of student progress. For example, almost equal numbers of parents relied on the MEAP (16 parents), report cards (16 parents) and portfolios (17 parents), while only a few more (23 parents) relied on teacher evaluations. All parents relied on one or more tools to evaluate their children's school-related progress. This pattern of multiple tool use by parents is reflected in the following survey response to the question concerning the kind of feedback parents like to have about their children's progress:

> "I find a parent-teacher conversation or conference to be most helpful along with samples of [my child's] work. The teachers know the situation firsthand and the student's work will support or not support their evaluation and the marks on the report card. With copies of the child's work you can then physically see the child's problems. "

This response describes the way that this parent drew on multiple sources of information to get an in-depth sense of the child's progress in school.

Of the 16 parents who used the MEAP to evaluate student progress, 11 parents reported using it to compare their child's progress to other students (norm-referenced use), five parents used it to evaluate student progress in terms of a "broader set of criteria" (criterion-referenced use), and one parent used it to make both kinds of evaluations. Two parents also used information from the MEAP to supplement school-related information on student progress. These two parents stated that they used MEAP scores to confirm (i.e., "It has given me a reinforcement of subjects I know he excels in or areas of weakness") and disconfirm (i.e., "It assured me that although my kids might be having problems in school they were basically quite bright by a wider set of judgement criteria") evaluations of their children's progress based on other assessment tools. These findings suggest that parents valued multiple assessment tools for evaluating student literacy progress both in school and in general.

Finally, only four parents used assessment information to plan support of student progress. These parents reported using information from classroom-based assessments (e.g., "specific strategies for improving problem areas works best in a conference") and the MEAP either to lobby for instructional support of weak areas at school (i.e., "I used the MEAP to stress more time on areas that were lower") or assist children in setting goals (i.e., " to help [my] child set realistic goals"). Nevertheless, at least one parent stated that a major weakness of the MEAP was a lack of information on "what can

be done to strengthen weak areas." Thus, while a few parents used assessment information to plan support of student learning, this use was not a priority for the vast majority of parents.

Students. Table 13 characterizes the uses students made of assessment tools and information. While students failed to use assessment information for reporting or identifying, they did evaluate their own progress and that of the school. They also used assessment information to plan their own learning.

Table 13-Student uses of assessment tools

USE	LEVEL	TOOLS	NUMBER OF STUDENTS
Evaluating	Student progress	MEAP Report cards Portfolio Teacher evaluations	19 15 26 17
	School programs	MEAP	1
Planning	Student progress	Report cards Portfolio Teacher evaluations	10 5 11
Reporting			
Identifying			

Table 13 reveals that while all students used information from at least one assessment tool to evaluate their own progress (e.g., "Looking through my portfolio lets me know how I'm doing), only one student reported using the MEAP to evaluate school programs (i.e., "The newspaper tells how the whole school did on the MEAP"). Not surprisingly, student-progress evaluation was

a priority with students while evaluation of the school was not.

Students drew on multiple assessment tools to evaluate their progress. While the MEAP was used to evaluate general reading ability (e.g., "That I'm very good at reading"), report cards and classroom-based assessment tools were used by students to evaluate school-related literacy progress (e.g., "The teacher tells you if you share more in Book Club or write more in your journal"). This finding suggests that information from the MEAP is redundant in the evaluation of student progress, and probably unnecessary for this use.

Almost one-half (45%) of the students who used report cards, portfolios and teacher evaluations to evaluate their progress also used these tools to plan progress. In contrast, none of the students who used the MEAP to evaluate their progress reported using this information for planning. These findings suggest that while students used the MEAP for evaluation, it was only used to confirm evaluations made based on other assessment information. This finding supports the conclusion that the MEAP provided redundant and unnecessary information for student uses. In general, students valued information from report cards, portfolios, and teacher evaluation for a wider range of important uses.

Overall, my analysis of tool use by consumers suggests two trends regarding the value of assessment tools: (1) valued tools were used by multiple consumers, and (2) information from valued tools was drawn on for

multiple and important uses both across and within consumer groups. The MEAP, report cards, the classroom-based portfolio and teacher evaluations were all used by three or more consumer groups. These tools were also used by individual consumer groups for multiple and important uses. These findings support the continued inclusion of these tool in Highmeadow's assessment program. In contrast, the basal test and school-wide archival portfolio were not reportedly used by any consumer group for any purpose, indicating that these tools were not valued; thus, the inclusion of these tools in the Highmeadow literacy assessment program should be re-considered.

The analyses of assessment-tool use by consumers also elucidated information redundancies in Highmeadow's literacy assessment program. Redundancies were evident when a consumer group reported drawing on multiple tools for a single use. These redundancies took two forms: (1) confirmatory and (2) complementary. Confirmatory redundancies were usually unnecessary and unjustified. Joan, for example, relied on information from the MEAP, CTBS and CogAT to evaluate school programs. Additionally, she drew on the MEAP when planning school improvements. Nevertheless, Joan only used the CTBS and CogAT to confirm evaluations made based on the MEAP, and these tools did not impact her planning at all. Thus, the use of the CTBS and CogAT was simply confirmatory and did not direct action. In contrast, complementary redundancies were warranted because of the integrated use of assessment information. For example, June

used information from a complex system of **complementary**, curriculum-relevant tools including report cards, portfolios, teacher evaluations, and parent-teacher conferences to evaluate, and plan classroom curriculum and instruction and report student progress.

Finally, findings support the assertion that different consumers need and value different kinds of assessments (Farr, 1992). Table 14 summarizes the specific uses consumer groups made of assessment tools and suggests that different consumer groups used and valued different assessments.
For example, Joan, the school administrator relied on standardized assessments, particularly the MEAP for predominately school-level uses. Furthermore, other consumers that engaged in school-level uses (i.e., parents, students) also made used of the MEAP. Thus, the MEAP appears to be of value to consumers for school-level uses. In contrast, June, the students and parents primarily engaged in classroom- and student-level uses. While parents and students drew on information from the MEAP for evaluating student progress, all of these consumer groups relied predominately on classroom-based assessments. But, what is it about the MEAP that makes it attractive to consumers for school-level uses and classroom-based assessment for classroom- and student-level uses? In the final section of this chapter I identify dimensions of assessments that were critical to consumers' valuing of and decisions to use them.

Table 14-Use by level made of assessment tools by consumers

LEVEL & USE	CONSUMER	TOOLS
SCHOOL Evaluating	ADMINISTRATOR	MEAP, CTBS, CogAT
	PARENTS	MEAP, Report cards, Parent-teacher conferences Teacher evaluations
	STUDENTS	MEAP
Planning	ADMINISTRATOR	MEAP
Reporting	ADMINISTRATOR	MEAP
Identifying		
CLASSROOM Evaluating	TEACHER	Portfolios, Teacher evaluations
Planning	TEACHER	Portfolios, Teacher evaluations
Reporting		
Identifying		
STUDENT Evaluating	TEACHER	Report cards, Parent-teacher conferences Portfolios, Teacher evaluations
	PARENTS	MEAP, Report cards, Parent-teacher conferences, Portfolio, Teacher evaluations
	STUDENTS	MEAP, Report cards, Portfolio Teacher evaluations
Planning	PARENTS	Portfolio, Teacher evaluations
	STUDENTS	Report cards, Portfolio, Teacher evaluations
Reporting	TEACHER	Report cards, Parent-teacher conferences Portfolios, Teacher evaluations
Identifying	ADMINISTRATOR	Botel, Report cards

Dimensions of assessment tools and their value to consumers

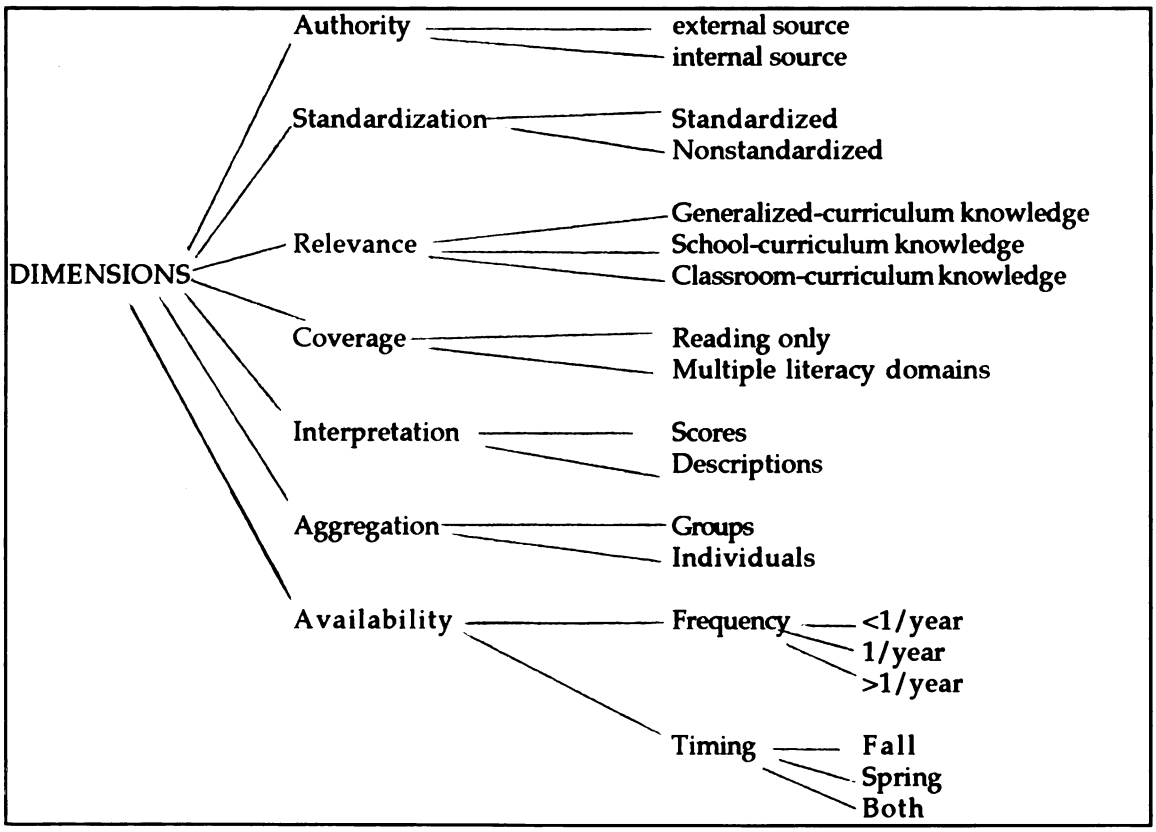
Through these analyses, I defined dimensions of assessments that consumer groups paid attention to when deciding whether to use available information. I characterized the assessment values of consumers through the identification of these dimensions, coupled with findings on consumer tool use. These analyses not only highlighted the values of assessment consumers but provided a framework for exploring assessment gaps (i.e., assessment properties desired by consumers but not available from established tool in the program). They also offered guidelines for examining the potential value of proposed assessment tools for filling those gaps (the focus of Chapter 6).

To answer the question about valued assessment dimensions, I conducted a domain analysis (Spradley, 1980). First, I identified an initial set of assessment dimensions that I defined in terms of contrasting assessment properties. These properties were defined as characteristics that distinguished between specific assessment tools (e.g., standardized administration/scoring versus nonstandardized administration/scoring). I selected dimensions and properties based the characteristics of assessments that I believed consumers would care about when deciding to draw on a specific tool/information for a particular use.

Second, I catalogued consumer interview and survey data according to assessment properties which lead consumers to use or ignore assessment information. Through this analysis, I expanded and refined the initial set of

assessment properties and dimensions. The resulting set of properties and dimensions, and their relationships, are illustrated in Figure 5.

Figure 5-Domain analysis: assessment dimensions and properties



Dimensions and properties of assessment tools

Through the domain analysis presented in Figure 5, I identified seven dimensions that influenced consumers' decisions to use assessment tools. The first dimension, **authority**, reflected the source mandating the assessment implementation. The source was either the teacher inside the classroom (i.e., internal) or an administrator or policy maker at the school, district, or state

level (i.e., external). Second, **standardization** reflected consistency in assessment items/tasks, and performance standards across students. Students either participated in the same tasks and were evaluated using a standard set of criteria, or they were given the freedom to select tasks and were evaluated individually. The third dimension of **relevance** referred to the knowledge domain that assessment items/tasks reflected. Tasks reflected the school curriculum, the classroom curriculum, or more general curriculum-related knowledge (e.g., verbal ability). Fourth, the **coverage** dimension reflected the breadth of literacy assessment tasks. Assessments were either limited to reading-specific tasks or included tasks which addressed multiple literacy domains (e.g., discussion, listening). The fifth dimension, **interpretation,** referred to the form of resulting assessment information. Interpretations included numerical scores and verbal descriptions. Sixth, **aggregation** reflected the level at which assessment information was available. Information was made available on individuals and/or groups. And finally, the dimension of **availability** referred to the *frequency* (i.e., how often the information was made available) and *timing* (i.e., when during the school year information was made available) of availability across and within consumer groups. Table 15 outlines the dimensions and properties of the four most frequently used assessment tools across and within consumers (see Appendix M for complete analysis of all tools in Highmeadow's literacy assessment program).

Table 15-Dimensions & properties of widely used tools

TOOL	DIMENSIONS	PROPERTIES
MEAP	Authority Standardization Relevance Coverage Interpretation Aggregation Availability	External source Standardized School curriculum Reading only Scores Groups & individuals Fall, 1/year
Report Card	Authority Standardization Relevance Coverage Interpretation Aggregation Availability	External source Standardized School (& classroom) curriculum Multiple domains Scores & descriptions Individuals >1/year
Portfolios AND Teacher evaluations	Authority Standardization Relevance Coverage Interpretation Aggregation Availability	Internal sources Nonstandardized Classroom curriculum Multiple domains Descriptions Individuals >1/year

Dimensions and properties valued by consumers

Analysis of survey and interview data revealed the dimensions which individual consumer groups considered in their decisions to use information from assessment tools making up the Highmeadow literacy assessment program. Table 16 summarizes the dimensions of assessment tools and information valued by each consumer group when deciding to use them.

Table 16-Dimensions of assessments valued by consumers

CONSUMER	VALUED DIMENSIONS
Administrator	Authority, standardization, relevance, interpretation, and aggregation
Teacher	Relevance, coverage, and availability
Parents	Relevance, coverage, and availability
Students	Availability

Joan, as school administrator, focused on dimensions of authority, standardization, relevance, interpretation, and aggregation when deciding to draw on information from assessment tools. For example, when I asked Joan to explain why she focused so heavily on the MEAP for school improvement evaluation and planning she relied:

> "Just because, yeah, because it's published in the papers and because...It was based on the state's core curriculum and our curriculum that was developed...But it was developed based on the state's core curriculum, too, so they're in line."

This interview turn suggests that Joan valued the MEAP because it was mandated by the state (external source) and provided her the opportunity to be accountable to the public. It also indicates Joan's focus on the assessment's relevance to the school curriculum. Joan's valuing of school-curriculum relevance is also reflected in her failure to use the CTBS, CogAT and the basal test for school program planning; she believed these two assessment tools were irrelevant to the school's curriculum. Joan also discussed her desire for

"standardized" information across students like that from report cards and the Botel to compare and ultimately identify students for special programs. Finally, Joan expressed a desire to have both aggregated and disaggregated "scores" for identifying individual students and evaluating and planning school programs. These conclusions are also supported by the fact that Joan did not use classroom-based assessments because they were not uniform and interpreted in terms of scores that could be aggregated across students.

While June, the classroom teacher, and parents focused on the same dimensions of relevance, coverage, and availability when deciding whether to use assessment tools, they often looked for different properties. For example, June was interested in assessment information that was relevant to her classroom curriculum when she evaluated and planned her instruction. In contrast, 10/14 parents were interested in relevance of the MEAP to the school curriculum (e.g., "we don't know how valid the MEAP criteria are in terms of the school curriculum") when using it to evaluate Highmeadow's literacy program. Nevertheless, because classroom-based assessments provided information frequently and addressed all domains of literacy including both oral and written language, both June and her students' parents valued these tools for evaluating student progress in school.

Finally, the only dimension of assessment tools that appeared to be of interest to students was that of availability. While students did not express valuing tools based on their availability, students did report using all

information that was made available.

Summary

In summary, my analysis of tool use by consumers suggests that information from valued tools is applied for multiple and important uses both across and within consumer groups. Highly valued tools included the MEAP, report cards, classroom portfolios, and teacher evaluations. Tools that were deemed less useful were those that were not used at all (e.g., basal test, archival portfolio), used for fewer or less important purposes (e.g., Joan's restricted use of the Botel for identifying students) or provided unnecessarily redundant information (e.g., CTBS, CogAT when used by Joan for school program evaluation). Thus, these findings suggest that the inclusion of several tools in Highmeadow's literacy assessment program could not easily be justified based on patterns of use.

Moveover, I discovered that assessments are used and deemed valuable by consumers as a function of the desired use and consumers' beliefs about what makes assessment tools and information meaningful. For example, June used assessment information to plan her classroom instruction and therefore, she valued assessment tools that provided her with information directly relevant to her curriculum. Thus, findings suggest that different assessment consumers value assessment tools differently. Thus, there is a need for a certain degree of diversity in Highmeadow's literacy assessment program.

Findings also include a set of eight valued assessment dimensions (i.e., authority, standardization, relevance, coverage, interpretation, aggregation, availability, openness), each with two or three contrasting properties (e.g., classroom-, school-, and generalized-curriculum coverage). These dimensions are the aspects of assessment tools and information which consumer groups pay attention to when deciding whether to use a particular assessment for a specific purpose. Different consumer groups focus on different dimensions. For example, the school administrator focused on assessment standardization and interpretation, teachers, parents, and even students paid attention to availability. Further, different consumer groups valued different assessment properties, even when they focused on the same dimension. For example, the classroom teacher was interested in assessment information that was relevant to her classroom curriculum for planning instruction, yet parents valued school-curriculum relevance when evaluating the school.

In Chapter 6, I apply the framework of consumer value dimensions to the evaluation of the performance-based assessment. Through this analysis, I identify dimensions and properties of assessment tools consumers believe they need but are not available from the program (i.e., assessment information gaps). I then examine the properties of the performance-based assessment and the potential of the performance-based assessment for addressing the information gaps.

CHAPTER SIX

PERFORMANCE-BASED ASSESSMENT VALUE FOR FILLING ASSESSMENT PROGRAM GAPS

The value of performance-based assessments for remedying many of the limitations of standardized tests has been widely acknowledged (e.g., Bisesi, & Raphael, 1997; Delandshere, & Petrosky, 1994; Linn, Baker, & Dunbar, 1991; Valencia, 1990). Bisesi and Raphael (1997) also compared the strengths and weaknesses of a performance-based assessment to classroom-based portfolios. Nevertheless, the value of performance-based assessment has not been explored in the context of an established literacy assessment program or from the perspective of assessment consumers. Thus, in this chapter, I investigate the value of a performance-based assessment for filling consumer perceived gaps in Highmeadow's literacy assessment program.

To evaluate the performance-based assessment, I first identified assessment gaps, defined in terms of the value dimensions and properties described in Chapter 5. I then explored the potential of the performance-based assessment for filling these gaps, based on the degree of match between desired assessment properties expressed by assessment consumers and those characterizing the performance-based assessment. Thus, this chapter is organized around two research questions: (1) What assessment gaps, defined in terms of consumer reported value dimensions and properties, were present in Highmeadow's literacy assessment program? (2) What is the

potential value of the performance-based for filling the assessment gaps?

Gaps in Highmeadow's literacy assessment program

To answer the first question concerning assessment gaps, I analyzed survey and interview data for patterns of consumer-stated limitations in the assessment tools constituting Highmeadow's literacy assessment program. I then categorized these limitations in terms of the dimensions and properties outlined in Chapter 5 (and added any necessary dimension and property categories) both within and across consumer groups.

Assessment program gaps by consumer group

In this section, I present the valued-assessment dimensions that each consumer group reported were not adequately addressed by the assessment program. I also discuss the gaps (defined in terms of valued-assessment dimensions) common across consumer groups.

School administrator. Joan, the school principal was concerned that most of the assessment tools that constituted the standardized assessment system did not align well with the school's curriculum. While Joan felt that many of the available assessments did not reflect outcomes targeted by the school's curriculum (e.g., "Current assessments in reading e.g., basal test, CTBS do not match the curriculum"), she believed that the reading MEAP was a reasonable reflection. Nevertheless, Joan stated that there were no equivalent measures (to the reading MEAP) for the other language arts including writing, speaking and listening (or their integration). This

assessment-program gap led Joan to believe that the other language arts were not receiving the instructional attention that they deserved (e.g., "Currently the reading component is taking precedent over the other areas because it is the only area measured").

Joan also reported that many skills targeted by the curriculum were not tapped by Highmeadow's assessment program (e.g., "These assessments do not help us understand the child's ability to problem solve, be a collaborative team member, etc. or tell us anything about their higher order thinking skills, concept development, or planning skills"). And while Joan praised June and the other teachers for their ability to create instructionally-embedded, alternative assessments to tap these important skills, she stated that the information from these assessments was not useful for evaluating, planning and reporting school-level progress. Nevertheless, Joan reported that she could make use of information from a more standardized form of alternative assessment that covered neglected curricular and skill areas, stating:

> "Now, if she [June] comes up with a holistic, you know, performance assessment, a snapshot in time about how our students are using, a whole list of scoring system so we could say that a 4 is a standard, you know, like say you have a 1 to 6 scale or something and 4 is grade level, and 6 is above, you know, whatever your standard is."

Thus, Joan perceived a need for additional, standardized information on

multiple, literacy domains and curriculum-relevant outcomes not targeted by Highmeadow's literacy assessment program. These findings suggest a need for an assessment which is standardized, addresses all four literacy domains, and is relevant to the school curriculum.

While Joan stated that academic progress in literacy was demonstrated by a "measurable change" in behavior, knowledge, and skills, she claimed that there was little information available which allowed her to judge student and school progress from year to year ("[The program] does not measure change year to year for each group of students."). Most of the standardized tests were only administered every couple of years (e.g., MEAP in 4th grade, CTBS in 3rd and 5th grades) and those that were administered every year (e.g., the basal test) did not measure change in valued aspects of the curriculum (e.g., writing). Joan needed an assessment that was administered at least once a year and provided information on all students.

Joan believed that report cards and other classroom-based assessments focused on valued, curricular outcomes and were available regularly, providing some information on progress from year to year. Nonetheless, she found this information cumbersome to use (e.g., "It's often difficult to see patterns over time and at the classroom or school level") and lacking the consistency and reliability of more standardized forms of assessment (e.g., "classroom assessments are not systematized school wide").

Finally, while Joan reported using Botel scores and report card data for

helping her to identify young, at-risk students for instructional placement, she stated that she could use additional information to help her identify students at-risk for poor performance on curriculum goals. Joan believed that the report cards were invaluable as a screening tool for "red flagging" students who were struggling to learn the curriculum, but that information from this source was cumbersome and time consuming to analyze which only made it possible to use for identifying a small group of at-risk students (i.e., kindergartners and first graders).

The Botel, on the other hand, was administered to all students and provided Joan with scores that lent themselves to the efficient identification of poorly performing students. Despite this efficiency, however, Joan pointed out that this test only addressed two aspects of reading--word recognition and word opposites. Joan was concerned that this test might not be sensitive enough to detect all students at-risk for poor school progress because it did not target all language arts areas (e.g., writing, speaking, and listening), and specific outcomes constituting the school curriculum. Joan also reported a desire to have information that would allow her to identify students who might perform poorly on the MEAP in order to better prepare them, stating:

> "So if we can like develop assessments that are more closely aligned with our curriculum, it will... also help us then identify, you know, kids that are at risk for the MEAP... we can predict better who will or will not do well on the MEAP. Or who or will not be at risk for our

curriculum, our goals.

Joan stated further that to address this gap she needed an assessment that would provide standardized information on student progress in school from year to year. She also stated that she needed an assessment that was validated for use in predicting performance on the MEAP and improving curriculum.

These findings suggest that the assessment tool that Joan needed was standardized, was relevant to the school curriculum, provided score-based interpretations covering all literacy domains, and could be used to identify individual students and aggregated to evaluate school progress. The dimensions and properties characterizing Joan's needs (and those of the other consumers) are listed in Table 17.

Table 17-Properties of assessment needed by consumers

CONSUMER	DIMENSIONS	PROPERTIES OF NEEDED ASSESSMENT
Administrator	Standardization Coverage Relevance Interpretation Aggregation	Standardized Multiple literacy domains School curriculum Scores Individual & groups
Teacher	Standardization Coverage Relevance Interpretation Availability	Standardized Multiple literacy domains Classroom curriculum Scores >1/year & both spring and fall
Parents	Standardization Coverage Relevance Availability Openness	Standardized Multiple literacy domains School curriculum >1/year Open
Students		

Classroom teacher. June obtained most of the information she used on a regular basis from her classroom-based assessments. With this information, June identified class-level patterns (e.g., many students were not writing point-of-view journal entries) and individual-student patterns over time (e.g., one student did not attempt to write a journal entry about the author's purpose) to plan instruction and guide individua-student learning. And while she found information from her classroom-based portfolio most relevant to her classroom curriculum (and available on demand), she reported that it lacked the standardization necessary to adequately evaluate the effectiveness of her Book Club curriculum.

Because June claimed that information from the portfolio was "often too varied" (e.g., types of artifacts, nature of evaluation, schedule of collection) and the analysis too time-consuming to determine the impact of the curriculum on student learning, she reported that she had begun to "standardize" the artifacts collected from each student and the types of evaluation provided (i.e., checklist). Furthermore, June said that the portfolio system did not include a standard set of criteria (only target outcome skills and knowledge) for evaluating performance across students and over time. While June, along with her students, had begun to generate criteria for evaluating progress on target outcomes (e.g., target for being a good listener: score of 1=playing with objects, looking around the room, talking to a neighbor, making faces at speaker, shout out differing opinion; and 5=hands

are free of objects, eyes on the speaker, listening quietly), she said she needed a constrained and standard set of criteria for efficiently documenting changes in student performance as well as curriculum and instructional effectiveness.

In contrast to her classroom-based assessments, June did not find any of the standardized assessments useful for this need because they did not align well with her classroom instruction, and were not available within a useful time frame (e.g., <1/year). For example, most standardized tests reflected generalized, literacy-curriculum knowledge (e.g., basal test, CTBS, CogAT) and were not directly relevant to her curriculum. And while June viewed the MEAP as relevant to the school's curriculum (and her's to the extent that she tried to cover school-target outcomes in her classroom instruction), it was only administered to fourth-grade students in the fall of the school year and did not provide June with information about the effectiveness of her 5th-grade literacy curriculum. These standardized tests offered June little information relevant to her curriculum and no basis for determining the impact of her instruction. Thus, June stated that she needed standardized assessment information that was relevant to her curriculum and available within an appropriate time frame (i.e., "two or three times per year") in order to adequately evaluate her curriculum.

Finally, June reported the fact that Highmeadow was attempting to implement the Book Club curriculum at all grade levels, second through fifth. She said it was a school improvement goal to develop and implement

an alternative assessment tool that would provide information about Book Club curriculum effectiveness across grade levels. These findings suggest that June needed an assessment tool that generated quantitative, standardized, curriculum-relevant information (i.e., standard set of curriculum-relevant tasks and artifacts, standard set of performance criteria, performance levels defined in terms of numbers) on a twice-yearly basis (see Table 17).

Parents. Table 17 reveals that parents focused on the dimensions of standardization, coverage, relevance, availability, and openness.[4] Seventeen out of 26 parents reported using student artifacts (e.g., portfolio) to evaluate student progress (e.g., "You can actually see what students are doing in school"). Nine parents reported that they could only evaluate student achievement and progress if the artifacts were accompanied by guidelines for judging their quality (e.g., "but I still need to have some way of measuring the quality of work they are producing as well as whether or not they are working at grade level," "I basically like these [student artifacts] but need to be able to have more guidelines that tell me how my child is doing"). While parents liked teacher-narrative evaluations (e.g., "I would like to see more comments on assignments...and a note to parents indicating the teacher's evaluation"), they reported the need for a standardized set of criteria for evaluating student progress (e.g., "Having a set of criteria all students in a particular class are

[4] Openness refers to clarity in the explication of procedures used to collect assessment information and standards, and performance criteria used to interpret and evaluate performance (Bisesi, Brenner, McVee, Pearson, Sarroub, in press). Assessments can be open (procedures and standards made clear to consumers) or closed.

judged on is essential to our understanding of how well our children are doing"), suggesting a focus on the dimension of standardization.

Twenty-one out of 26 parents reported wanting assessment information on listening and speaking in addition to reading (e.g., "yes I feel the four areas are equally important," "Yes, because I think communication skills are important all through life," "Listening and speaking are abilities that will also be important but have not gotten the focus that reading and writing have"). One parent criticized the MEAP in particular, for not providing information on "writing proficiency and possibly speaking proficiency." These findings suggests the parents perceived a gap in the coverage dimension and the need for more assessment information on all four literacy domains.

While two parents were content with the current reporting schedule (e.g., "The current schedule is adequate"), ten parents indicated gaps in the availability of information. One parent stated that she would like the current forms of information more frequently (e.g., "Weekly feedback would be great, but this is time consuming"). Other parents expressed a more general desire to have more information, more frequently (e.g., "I have no reason to reject any information. I like to have all the information I can get about my child's progress as often as I can get it"). These findings indicate parents' focus on the dimension of availability when evaluating the gaps in Highmeadow's assessment program.

Finally, parents reported using report cards and student artifacts (e.g. portfolio) together to make sense of student progress. Five parents reported dissatisfaction with the current report card because it failed to provide enough specific information about how students were performing in school-related subjects (e.g. "didn't tell how compare to standard," "too vague," "need more levels," "I don't think the report cards are always reflecting what's going on [in school]"). In addition, eleven parents criticized the MEAP because it did not provide enough information about the knowledge it assessed, the scores provided, or how this knowledge was related to the school curriculum. Three parents explicitly reported needing greater clarity in information from the MEAP (e.g., "We'd like to know whether the MEAP, assess [sic] the same knowledge as is considered essential in other states and at Highmeadow") including more explanation of scores and performance categories (e.g., "I don't really understand the different categories..."). These findings suggest that parents needed an assessment that was more open.

Students. Students reported satisfaction with the information they received. Students relied almost exclusively on classroom-based sources of information (e.g., teacher, report cards), and/or their own assessments when trying to make sense of their progress in literacy. Students only mentioned standardized tests as a source of information when asked directly about their usefulness (i.e., most students reported using information from the MEAP when asked whether scores from the MEAP told them anything about their

learning).

When asked, none of the 26 students reported needing any additional information about their learning and progress. And while this response pattern may be a manifestation of students' lack of knowledge about potential information sources or a lack of critical analysis of current forms of information, students' level of sophistication in describing how they used specific kinds of assessment information suggested that this was not the case. In other words, student responses suggested that students did not perceive any need for additional assessment information about their literacy performance and progress, beyond that already available from the assessment program.

Assessment program gaps across consumer groups

All consumer groups expressed gaps in Highmeadow's literacy assessment program. While some of these gaps were unique to particular consumers (e.g., the need for more information that could be aggregated across students was only expressed by Joan, the school administrator), several were expressed across consumer groups (see Table 18 for these trends).

In Table 18, I listed the gaps (defined in term of assessment dimensions and properties) in ascending order from those cited by multiple consumer groups to those cited by single consumer groups. Table 18 reveals that there was agreement between two or more consumer groups across five of the nine properties deemed necessary in a valued assessment.

Table 18-Gaps in Highmeadow's assessment program across consumers

DIMENSIONS	PROPERTIES	CONSUMERS
Standardization	Standardized	Administrator, teacher, & parents
Coverage	Multiple literacy domains	Administrator, teacher, & parents
Relevance	School Classroom curriculum	Administrator, & parents Teacher
Interpretation	Scores	Administrator & teacher
Availability	>1/year	Teacher, & parents
Aggregation	Individual Group	Administrator
Openness	Open	Parents

For example, the properties of standardization, classroom- and school-curriculum relevance and multiple, literacy-domain coverage were all cited by Joan, June, and parents as gaps in Highmeadow's literacy assessment program. Thus, any assessment added to Highmeadow's program should possess these critical properties to increase its potential for meeting consumer needs; to maximize its potential value it should possess all nine properties. Thus, in the next section, I evaluate the performance-based assessment in terms of its value for filling the gaps in Highmeadow's literacy assessment program.

Value of the performance-based assessment

To address the question concerning the value of the performance-based assessment (PBA), I first defined the properties of the PBA in terms of their alignment with the needed properties expressed by consumers. I then drew

on consumer survey and interview data to explore the value of the performance-based assessment (as express by consumers) for filling the gaps in Highmeadow's literacy assessment program.

As described in Chapter 3, the PBA was a standardized tool (consistent set of tasks evaluated using a standard set of performance criteria and score levels) which targeted multiple literacy domains (e.g., **reading** tradebooks, **writing** journal entries, **discussing** textual content) and was designed specifically to reflect June's classroom curriculum (i.e., Book Club). We designed the assessment to be administered twice a year (in the fall and spring), to provide baseline and end-of-the-year performance data. Finally, a copy of the scoring rubric was made available to all consumer groups prior to the administration of the PBA so that they would be aware of the performance standards. Thus, as suggested by Table 19 the PBA was designed to align with the expressed needs of assessment consumers.

Given the fact the PBA was designed with the needs of these consumer groups in mind, the alignment revealed in Table 19 was not surprising to find. The consumers' survey and interview response further supported these identified properties of alignment and confirmed the value of the PBA for filling the gaps in Highmeadow's literacy assessment program. In terms of standardization, most parents (15/26 parents) said they liked the PBA because it established a standardized set of performance criteria on which to evaluate journal entries and discussions (e.g., "scores describe student work using an

objective set of criteria, not just a best guess," "the scores set out clear, measurable standards"). These examples also demonstrate parents valuing of "clear," open standards.

Table 19-Alignment between needed and PBA properties

NEEDED ASSESSMENT PROPERTIES	PBA DESIGN PROPERTIES
Standardized	Standardized
Multiple literacy domain coverage	Multiple literacy domains coverage
School curriculum relevance	
Classroom curriculum relevance	Classroom curriculum relevance
Scores	Scores
Available >1/year	Available >1/year
Individuals	Individuals
Groups	Groups
Open	Open

Nevertheless, a few parents expressed a desire for additional openness. One parent stated she required more information about how work was collected and evaluated. She was concerned about the potential for putting "too much stock" in such a small sample of student work. Two other parents were concerned with how reliably the rubric was applied, requesting a copy of the work being evaluated and "sample" (i.e., anchor) journal entries representing each performance level. One final parent requested a summary of the text the student read to better judge the meaning of their child's scores. While parents requested greater openness in the PBA's administration

procedures and scoring, they valued the openness in performance standards.

Joan also expressed satisfaction with the standardization of the scoring system as illustrated in the following interview turn:

> "Oh, I see a big change [from fall to spring]. Boy would that be depressing if you didn't. I mean, really, if you think about it. But you're using the same rubric, right, so the rubric hasn't changed. It's not more difficult so therefore you should be able to see progress."

Joan's reference to "the same rubric" and "hasn't changed" suggests that she sees the PBA as standardized (in terms of performance criteria) and believes this is of value to "see progress." And while Joan did not report any gaps in information availability from the established assessment program, this comment suggests that she found value in having access to assessment information more frequently (i.e., biannually) to evaluate progress across the school year.

In terms of literacy domain coverage and curriculum relevance (both classroom and school), Joan, June and the parents all expressed satisfaction with the PBA. When asked about what information the PBA provided, four parents explicitly stated that it supplied information about multiple literacy domains (e.g., "how well student understands text and expresses their interpretation in written and oral form"). In addition, six parents indicated literacy skills it did not address like oral reading, effort and interest level.

In terms of classroom-curriculum relevance, June explained that the

PBA was relevant to her classroom curriculum because it focused on the same "components" (e.g., artifacts, objectives), as illustrated in the following interview exchange:

June: Right...and then on a specific three-day week created this performance assessment to really specifically look at all these components that I'm collecting on a daily basis. When you look at a fall and then a spring how are they growing and how effective my instruction has been. So I think it would help to support it, and be used to summarize performance patterns across the class.

Tanja: So you could use it as a piece of a portfolio for example?

June: Yeah and I think at the school level too. Because our school, every grade level, second through fifth grade, is going, is trying Book Club, we all have the same goals that we're working toward you know. And then if we could establish a second, third, fourth, and fifth performance assessment it would help track how are the kids doing. What are the strong areas? What are the weak areas the next grade needs to really focus on? And how effective our instruction has been?

This exchange highlights June's belief that the PBA reflected her classroom curriculum. June's reference to reviewing "fall and then spring" assessment artifacts and information suggests that the PBA met her expectations for availability to judge the effectiveness of her instruction. The exchange also illustrates the potential June sees in the PBA for assessing school-level curriculum (i.e., "I think at the school level too"). Joan, the school administrator, validated this potential use by examining the performance rubrics and making the following comment:

"That's the curriculum? Everything's there. You've got it all.

Nothing's missing!"

Joan also reported what she saw as "great promise" in the PBA for providing progress information on a yearly basis. While Joan did not explicitly refer to availability as a dimensional gap in Highmeadow's program, she did express her belief that this assessment, expanded to target school-wide language arts curriculum goals, could demonstrate measurable change in student behavior, knowledge and skills, within and across grade levels (i.e., "the standardized performance assessment...could help us monitor students learning from fall to spring and from grades 1-5"). In contrast to Joan, most (23 of 26) parents reported that they would prefer information from the PBA more frequently than biannually (i.e., three times a year or more).

In terms of aggregation, Joan referred to the value of the PBA for providing individual-student and group-level scores as indicating by the following interview turn:

> "Well, other than individual student scores, seeing how they're doing, again, just a total...Because that's what a school looks like...But you can easily look at these scores...to see how females compared to males. How are different ethnic or gender groups doing?"

Finally, while students did not indicate any explicit gaps in Highmeadow's literacy assessment program, they did express satisfaction (and dissatisfaction) with particular properties of the PBA. For example, 20 students reported that the rubric helped them to understand how they could

improve and use it to set goals (e.g., "So you know what you need to improve on. By looking at the target"), suggesting a degree of valued openness in the performance criteria. In contrast to this satisfaction, 10 students reported that they did not value performance scores (e.g., "I have no need for numbers. They are just things you count with. I see no value in them").

Summary

In this chapter, I applied the framework generated in Chapter 5, to identify important dimensions and properties of assessments that consumers believed they needed, but did not have available (i.e., assessment gaps) from Highmeadow's literacy assessment progrm. I then evaluated the performance-based assessment in terms of its potential for addressing these gaps.

My analysis of assessment-program gaps (defined in terms of assessment properties) both within and across consumer groups suggests that the most critically needed assessment properties were cited by multiple consumer groups. These properties included standardization, multiple literacy domain coverage, classroom- and school-curriculum relevance, increased availability frequency, and score-based interpretation. Group and individual aggregation and openness were cited by select consumers as important, yet missing properties.

Through my analysis of desired, yet missing, assessment properties, I created an assessment tool profile that reflected the properties of a tool that

possessed the potential to fill the assessment gaps in Highmeadow's literacy assessment program. Because the PBA was designed to address assessment consumer needs, the profiled properties matched those of the PBA. Nevertheless, interview and survey data from consumers confirmed that the profiled properties were indeed instantiated in the PBA and that assessment consumers valued these properties for their assessment uses.

CHAPTER SEVEN

DISCUSSION

I conducted this study to explore the value of the assessments that constitute Highmeadow's literacy assessment program from the perspective of assessment consumers (i.e. social validity). I was particularly interested in how consumers used assessments and what dimensions they focused on when deciding to use assessments. In Chapter 4, I described Highmeadow's literacy assessment program in terms of its constituent tools and available information. The purpose of Chapter 4 was to provide a context for understanding assessment-tool use and value and to establish the evolution of Highmeadow's dual-system literacy assessment program as typical of the trend toward expanding, additive assessment programs in education. In Chapter 5, I analyzed patterns of assessment use both across and within consumer groups to evaluate the tools making up the program and identify the dimensions and properties of assessment tools valued by assessment consumers. In Chapter 6, I explored the value of a performance-based assessment in terms of its potential for meeting the assessment needs of consumers. Findings have both practical and theoretical implications. They can inform the integration of Highmeadow's literacy assessment program and, more generally, literacy assessment program design. The theoretical implications include how we study and evaluate the assessment tools and programs we develop.

Implications for Highmeadow's literacy assessment program

Highmeadow's program reflected an additive model of assessment implementation that resulted in information redundancies. Assessment tools were gradually added over time at different levels of the educational system by different groups of policy makers with little or no consideration of the established program as a whole. This mindless accumulation of assessment tools resulted in information redundancies. Some redundancies were well-justified, involving the use of information from a complex system of complementary tools for multiple purposes (e.g., June used report cards, portfolios, teacher evaluations, and parent-teacher conferences to evaluate and plan instruction). A few redundancies were unwarranted, involving the use of information from less-valued assessment tools to simply confirm evaluations made based on more-valued tools (e.g., Joan used information from the CTBS to confirm evaluations of school programs made based on MEAP scores). Redesigning Highmeadow's literacy assessment program to reduce unnecessary redundancies would result in a more integrated, and efficient assessment program.

Findings also support the assertion that assessment consumer groups value different assessment tools and information (Farr, 1992) for different uses. The school administrator valued standardized assessments, particularly the state-mandated MEAP, to address school-level program evaluation, planning and reporting uses. She also relied on information from the Botel

and standardized report cards to identify at-risk students. In contrast, the classroom teacher relied almost exclusively on classroom-based sources of assessment information. She used a complex, classroom portfolio system including student-generated artifacts, anecdotal records, and checklists, as well as parent-teacher conferences and standardized report cards to address classroom-curriculum and student-level evaluation, planning and reporting needs. Parents and students used classroom-based assessment information and test scores from the MEAP. While parents used this information to evaluate student progress and plan support of student learning, students used these tools to evaluate and plan their own learning. Thus, each consumer group used and valued assessment information and tools differently, supporting some degree of diversity in the tools constituting Highmeadow's literacy assessment program (e.g., the dual-assessment system).

The inclusion of most classroom-based assessment tools (i.e., portfolios, teacher evaluations) and the MEAP was justified by their multiple, important uses both across and within consumer groups. In contrast, the inclusion of other standardized assessments was clearly not justified because of the lack of important consumer use. The reading/writing archival portfolio, and the basal test, for example, were not used at all by consumers. Consequently, the inclusion of these assessment tools in Highmeadow's literacy assessment program should be reconsidered.

While the inclusion of these tools was not justified because of the lack

of use by assessment consumers, the value of other tools in the program was less clear. The CTBS and CogAT were reportedly used by Joan only to confirm information from the MEAP when evaluating the school reading program. Joan also was the only consumer to use information from the Botel. The Botel was used as one piece of information along with report cards in Joan's identification of at-risk students, but it was not clear if one was subordinate. These findings raise the following question: does limited use of assessment information justify a tool's inclusion in the assessment program?

One approach I used in the evaluation of limited-use tools was to explore the value of the tools' properties from the perspective of consumers. For example, Joan used information from the Botel because of its availability and standardization. The Botel was administered to all students every fall, providing scores which allowed Joan to evaluate student growth on a standard set of tasks and compare performance to a normative sample. Nevertheless, the Botel covered constrained literacy skills (within the domain of reading) and was not as relevant to the school curriculum as the report card (which was also used for many other purposes). Moveover, Joan referred to the Botel as "very old fashioned" which suggests some reservation in using information from this tool. Because Joan valued the standardization and availability of the Botel but not its constrained-nature or relevance, the selection (from the established assessment program if possible) or development of a broader, more relevant standardized assessment tool seems

appropriate.

Another approach to addressing this question would be to explore the uses of additional consumer groups (e.g., other teachers, policy makers). For example, while June failed to find a personal use for information from the CTBS and CogAT, she indicated that the teachers from the gifted and talented program used these tools to identify potential students for their program:

> "And what we use umm I don't really use the CogAT and the CTBS for anything...they are only used by the gifted and talented teaching staff to identify students for the gifted and talented program."

Further justification for the continued implementation of these test at Highmeadow comes from the interview with Joan:

> "For the district's purposes, I think it's a way to monitor, you know, student achievement across the board. You know, in the areas that the tests measure. It doesn't really match the curriculum as well, you know, so you can't really say it measures everything you teach. But the things it does measure, you know, they can use it for accountability or, you know, to monitor student achievement district wide."

These findings suggest that, while the consumers in this study did not use the CTBS and CogAT, additional consumer groups may have used information from these tools providing support for their continued inclusion. Findings also suggest a limitation of this study, the fact that all relevant consumer groups were not represented.

Findings on the performance-based assessment suggested that it has enormous potential for filling information gaps in Highmeadow's literacy assessment program. For example, Joan and June found the curriculum-oriented, performance-based assessment useful for evaluating curriculum effectiveness (in June's classroom). Joan believed the performance-based assessment would be most useful to her if its application was expanded. She saw prospects for the performance-based assessment's expanded use as a school-wide tool for use in judging curriculum-effectiveness, improving school programs, and identifying at-risk students (potentially taking the place of the Botel). These findings are not surprising given the fact that the assessment was designed with the needs of these assessment consumers in mind.

This social validity case study of Highmeadow's literacy assessment program provided insight on the value of assessment tools from the perspectives of the school administrator, a fifth-grade classroom teacher, her students, and their parents. Moveover, my analyses helped to identify highly-valued tools that were used by multiple consumers for important purposes (e.g., report cards), and those that were of limited utility (e.g., basal test). Analyses also resulted in a framework of eight value dimensions (i.e., authority, standardization, relevance, coverage, interpretation, aggregation, availability, openness), and each with two or three contrasting properties (e.g., classroom-, school-, and generalized-curriculum coverage) that assessment

consumers focus on when deciding to use information from a particular assessment tool for a specific use. Overall, these analyses provided evidence justifying the inclusion and exclusion of various assessment tools in Highmeadow's literacy assessment program. Finally, findings suggest that the performance-based assessment possesses great value for filling the gaps in the assessment program.

Implications for assessment program & performance-based assessment design

Findings indicate that no single assessment tool or type of assessment (e.g., standardized) will serve all the needs of any one consumer group, let alone multiple consumers. This conclusion supports the implementation of complex, literacy assessment programs which include multiple tools like the program at Highmeadow. Nevertheless, assessment rarely occurs without negative consequences (e.g., Paris, Lawton, Turner, & Roth, 1991). The simple, additive approach to assessment-program implementation which occurred at Highmeadow increased the amount of assessment and decreased the amount of time remaining for instructional activities. Thus, findings substantiate Farr's (1992) conclusion that "what is needed is an integrated [assessment] system" (p. 36). Assessment designers and policy makers must balance diverse demands of assessment consumers, with a concern for keeping assessment (and its potential negative consequences) to a minimum.

Because different assessment consumer groups have different assessment needs and use assessment information in different ways, it is

important to understand those needs before planning an integrated assessment program. The dimensional framework generated as part of this study describes the aspects of assessment tools and information which consumer groups pay attention to when deciding whether to use a particular assessment for a specific purpose. Different consumer groups focus on different dimensions. For example, while the school administrator focused on assessment standardization and interpretation, teachers, parents, and even students paid attention to availability. Further, different consumer groups valued different assessment properties, even when they focused on the same dimension. For example, the classroom teacher was interested in assessment information that was relevant to her classroom curriculum for planning instruction, yet parents valued school-curriculum relevance when evaluating the school.

These findings support the argument made by Farr (1992) that no single source or type of assessment information will serve the educational performance information needs of all consumer groups. Assessment designers and policy makers need to understand the assessment dimensions valued by each consumer group when developing assessment programs. Care should be taken to balance the values and needs of all consumer groups. Efforts to reduce the overall amount of assessment should not result in program designs that privilege specific assessment tools (or consumers). Determining what assessment information is available, how it is used and

valued, and what information is needed will facilitate the design of well-balanced assessment programs.

In terms of developing performance-based assessments, findings from this study suggest that these assessments can be designed to fill the gaps in assessment programs and support information already available. While standardized test scores were valued most by administrators and parents, and classroom-based assessments were valued most by teachers, students and parents, the performance-based assessment was valued by all consumer groups. In other words, the performance-based assessment was the *only* assessment tool valued by all consumers groups. This finding suggests that performance-based assessments have the potential to be highly-efficient tools for collecting useful information about student literacy performance. Thus, as suggested by Farr (1992), criterion-referenced (e.g., standards-based rubric) performance-based assessments like the one implemented here, might be the key linkage between consumer groups, potentially addressing needs ranging from accountability and comparability, to informing instruction and learning.

To effectively design these assessments, it is important to define the established assessment program including tools and types of information available as well as patterns of use by relevant consumer groups. These data should provide insight into the properties that ought to be instantiated in performance-based assessment tools that are widely valued by consumers. While performance-based assessments have great potential, their design,

implementation, and management require a strong commitment from administrators and teachers. As Farr (1992) stated, "the teachers who have been most successful in using this [performance assessment] approach have had the support of administrators who could see over the assessment wall. Their support generated public interest and support" (Farr, 1992; p. 34). Thus, assessment program designers who are considering the implementation of performance-based assessments should balance the challenges of designing and implementing this form of assessment with its potential value for meeting the needs of consumers.

Finally, findings support the use of social validity research to evaluate assessment needs and values toward the design of balanced, integrated school-wide literacy assessment programs. Through this approach, I was able to identify assessment redundancies, valued dimensions, and program gaps. Social validity data allowed me to make recommendations for program redesign with the goals of reducing assessment time and increasing the value of the program to assessment consumers. These data also contributed to the development of a highly-useful performance-based assessment.

Implications for validity research

This study also has implications for research on the validity of assessments. The term "validity" stems from the Latin root *valere* which means "worth" (Johnston, 1992). The worth, or validity, of assessments has historically been defined in terms of technical, psychometric criteria (e.g.,

correlations). An assessment had validity "if it measured what it purported to measure" (Allen, & Yen, 1979; p. 95). This constrained definition of validity stressed the value of assessments as scientific measurement.

In a broader sense, however, "validity is concerned with making sense of a situation" (Cherryholmes, 1988; p. 425). The construct of assessment validity has been expanded over time to include multiple perspectives on making sense of assessments. Assessment tools have been valued in terms of their technical soundness (e.g., reliability, criterion-related validity), their worth for informing theory (i.e., construct validity) and their impact on the educational system (i.e., consequential validity). While some validity researchers have suggested that the validity construct has become overburdened beyond its usefulness (e.g., Wiley, 1991; Cole & Moss, 1989), others have argued that "this expansion suggests a subtle change in what, exactly, the focus of validity research is" (Moss, 1992; p. 235).

While the technical, theoretical, and consequential validity lenses have provided guidelines for judging the value of assessment, they have been limited. First, these lenses highlight the values of assessment researchers and designers. Those consumers who actually use assessment information have not had a voice in evaluating and validating assessment worth. The agenda of the assessment research community has dictated what evidence counts toward making the case for an assessment's validity. Discourse about the value of assessments has been limited to the research community. Expert

recommendations for valuing and using assessments have been "disseminated" to guide assessment consumers.

Second, these validity lenses have failed to address the social context of assessment use. The validity of scientifically-defined assessment constructs and interpretations has been examined, but the social (e.g., historical, cultural, political) forces which impact assessment interpretation and use in real-life contexts have not been explored. The push to consider the validity of assessments in terms of their consequences recognizes the importance of assessment use in context by addressing the effect of assessment implementation on the attitudes and behaviors of consumers. This consequential lens, however, does not consider the opinions and values of consumers and their reciprocal impact on the design, implementation and use of assessments.

Finally, researchers who have studied assessment validity from these perspectives have discounted the "sense" that students, teachers, and parents make of assessment tools and the information disseminated to them. They have criticized assessment users for what they perceive to be "misinterpretations" or "misuses" of assessment information (Anastasi, 1986). They claim that these consumers do not understand the meaning of assessment data. While a lack of understanding may partially account for the problem of assessment misuse, the failure to recognize the values, needs, and interpretations of assessment consumers also contributes to the problem.

The social validity lens sensitized me to the importance of considering the values and needs of assessment users in the design, implementation and refinement of assessment programs. As Schwartz and Baer (1991) suggested, social validity evidence encourages consumer program use by anticipating potential reasons for rejection or misuse. Considering the voices of consumers in assessment design increases the likelihood that consumers will support the implementation and use of assessment program tools (Shepard & Bliem, 1995). Thus, social validity inquiry, by recognizing the voices of consumers, may discourage "misuse" of assessment information, encourage consumer program support, and attenuate the (potential) negative impact of assessments on teaching and learning.

The construct of social validity as introduced by behavior analysts, however, was limited to the measurement of consumer satisfaction. The present study expanded the social validity construct to include a genuine concern for the discourse of consumers and their understandings of educational assessment in context (Cherryholmes, 1988). If this perspective on social validity inquiry is recognized as a legitimate approach for exploring assessment validity, it would empower assessment consumers and ensure that evidence of "constructs" underlying assessments in use (i.e., the understandings of assessment users) were given as much consideration and priority as those defined by the scientific community.

Limitations and future directions

Some potential limitations of this study revolve around the nature of the data collected, and the consumers and assessment uses explored. First, because of the *subjective* nature of self-report data (e.g., interviews, surveys), they have often been cited as suspect (e.g., Winett, Moore, & Anderson, 1991) in the study of psychological and social phenomena. My heavy reliance on these data as a basis for my analyses may be perceived as a limitation of this study. In other words, are consumers' reported uses and perceptions of value, reflected in self-report data, as valid as evidence of actual use?

To remedy this perceived limitation, I triangulated self-report data whenever possible with observational data, a strategy suggested by qualitative researchers to validate research findings and conclusions (e.g., Bogden & Biklen, 1992). For example, most assessment uses reflected in consumer self-report data (e.g., interview, survey) were confirmed by my own observations of consumer use. Nevertheless, this was not possible with all uses, particularly those that did not involve action-based *observable* decisions (e.g., parents evaluating school programs). Wolf (1978) also suggests that as social validity researchers "we must establish the set of conditions under which people can be assumed to be the best evaluators of their own needs, preferences, and satisfaction" including, "education about options, lack of coercion, and anonymity" (Wolf, 1978; p. 221). Because consumers were informed about the purpose of the study, the study had little or no negative consequences for consumers, and the consumers were promised anonymity

by the requirements of human subject protection, all of these conditions were in place in the present study.

It is also critical to keep in mind that the failure of any consumer group to mention using a particular assessment in a specific way does not necessarily indicate that the assessment was not used or valued. For example, all consumer groups failed to mention the archival portfolio. While this finding may suggest consumers' failure to value and use this tool, it may just as likely reflect an oversight which should be explored further.

A second potential limitation of this study is the focus on a constrained set of consumers. This study provided strong evidence of the assessment values of Highmeadow's school principal, one fifth-grade teachers, her students and their parents. It did not, however, offer insight into the values of other administrators, students, teachers, or parents at Highmeadow or consumers at other schools. For example, the interviews with June and Joan (reflected in dialogue presented in the first section of this chapter) suggested that the gifted and talented teachers and district policy makers used the CTBS and CogAT for student identification and curriculum evaluation, respectively. Furthermore, the fifth-grade students in June's classroom were all average or above average achievers, making it risky to assume the same needs and values for low-achieving, special-needs students.

Another potential limitation is the primary focus on direct, rather than indirect uses of assessments. I have defined direct uses as evaluations or

actions that immediately follow from available assessment information. For example, Joan used information from the MEAP to directly evaluate and plan school curriculum. I did not explore indirect uses (actions that follow from policies resulting from direct uses) and the perceived value of assessment tools for these uses. For example, June reported receiving information from the MEAP as part of her participation in a school-wide student performance analysis. June reported that, while she did not directly use the information from this analysis, it did influence her classroom curriculum and instructional planning via school improvement goals, as suggested by the following statement:

> "Yes. And the MEAP scores what we have done as a school is we're really analyzing umm patterns and types of umm questions that students would miss on the MEAP to see if our instruction is lacking in some way. Are we not spending enough time on informational text you know that type of thing to kind of question our teaching. That's how we ended up emphasizing informational texts school wide and I use them a lot in my classroom."

June went on to say that she found MEAP information useful for this need because she felt the MEAP reflected her curriculum. These findings suggest that June did not directly use MEAP scores she received. Nevertheless, her involvement in the analysis of school-wide student performance patterns indirectly helped her to plan classroom curriculum and instruction,

providing additional support for the value of the MEAP.

Overall, this work supplies initial answers to questions concerning the social value of assessments from the perspective of assessment consumers. Future research should extend this work, addressing the values of additional consumer groups and indirect uses of assessment. For example, future research might focus on different grade levels (e.g., a first-grade classroom where not so much assessment is done but where identification of at-risk students is more of a priority) or achievement levels (e.g., low-achievement, special-needs students), and additional consumer groups (e.g., district-policy makers, gifted and talented teachers) to better understand the values and needs of literacy assessment program consumers.

Furthermore, future research might redefine assessment consumers as assessment "clients." The term consumer, historically used in the social validity literature, implies the user of a standardized, consumable good or product. The term client, on the other hand, suggests the receiver of a tailored and individualized, professional service. This shift in terminology might help to cultivate an approach to assessment program design which is built on an ongoing dialogue between professional designers, educational policy makers, and the assessment clients they serve.

Finally, to fully address the "social" validity of assessments, future research should focus on the historical evolution of assessment-consumer values and the political forces which impact these values. Studying

consumer values over an extended period of time (e.g., several years) would provide insight into how programs and perceived values and needs evolve and the forces that impact their development and change. This research could also begin to evaluate the consequences (i.e., consequential lens) of assessment programs (and performance-based assessments) for teaching and learning. More importantly, this research would further expand the construct of social validity beyond a concern for the phenomenological perspective of assessment users toward a consideration of the social forces (e.g., time, power) impacting consumer value development and change (Cherryholmes, 1988).

APPENDICES

APPENDIX A

APPENDIX A

Novel, Short Story, & Informational journal entries assigned scores of "3," "2" and "1"

Text types	ENTRIES ASSIGNED A SCORE OF 3	ENTRIES ASSIGNED A SCORE OF 2	ENTRIES ASSIGNED A SCORE OF 1
NOVEL	6/1/94 I don't think they are all monsters as Hannah said. They are forced to be there and they have no control over what happens there. They are forced by the Nazi monsters to work. They do not practically starve each other to death, the monsters do. They do not choose to be cremated. The monsters do that to them. How can Hannah think such a thing? She and the others have done nothing wrong. They are forced to work. Why? The monsters?	5/26/94 I did not like these two chapters because they didn't make sence [sic] to me. All of a sudden, Ellens [sic] parents were there, and I had no idea where they came from. I think the book is getting boring now because there is nothing exciting happening. In the casket I don't think there will be a person in it, I also think it was rude for that soldier to hit Annemaries [sic] mom in the face.	Now, I think all the people who came are Jews, and they're getting help with their escape. Still, this is very weird.
SHORT STORY	6/8/94 If I was Sadako, I would be really scared, and making those paper cranes would probably keep my spirits up. It would give me something to hope for, something to keep trying more for. I mean, if she didn't have those cranes, what would she do all day? Sit and rot in her hospital bed? I wish she could have made the thousand cranes so she could get well again. I wish the Americans had never dropped the A-bomb on Hiroshima, and none of this terrible disease stuff would never had happened	.June 8, 94 Sadako was a fighting, stubborn person who won't give up. She is very independent-like my friend Molly. If I were Sadako I wouldn't think about dieing [sic] so much and be gad [sic] I was alive at all. I would have made the cranes.	If I were in Sadako's place I would feel terrible! But I would spend most of my time either drawing or writing stories.

APPENDIX A (cont'd)

Text types	ENTRIES ASSIGNED A SCORE OF 3	ENTRIES ASSIGNED A SCORE OF 2	ENTRIES ASSIGNED A SCORE OF 1
INFOR-MATIO NAL TEXT	5/16/94 Before the End of May, 1940 Hitler's troops took over Czeckoslovakia [sic], Poland, Denmark, Norway, Belgium, Netherlands, and Luxembourg. Hitler is such a Pig! Even though he's trying to take over the world (and he came close in Europe) there is no way all people on earth would let him take over. Example: He invades Canada. Canada fights. The U.S. helps Canada fight. US gets Africa to help fight. Hitler stinks!	5/17/94 I used to think World War II was just a couple bombings here and there. I also thought he was just some big past military leader. Now I know that he also ruled Germany as he was their dictator. I was amazed when I found out that Jews were forced to work as slaves! Then I found out eventually, Hitler just started killing all the Jews!	The day pearl Harbor was bomed [sic] It happend in March 1939, Germany took over the rest of Czeckoslovakia [sic] the tiring six months of calm that followed the german [sic] quest of Poland that ended suddenly early in April, 1940. When Hitler's forces struck again, on June [sic] 9, 1940, the Norwagien [sic] Army surrenderd [sic] to the germans [sic]. Befor [sic] the end of May the Belgium surrenderd [sic] across the English channel, the British government orginzed [sic] everything that would float. Destroyers, minisweepes, tugboats, ferryboats, fishing boats, yahts [sic], dories, dinghies, and motor launches set out for Bunkirt 32 miles away. More than 330,000 troops escaped.

APPENDIX B

APPENDIX B

Fall 1994 Administrator Survey

Name_____ School_____ Date_____

The following survey was designed to help us, along with teachers, to design an assessment system that will document the literacy progress and achievement of students participating in the Book Club reading program. One of the goals is to create an assessment system that meets the needs of administrators, providing them with the information they want and need about students' academic progress. We would appreciate you taking a few minutes to fill out this brief survey and returning it in the enclosed envelop. Thank you for your time.

1. What do you believe constitutes academic progress in literacy for your students?

2. What aspects of students literacy progress (e.g., reading, writing, listening, speaking) do you want/need information about?

   a) Which ones do you believe are most important and why?

3. How often and what specific kinds of information do you need to talk about progress with the following:

   a) Policy makers?

   b) Parents?

   c) Teachers?

4. Briefly describe the current literacy assessment system in place in your school?

   a) How do you use results from the current system?

   b) What about the current system do you find MOST useful?

   c) What about the current system do you find LEAST useful?

5. **What aspects of literacy learning do you believe the current system taps and does not tap?**

   a) Do you believe the current system taps valuable literacy learning and why?

   b) How well do you believe the system taps the types of learning going on in your school's classrooms?

   c) What specific gaps do you see in the current assessment system?

6. How would you characterize alternative assessment?

a) What do you see as its strengths and weaknesses?

b) How do you see alternative assessment fitting into your current assessment system/program?

c) What skills do you believe teachers need to be successful with alternative assessment?

d) What do you see as your role in helping them obtain these skills?

PLEASE SIGN AND RETURN WITH YOUR COMPLETED SURVEY

I, _____ am willing_____ NOT willing_____ to participate further in the assessment project described in the attached letter. I can be contacted at_____.
(phone number)

_____
(administrator signature)

_____
(date)

APPENDIX C

APPENDIX C

Fall 1994 Teacher Survey

Name_____ School/Grade_____ Date_____

1. What do you believe constitutes academic progress in literacy for your students?

   a) What are the specific goals you have for student literacy learning in your classroom?

2. What aspects of students literacy progress do you want information about?

READING WRITING LISTENING SPEAKING (circle all that apply)

   a) Which aspects do you believe are most important and why?

   b) What type of information tells you if students are making progress in these areas?

3. What types of information on student learning do you currently use to make instructional decisions?

4. What types of information about your students' literacy progress would you like to have that you don't have available currently?

   a) For what purposes is this information needed?

5. How often and what information do you need to talk about progress with the following:

   a) Students?

   b) Parents?

   c) Administrators?

   d) Others?

6. Please describe the assessments (e.g., artifacts, tools) you currently use in your classroom.

   a) What features of these assessments do you find MOST useful?

   b) What features of these assessments do you find LEAST useful?

7. What literacy assessment system is currently used by your school (e.g., tests)?

   a) What aspects of literacy learning do you believe the current system taps and does not tap?

   b) Do you believe the aspects of literacy reflected on this assessment represent valuable learning and knowledge?

   c) How well do you believe the system taps the types of learning going on in your classroom and what gaps, if any, do you see?

8. How would you characterize alternative assessment?

a) Would you/do you like to use alternative assessment in your classroom?

YES SOMETIMES NO (circle one)

b) Why or why not?

If you answered NO, then stop here. If you answered YES or SOMETIMES, go on to questions 8b-8g.

b) How would/do you use information from alternative assessment (e.g., determine grades, direct instruction, report to parents)?

c) What benefits do you believe alternative assessment will/does have for students, teachers, parents, administrators?

d) How do you (plan to) determine if your assessments are valid/reliable?

e) How do you (plan to) do (e.g., manage, implement, use) alternative assessment in your classroom?

f) How do you (plan to) communicate the changes in your program to parents, students, colleagues, administrators?

g) How does alternative assessment fit into the current testing/assessment program you have in place in your school?

APPENDIX D

APPENDIX D

Fall 1994 Parent Survey

Child/Parent names_____ Date_____

The following survey was designed to help teachers develop an assessment system that will document the literacy progress of students participating in the Book Club reading program. One of the goals is to create an assessment system that meets the needs of parents, providing them with the information they want and need about students' academic progress. We would appreciate you taking a few minutes to fill out this brief survey and returning it to your child's teacher. Thank you for your time.

1. When you want to know how your child is doing in school in subjects such as reading, math, science, etc., what kind of information do you find most helpful to receive from the teacher and why?

2. What kinds of information tell you your child is doing well in reading and writing?

3. Would you like information about your child';s progress in areas like listening and speaking?

   YES NO

   a) Is it more important to have information about listening/speaking or reading and writing and why?

4. To have a good sense of how your child is doing in school, how often would you like feedback from the school and what kind of feedback do you want (be as specific as possible)?

5. In Michigan, we have a statewide tests called the Michigan Educational Assessment Program, or the MEAP. Do you think the MEAP provides you with useful information about your child's progress?

YES SOMETIMES NO (circle one)

If you circled SOMETIMES or YES, answer questions 5a. If you answered NO, move on to question 5b.

a) How do you use this information to understand your child's progress?

b) What else do you wish the MEAP would tell you?

6. Many schools are trying different ways of measuring students' progress and letting parents know how their children are doing. For example, some schools have asked students and teachers to keep collections of student work. Others have stopped using traditional report cards. Would you briefly describe any experiences you've had with newer forms of learning about your child's progress.

   a) Did you feel these were valuable experiences? Why or why not?

7. Would you like to be involved in documenting you child's progress?

   NO SOMETIMES YES (circle one)

8. Do you want to know how your child's classmates or school building is doing?

NO SOMETIMES YES (circle one)

If you answered SOMETIMES or YES go on to questions 8a and 8b. If you answered NO, then go on to question 9.

a) Why do you like to have this information?

b) How do you typically get this kind of information?

9. Please describe any additional kinds of information that you wish you had about your child's progress.

PLEASE SIGN AND RETURN WITH YOUR COMPLETED SURVEY

I, _____ am able_____ am NOT able_____ to participate
(parent name)
further in the assessment project described in the attached letter. I can be contacted at_____.
(phone number)

_____
(parent/guardian signature)

_____
(date)

APPENDIX E

APPENDIX E

Fall 1994 Student Survey

Name:_____ School:_____Date:_____

Instructions: The following survey was designed to help your teacher develop ways to find out what students have learned during Book Club. Your teacher also wants to collect information that will help students make decisions about their own learning. Your responses to these questions will help your teacher provide you with the information you need. Take a few minutes to fill out this brief survey. Answer each question as completely as possible. Then, return the completed survey to your teacher. Thank you for your time.

1. What makes someone a good reader?

a) What makes someone a good writer?

2. How do you know if you are getting better at reading?

a) How do you know if you are getting better at writing?

3. How does your teacher figure out if you are getting better at reading?

a) How does your teacher figure out if you are getting better at writing?

4. Have any of your teachers ever had you talk about books? YES NO

If you circled YES, answer questions 4a, 4b, &4c. If you answered NO, go on to question 5.

a) What was it like?

b) Did talking about books help you read better? YES NO

c) How did it help you?

5. Do you like to know how you're doing in school? YES NO

a) Why?

b) Who and what can help you get this information?

6. What do you want your parents to know about how you're doing?

a) Who and what can give them this information?

7. Have you ever taken a test called the MEAP? YES NO
If you circled YES, answer questions 7a-7e. If you answered NO, go on to question 8.

a) What was it like?

b) Did you find out how you did on the test? YES NO

c) How did you find out?

d) Did it help you learn about yourself as a reader? YES NO

e) What did you learn about yourself as a reader?

8. Would you like to know more about how you're doing in school?
YES NO

If you circled YES, answer questions 8a & 8b. If you answered NO, stop here.

a) What would you like to know more about?

b) Who and what could help you get this information?

APPENDIX F

APPENDIX F

Spring 1995 Parent Survey

Student Name:_____

Date:_____

Parent Name:_____

1. Review the three-point scoring scales attached to this survey. One scale is for evaluating student journal writing and the other is for evaluating student performance during literature discussions. These scales were developed by teachers at Highmeadow along with researchers at Michigan State University. They were created to provide a sense of how students are performing in the Book Club reading program which Mrs. F. uses in her classroom.

a) Would having scores such as these on your child tell you more about how your child is doing in reading?

YES NO (circle one)

Why or why not?

b) Do you believe these score descriptions represent fair expectations for your child?

YES NO (circle one)

Why or why not?

c) What would you change or add to these scoring scales to make them more useful to you?

2. Look at the score reporting form for your child. The scores provided are based on your child's journal writing performance at the beginning of the school year.

a) Would these scores be something you would like to have on your child?

YES NO (circle one)

b) How often and at what points during the year?

c) What information does this provide you about your child's progress (if any)?

d) Would these score be something to which you would want your child to have access?

YES NO (circle one)

Why or why not?

e) What other information would you like to have about your child's progress in reading that is not included in the described scoring systems?

APPENDIX G

APPENDIX G

Spring 1995 Student Discussion Survey

Student Name:_____ Date:_____

1. a) What makes someone a good reader in Mrs. F.'s classroom?

b) What makes someone a good writer?

2. a) Read the descriptions of the three discussion participation scores (1, 2, 3). Look at and listen to the discussions about Hatchet and acid rain your group had this fall. Using the score descriptions, score your participation in these discussions.

Hatchet Acid Rain

b) Briefly explain why you gave yourself these scores, using evidence from the discussions to defend your scores.

Hatchet:

Acid Rain:

3. Do you feel these discussions were good ones for you or did you usually participate better? (remember: this was the beginning of the school year when you first started Book Club).

4. a) Would you like to receive these scores on your discussions on a regular basis?

YES NO (circle one)

b) Explain why or why not.

5. a) Would receiving these scores from your teacher help you to contribute more to discussions or have better discussions?

YES NO (circle one)

b) Explain why or why not.

6. a) Would receiving these scores make you enjoy discussions more or less than you already do?

MORE LESS (circle one)

b) Explain why.

7. a) Think back to a recent Book Club discussion you had, have your discussions improved since the fall?

YES NO (circle one)

b) In what specific ways?

8. How could you improve your Book Club discussions? What specifically could you work on?

APPENDIX H

APPENDIX H

Spring 1995 Student Journal Entry Survey

Student Name:_____ Date:_____

1. a) Read the descriptions of the three journal scores (1, 2, 3). Look at the Hatchet journal entries for 9/27 and 9/29 and your acid rain entries for 10/7 and 10/10. Using the score descriptions, score each of the entries that you wrote at the beginning of the year.

Hatchet		Acid Rain	
9/27	9/29	10/7	10/10

b) Explain why you gave yourself these scores, using evidence from your journal entries to defend your scores.

9/27:

9/29:

10/7:

10/10:

2. Do you feel these entries were good ones or did you usually write better ones than these? (remember: this was the beginning of the school year when you first started writing journal entries).

3. Look at the scores assigned by teachers. Are they the same as the scores you gave yourself?

YES NO (circle one)

4. a) Do you believe the teachers' scores accurately reflect the quality of your entries?

YES NO (circle one)

b) If not, do you believe the scores are too high or too low?

TOO HIGH TOO LOW (circle one)

c) Why ?

5. a) Would you like to receive these scores on your entries on a regular basis?

YES NO (circle one)

b) Why or why not?

6. a) Would receiving these scores from your teacher help you learn more or write better journals?

YES NO (circle one)

b) Explain why or why not.

7. a) Would receiving these score make you enjoy writing journals more or less than you already do?

MORE LESS

b) Explain why.

8. Look at a recent journal entry. Score it using the same scale. Has your journal writing improved since the fall?

YES NO

9. How could you improve your journal writing? What specifically could you work on?

APPENDIX I

APPENDIX I

Fall 1994 Student Interview Protocol

The interview will cover four primary areas: (1) personal background, (2) curriculum/instruction, (3) literacy, and (4) assessment. There are a cluster of questions within each of the four areas to guide the interview, but the interviewer should feel free to obtain the information within categories in a way that feels like a more natural conversation.

Important things to keep in mind:

•The student should do the talking. Try to ask the question, then follow up to help the child expand his or her response. Questions such as "Can you tell me a little more about that?" "I'm not sure what you mean, can you give me an example?" "What else can you tell me about this?"

•Avoid the temptation to put words in the student's mouth. That means, when a student's response isn't very clear, it's tempting to rephrase their answer and ask if that's what they meant. Students tend to say "yes." It is critical to have their own words, so if something isn't clear, ask them the question using a different phrasing, or just tell them that you're confused and need some help understanding what they're saying.

•The interview should take from 15 - 30 minutes at the most. Knowing this can help you pace your questions. This means it's important to help students stay focused on the questions asked and not go off in other directions.

•Ask the student to say his name, school, and classroom into the tape recorder and play it back so he or she can hear how they sound. Make sure they are talking loudly enough to be picked up by the tape. Also, always watch to make sure the tape is moving and batteries are operating well.

•Begin the interview by introducing yourself, and telling the student that you're very interested in learning more about how students read and write and about ways that we can tell how we can help to make reading and writing instruction better. We are going to be talking with students in four different schools and s/he was selected by the teacher because she thought s/he might have interesting ideas about their reading and writing program and would enjoy talking with us about their ideas. This isn't a test and there aren't right or wrong answers. Any time a question is confusing, they should just ask you

to explain it more. If they get tired, all they have to do is tell you and you can stop the interview. End by saying you're looking forward to talking with them and that the chat is likely to take about 20 minutes.

Questions by Category

Personal Information

1. Can you tell me a little about yourself? How old you are? How long you've been at _____school? How do you like school? [*Note: these are primarily to help the interviewee become comfortable in the setting. Let him or her talk a few minutes guided by question such as the above.*]

2. Literacy in the home questions: Can you tell me the kinds of things you do at home with reading and writing? For example, do your parents read to you at home? Do you ever write letters or stories for members of your family? To friends you have from other places? [*note: we'd like information about things like the language(s) spoken in the home, siblings that the child might have, literacy practices in the home.*]

Children's Views about Literacy Curriculum

[***Within each of these questions/responses, try to elicit information from the students about how they feel about each of these, how much they value them, whether or not they see them as important things to learn?***]

3. Can you tell me what kind of reading you do in school?
   - favorite activities
   - stories read
   - typical activities

4. Can you tell me what kind of writing do you do in school?
   - favorite activities
   - stories read
   - typical activities

5. Can you tell me the kinds of things you do in school where you just talk about things -- like map work or math problems, like the books you've read

or the stories you've written? How do you do this?

6. Are there times you read or write or talk without the teacher being there? Are there times you talk about your reading and writing with your friends/peers in class?

Evaluation and Assessment

7. How do you know if you're doing a good job in reading? in writing? in talking about things in class?

8. How does the teacher figure out what it is that you've learned? How does she know if you need more help or if you've learned something really well?

9. **ACTIVITY:** **Ask students to bring to the session either something they've written or a book that they've read. Then ask the following:**

9a. Can you tell me what you think is particularly good about this [book/story]?

9b. Do you think this [book/story] has any problems that you think you (or the author) might change if you (s/he) were to work on it some more?

Probe to find out what sort of criteria they are using to judge the quality of the piece of work -- whether it is their own or something that has been published.

2. How do you know how you're doing in reading and writing--if you are improving? What tells you this? What do you look at? Who do you talk to? What do you think about?

3. What are the sections of your portfolio? What goes into your portfolios? Who decides what goes in them? What are they used for? How often do you look at them and why?

4. Do your parents see your portfolios? When? Do they like to see them? Why?

5. How would your score discussions? How would you score your journals?

Performance Assessment

1. How would you like it if your teacher gave you a "score" (like a 1, 2, 3) on your journal entries and discussions? Would it help you learn more?

2. If your teacher asked you to help develop scores, what would you say would make a good journal entry--what would it look like? What would make a good discussion--what would it look like, sound like?

3. Show students score rubric. Here is a scoring system that has been developed for scoring journal entries. Read the descriptions of each score. What do you think is good about these descriptions for journal entries? Discussions? What would you change?

4. Would receiving scores like this on your journal entries and discussions help you learn more? Why or why not?

5. Do you think your parents would like to see some scores like this? Why?

6. Who else do you think might like to see these scores? Why?

7. Using the scoring systems provided, guess how you think you would be scored on the journal entries you write in class. On your discussion participation.

Journals	Discussions

APPENDIX K

APPENDIX K

Spring 1995 Administrator Interview Protocol

Curriculum:

1. What do you know about June's reading curriculum? Do you feel it is consistent with the literacy goals for the school, why or why not?

Standardized assessment:

CTBS
CogAT
Basal Reading Test
MEAP--Reading

1. When are these tests administered? Are any others given and if so when?

2. How, specifically, is the information from these tests used? Who uses the information from these tests?

3. What are your school improvement goals regarding the MEAP? What other goals do you have and where do they come from?

Classroom assessment:

1. Are you familiar with June's system of classroom portfolio assessment? Is her system consistent with school goals? Why or why not?

2. Do you ever use information June collects? What information and how do you use it?

Performance-based assessment:
Describe the performance assessment, including: 1) two-day procedures, 2) journal and discussion activities, 3) two text types, 4) goals of the assessment and instruction, 5) holistic score descriptions.

1. Do you feel the performance assessment format and goals are consistent with good literacy instruction? How? In what ways? Explain.

2. What do you see as the strengths and weaknesses/limitations of the performance assessment? Relative to portfolio assessment? Relative to standardized tests?

3. Looking at the results for the students on their journal entries, how might/could YOU use these data? Do you believe these data would be of value to anyone (e.g., students, teachers, parents, other administrators)? Explain.

4. What do you believe might be the consequences, for you, teachers, or students, of using these data in the ways you have indicated above? Indicate both potential positive and negative impact and explain why you believe this would result.

5. This performance-based assessment has been approved to be given as a replacement for the end-of-the-year reading test in June's classroom? Who decided this? How will the information be used?

APPENDIX L

APPENDIX L

Winter/Spring 1995 Teacher Interview Protocol

1. Note that we're interested if they're trying to do anything different in their literacy programs (i.e., some special unit or something that is not a part of their regular literacy program that we may want to observe -- find out times for observations of such activities).

2. What are the different literacy activities you do in your classroom in terms of reading, writing, speaking, and listening?

topics covered
titles of books used
kinds of writing
how do you deal with skills?

3. What are your goals for your literacy program?

4 How do you decide what types of novels/basal stories you will use? Do you make connections between stories or across subjects? If so, what sorts of connections do you try to include?

5. RE: resources...where do you get materials you're using and how difficult is it to get materials?

6. What different kinds of formal and informal assessments do you use? Where do your ideas for assessment come from? What do you know about the different kinds of assessment (e.g., performance based, portfolio, etc.)?

7. How does assessment fit in with your curriculum? standardized tests? other forms of assessment?

8. What things have had a particular influence on the kinds of decisions you make in your teaching (e.g., books, classes, peers, workshops, etc.)?

9. What tools/information do you use to make decisions about whether students are learning? What does each tool tell you--how do you use them? If you could have any additional information, what would it be?

10. What information most impacts your instruction? How?

11. How do you determine a student's performance level in book club? How do you specifically adjust your instruction for students of different performance levels based on assessment data?

12. What are your primary target outcomes in the classroom? Are they pretty much the performance assessment standards? How has developing standards impacted your instruction?

13. How do you translate portfolio data into report card scores? Do they translate directly from your outcomes and tool? How?

14. Do you feel the performance assessment format and goals are consistent with your instruction? How? In what ways? Explain.

15. What do you see as the strengths and weaknesses/limitations of the performance assessment? Relative to portfolio assessment? Relative to standardized tests?

16. What results do you expect? Generate estimated scores for students. Will results look more like portfolio or standardized tests results? Why do you think this?

17. Looking at the results for the students on their journal entries, are they what you expected? How are they the same/different?

18. How might/could YOU, YOUR STUDENTS, PARENTS or YOUR SCHOOL ADMINISTRATOR(S) use these results? Do you believe these data would be of value to anyone? Explain.

19. What do you believe might be the consequences, on you and your students, of using these data in ways you have indicated above? Indicate both potential positive and negative impact and explain why you believe this would result.

APPENDIX M

APPENDIX M

Dimensions and properties of assessment tools

TOOL	DIMENSIONS	PROPERTIES
MEAP	Authority	External source
	Standardization	Standardized
	Relevance	School curriculum knowledge
	Coverage	Reading only
	Interpretation	Scores
	Aggregation	Groups & individuals
	Availability	Fall, 1/year
BOTEL	Authority	External source
	Standardization	Standardized
	Relevance	Generalized knowledge
	Coverage	Reading (related)
	Interpretation	Scores
	Aggregation	Individuals
	Availability	Fall, 1/year
CTBS	Authority	External source
	Standardization	Standardized
	Relevance	Generalized knowledge
	Coverage	Reading
	Interpretation	Scores
	Aggregation	Groups & individuals
	Availability	Spring, 1/year
CogAT	Authority	External source
	Standardization	Standardized
	Relevance	Generalized knowledge
	Coverage	Reading (related)
	Interpretation	Scores
	Aggregation	Groups & individuals
	Availability	Spring, 1/year
Basal Test	Authority	External source
	Standardization	Standardized
	Relevance	Generalized knowledge
	Coverage	Reading
	Interpretation	Scores
	Aggregation	Groups & individuals
	Availability	Spring, 1/year

APPENDIX M (cont'd)

Reading/writing archival portfolio	Authority Standardization Relevance Coverage Interpretation Aggregation Availability	External source Standardized School & classroom curriculum Multiple domains Descriptions Individuals Spring, 1/year
Report Card	Authority Standardization Relevance Coverage Interpretation Aggregation Availability	External source Standardized School & classroom curriculum Multiple domains Scores & descriptions Individuals >1/year
Parent-Teacher Conference	Authority Standardization Relevance Coverage Interpretation Aggregation Availability	Internal & external source Nonstandardized School & classroom curriculum Multiple domains Descriptions Individuals >1/year
Portfolios & other classroom/homework artifacts or behaviors	Authority Standardization Relevance Coverage Interpretation Aggregation Availability	Internal sources Nonstandardized Classroom curriculum Multiple domains Descriptions Individuals >1/year
Teacher written evaluations & informal communications	Authority Standardization Relevance Coverage Interpretation Aggregation Availability	Internal sources Nonstandardized Classroom curriculum Multiple domains Descriptions Individuals >1/year

LIST OF REFERENCES

LIST OF REFERENCES

Abruscato, J. (1993). Early results and tentative implications from the Vermont Portfolio Project. Phi Delta Kappan, 74, 474-477.

Allen, M. & Yen, W. (1979). Introduction to measurement theory. Monterey, CA: Brooks/Cole Publishing Company.

American Educational Research Association, National Council on Measurement in Education. (1955). Technical recommendations for achievement tests. Washington, DC: National Education Association.

American Psychological Association. (1954). Technical recommendations for psychological tests and diagnostic techniques. Psychological Bulletin, 51(2, Pt. 2).

American Psychological Association. (1966). Standards for educational and psychological tests and manuals. Washington, DC: Author.

American Psychological Association. (1974). Standards for educational and psychological testing. Washington, DC: Author.

American Psychological Association, American Educational Research Association, National Council on Measurement in Education. (1985). Standards for educational and psychological testing. Washington, DC: Author.

Anastasi, A. (1954). Psychological testing (1st ed.). New York: Macmillan.

Anastasi, A. (1961). Psychological testing (2nd ed.). New York: Macmillan.

Anastasi, A. (1986). Evolving concepts of test validation. Annual Review of Psychology, 37, 1-15.

Anastasi, A. (1993). A century of psychological testing: Origins, problems, and progress. In T. Fagan and G. VandenBos (Eds.), Exploring applied

psychology: origins and critical analyses. Washington, DC: American Psychological Association.

Baker, E., O'Neil, H., & Linn, R. (1993). Policy and validity prospects for performance-based assessment. American Psychologist, 48, 1210-1218.

Bisesi, T. (1993, December). Envisionment Building: Diverse learners constructing meaning during reading of narrative and expository texts. Paper presented at the annual meeting of the National Reading Conference, Charleston, SC.

Bisesi, T. (1996). Upper-elementary students' written responses to text: A holistic scoring rubric for evaluating journal entries. In D. Lue, C. Kinzer, and K. Hinchman (Eds.), Literacies for the 21st Century (pp. 76-87). Chicago: National Reading Conference.

Bisesi, T., Brenner, D., McVee, M., Pearson, P.D., & Sarroub, L. (in press). Assessment in literature-based reading programs: Have we kept our promises? In T. Raphael, and K. Au, Literature-based instruction: Transforming the curriculum. Norwood MA: Christopher Gordon.

Bisesi, T., & Raphael, T. (1997). Assessment in the Book Club program. In, S. McMahon, T. Raphael, T. Goatley and L. Pardo (Eds.), The Book Club Connection: Literacy learning and classroom talk (pp. 184-204). New York: Teachers College Press.

Bogden, R. & Biklen, S. (1992). Qualitative research for education: An introduction to theory and methods. Boston, MA: Allyn and Bacon.

Botel, M. (1970). Botel Reading Inventory. New York: Follett Publishing.

Campbell, D. (1960). Recommendations for APA test standards regarding construct, trait, or discriminant validity. American Psychologist, 15, 546-553.

Campbell, D. & Fiske, D. (1959). Convergent and discriminant validity in the multitrait-multimethod matrix. Psychological Bulletin, 56, 81-105.

Cherryholmes, C. (1988). Construct Validity and the Discourses of Research. American Journal of Education, 96, 421-457.

Coerr, E. (1977). Sadako and the Thousand Paper Cranes. New York: Dell Publishing.

Cole, N. & Moss, P. (1989). Bias in test use. In R.L. Linn (Ed.), Educational

Measurement (3rd ed., pp. 201-219). Washington, DC: American Council on Educational and National Council on Measurement in Education.

Cronbach, L. (1971). Test validation. In R. L. Thorndike (ed.) Educational Measurement (2nd ed., pp. 443-507).

Cronbach, L. (1988). Five perspectives on validity argument. In H. Wainer (Ed.), Test validity (pp. 3-17). Hillsdale, NJ: Erlbaum.

Cronbach, L. & Meehl, P. (1955). Construct validity in psychological tests. Psychological Bulletin, 52, 281-302.

Curtis, M. & Glaser, R. (1983). Reading theory and the assessment of reading achievement. Journal of Educational Measurement, 20(2), 133-147.

CTB/McGraw-Hill. (1989). Comprehensive test of basic skills (4th ed.). Monterey CA: Author.

Delandshere, G., & Petrosky, A. (1994). Capturing teachers' knowledge: Performance assessment and post-structuralism. Educational Researcher, 23(5), 11-18.

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

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

Frederiksen, J. and Collins, A. (1989). A systems approach to educational testing. Educational Researcher, 18(9), 27-32.

Freedman, S. (1979). How characteristics of student essays influence teachers' evaluations. Journal of Educational Psychology, 71, 328-338.

Freedman, S. (1993). Linking large-scale testing and classroom portfolio assessments of student writing. Educational Assessment 1(1), 27-52.

Gavelek, J. (1986). The social context of literacy and schooling: A developmental perspective. In T.E. Raphael (Ed.), The contexts of school-based literacy (pp. 3-26). New York: Random House.

Glaser, G. & Strauss, A. (1967). The discovery of grounded theory: Strategies for qualitative research. Chicago: Aldine.

Gloucester Press. (1987). Solutions, cleaner smoke, and acid trade. Issues, issues, issues: Acid rain (pp. 20-25). New York: Author.

Gulliksen, H. (1949). Intrinsic validity. American Psychologist, 5, 511-517.

Haertel, E. (1991). New forms of teacher assessment. Review of Research in Education, 17, 3-29.

Hallam, PJ (1995, November). Exploring emerging paradigms in reading assessment. Paper presented at the annual meeting of the National Reading Conference, New Orleans, LA.

Hawkins, R. (1991). Is social validity what we are interested in? Argument for a functional approach. Journal of Applied Behavior Analysis, 24, 205-213.

Johnston, P. (1989). Constructive evaluation and the improvement of teaching and learning. Teachers College Record, 90, 509-528.

Johnston, P. (1992). Constructive evaluation of literate activity. White Plains, NY: Longman Publishing Group.

Langer, J. (1990). Understanding literature. Language Arts, 67, 812-816.

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

Loevinger, J. (1957). Objective tests as instruments of psychological theory. Psychological Reports, 3, 635-694.

Lowry, L. (1989). Number the Stars. New York: Dell-Yearling.

McMahon, S., Raphael, T., Goatley V., & Pardo, L. (1997). The Book Club Connection: Literacy learning and classroom talk. New York: Teachers College Press.

Merriam, S. (1988). The case study research in education. San Francisco: Jossey-Bass.

Messick, S. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Messick, S. (1989a). Meaning and values of test validation: The science

and ethics of assessment. Educational Researcher, 18(2), 5-11.

Messick, S. (1989b). Validity. In R. Linn (Ed.), Educational Measurement (3rd ed., pp. 13-103). Washington, DC: American Council on Educational and National Council on Measurement in Education.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), 13-23.

Mosenthal, J., Lipson, M., Mekkelsen, J., Daniels, P., & Jiron, H. (1996). The meaning and use of portfolios in different literacy contexts: making sense of the Vermont Assessment Program. In D. Lue, C. Kinzer, and K. Hinchman (Eds.), Literacies for the 21st Century (pp. 113-123). Chicago: National Reading Conference.

Moss, P. (1992). Validity in Educational Measurement. Review of Educational Research, 62, 229-258.

Moss, P. (1994). Can there be validity without reliability? Educational Researcher, 23(2), 5-12.

Moss, P. (1996). Enlarging the dialogue in educational measurement: voices from interpretive research traditions. Educational Researcher, 25(1), 20-28, 43.

National Center for Education Statistics (1994). NAEP reading report card for the nation and the states. Washington DC: United States Department of Education.

Paris, S., Lawton, T., Turner, J., & Roth, J. (1991). A developmental perspective on standardized achievement testing. Educational Researcher, 20(5), 12-20.

Paulsen, G. (1988). Hatchet. New York: Puffin Books.

Pearson, P.D. (1997). Commentary. In S. McMahon, T. Raphael, V. Goatley and L. Pardo (Eds.), The Book Club Connection: Literacy learning and classroom talk (pp. 222-223). New York: Teachers College Press.

Pearson, P.D. & Garavaglia, D. (1997). Improving the information value of performance items in large scale assessments. Unpublished manuscript.

Raphael, T., Pardo, L., Highfield, K., & McMahon, S. (1997). Book Club: A Literature-based curriculum. Newton, MA: Small Planet Communications

Inc.

Raphael, T., Wallace, S., & Pardo, L. (1996, April). Paper presented at the annual meeting of the American Educational Research Association, New York.

Reiss, J. (1972). The Upstairs Room. New York: Harper and Row.

Resnick, D. (1982). History of educational testing. In A Wigdor & W. Garner (Eds.), Ability testing: Uses, consequences and controversies: Part II (pp. 173-194). Washington, DC: National Academy Press.

Rosenblatt, L. (1991). Literary Theory. In J. Flood, J. Jensen, D. Lapp, & J. Squire (Eds.), Handbook of Research on Teaching the English Language Arts. New York: MacMillan Publishing Company.

Routman, R. (1991). Invitations. Portsmouth, NH: Heinemann.

Schwandt, T. (1989). Recapturing moral discourse in evaluation. Educational Researcher, 18(8), 11-16.

Schwartz, I. & Baer, D. (1991). Social validity assessments: Is current practice state of the art? Journal Of Applied Behavior Analysis, 24, 189-204.

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

Shepard, L. (1991). Psychometrician's Beliefs about Learning. Educational Researcher, 20(6), 2-16.

Shepard, L. (1993). Evaluating test validity. Review of Research in Education, 19, 405-450.

Shepard, L. & Bliem, C. (1995). Parents' thinking about standardized tests and performance assessments. Educational Researcher, 24(8), 25-32.

Silver, Burdett, & Ginn (1993). Dream chasers skill progress tests. New York: Author.

Sizemore, J. (1988). Acid rain: the unsettled question. Cobblestone, (23) 37-40.

Smith, M. (1991). Put to the test: The effects of external testing on teachers. Educational Researcher, 20(5), 8-11.

Smith, N. (1965). American Reading Instruction. Newark, DE: International Reading Association.

Spradley, J. (1980). Participant observation. Orlando, FL: Harcourt Brace Jovanovich.

Stewart, R., Paradis, E., & Aegerter, J. (1992, December). Portfolios empowering teachers. Paper presented at the 42nd annual meeting of the National Reading Conference, San Antonio, TX.

Strauss, A. & Corbin, J. (1990). Basics of qualitative research: Grounded theory procedures and techniques. Newbury Park, CA: Sage.

Stiggins, R. (1987). Design and development of performance assessments. Educational Measurement: Issues and Practices, 6(3),33-42.

Stiggins, R. (1991). Facing the challenges of a new era of educational assessment. Applied measurement in education, 4(4), 263-273.

Sullivan, G. (1993). Aggression on the March. The day pearl harbor was bombed: A photo history of world war II. New York, NY: Scholastic.

Thorndike, R. & Hagen, E. (1986). Cognitive Abilities Test (form 4). Chicago: Riverside Publishing.

Tierney, R., Carter, M., & Desai, L. (1991). Portfolio Assessment in the Reading-Writing Classroom. Norwood, MA: Christopher-Gordon.

Valencia, S. (1993, December). Reliability and validity of literacy portfolios across classrooms. Paper presented at the annual meeting of the National Reading Conference, Charleston, SC

Valencia, S. (1990). A portfolio approach to classroom reading assessment: The whys, whats, and hows. The Reading Teacher, 43, 338-340.

Valencia, S., Hiebert, E., & Afflerbach, P. (1994). Authentic reading assessment: Practices and possibilities. Newark, DE: International Reading Association.

Vygotsky, L. (1978). Mind in society: The development of higher psychological processes. Cambridge: Harvard University Press.

Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a marxist theory of test construction.

Applied Measurement in Education, 6, 103-118.

Wells, G. & Chang-Wells, L. (1992). Constructing meaning together. Portsmouth, NH: Heinemann Educational Books.

Wertsch, J. (1985). Vygotsky and the social formation of mind. Cambridge: Harvard University Press.

Wiley, D. (1991). Test validity and invalidity reconsidered. In R.E. Snow & D. E. Wiley (Eds.), Improving inquiry in the social sciences: A volume in honor of Lee J. Cronbach. Hillsdale, NJ: Erlbaum.

Winett, T., Moore, J. & Anderson, E. (1991). Extending the concept of social validity: Behavior analysis for disease prevention and health promotion. Journal of Applied Behavior Analysis, 24, 205-213.

Wolf, M. (1978). Social validity: The case for subjective measurement or How applied behavior analysis is finding its heart. Journal of applied behavior analysis, 11, 203-214.

Yolen, J. (1990). The Devil's Arithmetic. New York: Puffin Books.