

TRUSTING TEACHERS' JUDGMENTS: A VALIDITY STUDY OF A CURRICULUM-EMBEDDED PERFORMANCE ASSESSMENT IN KINDERGARTEN—GRADE 3**Samuel J. Meisels²****Donna DiPrima Bickel³****Julie Nicholson²****Yange Xue²****Sally Atkins-Burnett²****ABSTRACT**

Teacher judgments of student learning are a key element in performance assessment. This study examines aspects of the validity of teacher judgments that are based on the Work Sampling System (WSS)—a curriculum-embedded, performance assessment for preschool (age 3) to Grade 5. The purpose of the study is to determine whether teacher judgments about student learning in kindergarten—third grade are trustworthy if they are informed by a curriculum-embedded performance assessment. A cross-sectional sample composed of 345 K—3 students enrolled in 17 classrooms in an urban school system was studied. Analyses included correlations between WSS and an individually-administered psychoeducational battery, four-step hierarchical regressions to examine the variance in students' spring outcome scores, and Receiver-Operating-Curve (ROC) characteristics to compare the accuracy of WSS in categorizing students in terms of the outcome. Results demonstrate that WSS correlates well with a standardized, individually administered psychoeducational battery; that it is a reliable predictor of achievement ratings in Kindergarten—Grade 3; and that the data obtained from WSS have significant utility for discriminating accurately between children who are at-risk (e.g., Title I) and those not at-risk. Further discussion concerns the role of teacher judgment in assessing student learning and achievement.

¹ We acknowledge the invaluable assistance of Sandi Koebler of the University of Pittsburgh and Carolyn Burns of the University of Michigan in collecting and coding these data and Jack Garrow for assisting us with school district data. We are also deeply grateful to the principals, teachers, parents, and children who participated in this study, and to the staff and administrators of the Pittsburgh Public Schools. This study was supported by a grant from the School Restructuring Evaluation Project, University of Pittsburgh, the Heinz Endowments, and the Grable and Mellon Foundations. The views expressed in this paper are those of the authors and do not necessarily represent the positions of these organizations. Dr. Meisels is associated with Rebus Inc, the publisher, distributor, and source of professional development for the Work Sampling System[®]. Corresponding author: Samuel J. Meisels, School of Education, University of Michigan, Ann Arbor, MI 48109-1259; smeisels@umich.edu.

² University of Michigan

³ University of Pittsburgh

TRUSTING TEACHERS' JUDGMENTS: A VALIDITY STUDY OF A CURRICULUM-EMBEDDED PERFORMANCE ASSESSMENT IN KINDERGARTEN—GRADE 3

Assessments that rely on teacher judgments of students' academic performance are used widely in both research and applied settings. In research settings, they contribute to evaluations of intervention studies, classroom processes, and children's intellectual, socio-emotional, and behavioral development (Hoge, 1983). In applied settings, teachers rely at least as often on their own judgments as they do on more conventional standardized measures in evaluating students' achievements, planning instruction, and reporting to parents (McTighe & Farrara, 1998; Popham, 1996; Sharpley & Edgar, 1986; Stiggins, 1987, 1998). Teachers' judgments are also used for screening and diagnostic decisions about referrals and special placements for individual students (Hoge, 1984). Even district and state level assessments rely increasingly on teacher observation and judgment as means of evaluating students' performances in such areas as writing, science, and visual or performing arts (Baron & Wolf, 1996; Mills, 1996; Stiggins, 1987).

Some researchers argue that teachers can be valid assessors of their students (see Perry & Meisels, 1996, for a review of this issue). They claim that since teachers observe and interact with students on a daily basis, they are in the best position to evaluate their students' intellectual, socio-emotional, and behavioral accomplishments (Calfée & Hiebert, 1991; Hopkins, George, & Williams, 1985; Kenny & Chekaluk, 1993). Others express concerns about the trustworthiness and consistency (i.e., validity and reliability) of these assessments (Hoge & Coladarci, 1989). Specifically, they question whether teachers have sufficient knowledge about the domains that are tested and the tasks they are asked to judge. Also questioned are teachers' abilities to discriminate such constructs as achievement and motivation, and such individual differences as low achievement and specific learning disabilities (Hoge & Butcher, 1984; Salvesen & Undheim, 1994). Another area of concern is the subjectivity inherent in teachers' judgments (Silverstein, Brownlee, Legutki, & MacMillan, 1983) and the extent to which teachers' expectations and biases may influence student outcomes (Hoge, 1983; Hoge & Butcher, 1984; Sharpley & Edgar, 1986). Given these concerns, and their implications for students, it is reasonable to ask, "Can we trust teachers' judgments of student performance?"

Curriculum-embedded performance assessment (often called authentic assessment) is heavily reliant on teacher judgment. Such assessments are defined as integrated parts of the learning experience that differ from on-demand assessments which are often external to the classroom. Curriculum-embedded performance assessments are integrated into the day to day curriculum and instructional activities of a classroom. They consist of real instances of extended criterion performances, rather than proxies or estimators of actual learning goals (Shepard, 1991, p. 21). In contrast, on-demand assessments—whether performance-based or not—are not necessarily drawn from the actual repertoire of the

classroom, nor do they always occur during the process of teaching and learning. They call for students to perform at a certain place and time and in a certain way.

The literature on performance assessment has been growing for more than a decade (see Baker, O Neil, & Linn, 1993; Darling-Hammond & Aness, 1996; Fredricksen & Collins, 1989; Herman, Aschbacher, & Winters, 1992; Linn, 2000; Linn, Baker, & Dunbar, 1991; Meisels, Dorfman, & Steele, 1995; Resnick & Resnick, 1992; Wiggins, 1989; Wolf, Bixby, Glen, & Gardner, 1991). Researchers and practitioners alike have demonstrated a significant interest in this approach to evaluating student learning. Although extensive empirical support for the accuracy of teachers' judgments based on performance assessments is unavailable, evidence (derived largely from on-demand assessments) is continuing to build concerning performance assessments' potential to improve teaching and learning (Borko, Flory, & Cumbo, 1993; Darling-Hammond, 1994; Falk & Darling-Hammond, 1993; Gearhart, Herman, Baker, & Whittaker, 1993; Khattri, Kane, & Reeve, 1995; Linn, 1993; Meisels, 1997; Meisels, Dorfman, & Steele, 1994; Smith, Noble, Cabay, Heinecke, Junker, & Saffron, 1994; Taylor, 1994). Use of performance assessments in Vermont (Koretz, Stecher, Klein, McCaffrey, & Diebert, 1994), Kentucky (Kentucky Institute for Education Research, 1995), and Maryland (Almasi, Afflerbach, Guthrie, & Schafer, 1995; Koretz, Mitchell, Barron, & Keith, 1996) resulted in positive changes in instruction. Other researchers also reported that performance assessments led to better curricular integration (Khattri, Kane, & Reeve, 1995), positive changes in instructional practices (Stecher & Mitchell, 1996), greater use of higher-order thinking skills in instruction (Koretz, Stecher, Klein, & McCaffrey, 1994), and more awareness of and emphasis on individual growth and development in classrooms (Aschbacher, 1993).

Many of those who support performance assessment view it as a potential remedy for some of the frequently reported abuses of conventional, standardized, group-administered tests (Gardner, 1993; Wiggins, 1993). These and other authors (see Corbett & Wilson, 1991; Sternberg, 1996; Sykes & Elmore, 1989) point out that norm-referenced, achievement tests can be used to establish a system in which indicators of learning overwhelm attention to learning itself. Such tests, particularly when they are institutionalized in high-stakes district or state tests tend to draw attention primarily to what is measured, neglecting those elements of the curriculum that are not measured. They encourage a standardized pedagogy for use with a non-standard, diverse student population; offer few rewards for innovation or risk taking on the part of teachers or students; and distort the motivational climate for teaching and learning.

However, the promises of performance assessments have not always been realized and have rarely been documented empirically. Some have called into question the gains reported on performance assessments by certain states (e.g., Kentucky) (Green, 1998) and others have suggested that performance assessments actually work against some reform goals, such as constructivist approaches to teaching and learning (Murphy, Bergamini, & Rooney, 1997).

Critics (e.g., Mehrens, 1998) charge performance assessment with responsibility for narrowing the curriculum and decreasing the effectiveness of instruction (Khattri et al., 1995; Murphy, Begamini, & Rooney, 1997). Other commonly cited problems include inadequate reliability (Linn, 1994; but see Moss, 1994, for a response to this criticism), limited generalizability across tasks (Shavelson, Baxter, & Gao, 1993), the potential to widen achievement gaps (Linn, Baker, & Dunbar, 1991), and the cost and extensive time required to train teachers to administer and score such assessments (Cizek, 1991; U.S. General Accounting Office, 1993). Clearly, more evidence is needed to verify the claims made on behalf of curriculum-embedded performance assessments.

The present study contributes to the establishment of a research base for performance assessment by examining evidence about the relationship of curriculum-embedded performance assessment to other key indicators of student achievement. Its purpose is to investigate the validity of the Work Sampling System (WSS; Meisels, Jablon, Marsden, Dichtelmiller, & Dorfman, 1994), a performance assessment for preschool (3-year-olds)—Grade 5, by determining whether teacher judgments about student learning are trustworthy when those judgments are based on this curriculum-embedded performance assessment. Previous research on WSS (Meisels, Liaw, Dorfman, & Nelson, 1995) was limited to a cohort of 100 kindergarten children who were administered the field-trial version of the assessment. High internal reliability on the WSS checklists (Cronbach alphas ranging from .84 to .95), and moderate inter-rater reliability on the WSS Summary Report (zero order correlations between two external raters and 10 teachers of .68 and .73, $p < .001$) were reported in this study. Moderate to high correlations were also obtained between the fall WSS checklist and psychoeducational assessments given in the fall ($r = .74$) and spring ($r = .66$). Two-step hierarchical regressions demonstrated significant contributions of the fall WSS checklist to predictions of children's performance in the spring, even when the potential effects of gender, maturation (age), and initial ability (fall test scores) were controlled.

The cross-sectional, psychometric investigation presented here extends the previous study and represents the first investigation ever conducted of a curriculum-embedded performance assessment in the early elementary grades. Although many aspects of the validity of performance assessments besides their relationship to external criteria are important to consider (see Baker, O Neil, & Linn, 1993; Frederiksen & Collins, 1989; Linn, Baker & Dunbar, 1991), a design that demonstrates evidence for the validity of curriculum-embedded performance assessment and the trustworthiness of teacher judgments is a key ingredient in demonstrating to practitioners and policy makers the accuracy and practicality of their use.

SAMPLE, METHODS, AND PROCEDURES

This report is part of a larger study of WSS that collected data from students, parents, and teachers, using multiple means of measurement. Information is presented here

concerning the direct assessment of children, focusing primarily on validity evidence regarding the relationship of WSS to other achievement variables (for a discussion of additional aspects of validity related to performance assessment see Baker, O'Neill, & Linn, 1993 and Moss, 1992, 1996). Other studies will present additional validity evidence, including analyses of consequential aspects of validity based on extensive interviews with teachers (see Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 1998, for a preliminary report) and studies of parent reactions to the use of WSS (Meisels, Xue, Bickel, Nicholson, & Atkins-Burnett, in press).

Sample

The teachers (N = 17) in the WSS schools were all voluntary participants. Selection criteria for participation included 1) at least two years experience using WSS, 2) a rating within the highest quartile of teacher participants, based on a review conducted in the spring of 1996 by external examiners of WSS portfolios, and 3) a determination by the research team that the teachers' 1996-97 WSS materials were completed competently. These criteria contributed to our confidence in the fidelity of teachers' implementation of WSS and enabled us to focus on variability in children's learning rather than variability of implementation. All of the teachers in the sample were female. Thirteen percent were African American and 77% were Caucasian. Nearly half (47%) had completed a Master's degree and had more than 10 years' teaching experience.

The study presents cross-sectional data concerning students who were enrolled in kindergarten—Grade Three in five schools located in the Pittsburgh (PA) Public Schools (PPS). At the time the study took place (1996—97) WSS had been implemented in these schools for three years. The student study sample was composed of 345 children, all of whom were enrolled in WSS schools. Table 1 presents the demographic characteristics of these students. Most of them were African-American (69.9%) and received free or reduced lunch (79.4%). Gender was distributed fairly evenly (Male = 48.7%) and only a small number of children were classified as children with special needs (7.8%).

TABLE 1. WSS CROSS-SECTIONAL SAMPLE CHARACTERISTICS¹

Demographic Variable	Kindergarten (N = 75)	First Grade (N = 85)	Second Grade (N = 91)	Third Grade (N = 94)
Age range (in months)	57–73	69–94	80–104	93–120
Ethnicity	67%–AA* (50) 21%–White (16) 1%–Hispanic (1) 7%–Asian (5) 4%–Other (3)	60%–AA (51) 35%–White (30) 2%–Asian (2) 2%–Other (2)	62%–AA (56) 36%–White (33) 1%–Asian (1)	89%–AA (84) 10%–White (9) 1%–Other (1)
Disability Classification**	3%–LD (2) 1%–MH (1)	1%–LD (1) 4%–MH (3) 2%–OH (2)	3%–LD (3) 7%–MH (6) 1%–BD (1) 2%–OH (2)	5%–LD (5) 1%–MH (1)
Lunch Status	57%–Free (43) 3%–Reduced (2) 37%–Regular (28)	86%–Free (73) 2%–Reduced (2) 8%–Regular (7)	84%–Free (76) 3%–Reduced (3) 12%–Regular (11)	77%–Free (72) 3%–Reduced (3) 12%–Regular (11)
Sex	57%–Male (43) 43%–Female (32)	47%–Male (40) 53%–Female (45)	49%–Male (45) 51%–Female (46)	43%–Male (40) 57%–Female (54)

¹ Due to missing data, not all percentages total to 100.

* AA = African American

** LD = Learning Disabled, MH = Mentally Handicapped, BD = Behavior Disordered, OH = Other Handicap

Measures

Work Sampling System. WSS is a low-stakes, curriculum-embedded performance assessment whose primary purpose is instructional assessment. It is not designed to rank and compare students or to be used for high stakes decision-making. Rather, its value is linked to its impact on instruction. It is intended to clarify what students are learning and have begun to master by providing information relevant to understanding individual students' academic, personal and social, and other accomplishments. Accordingly, it guides instructional decision-making and provides instructionally relevant information to teachers that can be used to enhance teaching and improve learning. Extensive professional development is available to teach teachers how to use WSS, and many states and school districts have adopted WSS for use in the early years of school.

WSS is virtually unique in terms of its multi-dimensionality. It uses three forms of documentation: checklists, portfolios, and summary reports (see Dichtelmiller, Jablon, Dorfman, Marsden, & Meisels, 1997; Meisels, 1996; 1997). Checklists for each grade (preschool-fifth) list specific classroom activities and learner-centered expectations that were derived from national and state curriculum standards. The checklists consist of items (K = 67, first grade = 74, second and third grades = 75) that measure seven domains of development: *personal and social* (self concept, self control, approach to learning, interactions with others, conflict resolution), *language and literacy* (listening, speaking, literature and reading, writing, spelling), *mathematical thinking* (patterns, number concepts and operations, geometry and spatial relations, measurement, probability and statistics), *scientific thinking* (observing, investigating, questioning, predicting, explaining, forming conclusions), *social studies* (self, family, community, interdependence, rights and responsibilities, environment, the past), *the arts* (expression and representation, appreciation), and *physical development* (gross and fine motor, health and safety). For this study, only language and literacy and mathematical thinking ratings are reported. This is because these areas are assessed most adequately on the outcome measure we selected; they are the academic areas of greatest interest to policy makers; and many school districts implement only these two domains plus personal and social development.

Every skill, behavior, or accomplishment included on the checklist is presented in the form of a one-sentence performance indicator (for example, *Follows directions that involve a series of actions*) and is designed to help teachers document each student's performance. Accompanying each checklist are detailed developmental guidelines. These content standards present the rationale for each performance indicator and briefly outline reasonable expectations for children of that age. Examples show several ways children might demonstrate the skill or accomplishment represented by the indicator. The guidelines promote consistency of interpretation and evaluation among different teachers, children, and schools.

Portfolios illustrate students' efforts, progress, and achievements in a highly organized and structured way. Work Sampling portfolios include two types of work (core items and individualized items) that exemplify how a child functions in specific areas of learning throughout the year in five domains—language and literacy, mathematical thinking, scientific thinking, social studies, and the arts. Portfolio items are produced in the context of classroom activities. They not only shed light on qualitative differences among different students' work; they also enable children to take an active role in evaluating their own work.

The summary report replaces conventional report cards as a means of informing parents and recording student progress for teachers and administrators. The summary report ratings are based on information recorded on the checklists, materials collected for the portfolio, and teachers' judgments about the child's progress across all seven domains. Teachers complete the reports three times per year, completing brief rating scales and

writing a narrative about their judgments. The report is available in both hard copy and electronic versions. By translating the information documented on the checklists and in the portfolios into easily understandable evaluations for students, families, administrators, and others, this report facilitates the summarization of student performance and progress and permits this instructional evidence to be aggregated and analyzed. Examples of all WSS materials are available on line at www.rebusinc.com.

Teachers using WSS rate students' performance on each item of the checklist in comparison with national standards for children of the same grade in the fall, winter, and spring. They use a modified mastery scale: 1 = Not Yet, 2 = In Process, or 3 = Proficient. In the fall, winter, and spring, teachers also complete the hand-written or electronic summary report on which they summarize each child's performance in the seven domains, rating their achievement within a domain as 1 = As Expected, or 2 = Needs Development. Teachers rate students' progress separately from performance on the Summary Report as 1 = As Expected or 2 = Other Than Expected (distinguished as below expectations or above expectations), in comparison with the student's past performance.

Subscale scores for the checklist were created by computing the mean score for all items within a particular domain (i.e., language and literacy or mathematical thinking). Subscale scores for the summary report were created by computing a mean for a combination of three scores: students' checklist and portfolio performance ratings, and ratings of student progress. Missing data in the teachers' WSS ratings were addressed by using mean scores instead of summing teachers' ratings when computing the subscale scores.

Woodcock Johnson Psychoeducational Battery-Revised. The achievement battery of the Woodcock-Johnson Psychoeducational Battery-Revised (WJ-R; Woodcock & Johnson, 1989) is an individually-administered achievement test that was normed on a population of 6,359 individuals chosen in a random stratified sample procedure. Nine subtests were administered in this study: letter word identification, passage comprehension, dictation, writing sample, applied problems, calculation, and science and social studies (results of science and social studies are not described in this report). We will report several WJ-R cluster scores including broad reading (combining letter-word identification and passage comprehension), broad written language (dictation and writing samples), broad math (applied problems and calculation), skills (letter-word identification, applied problems, and dictation), and language and literacy (standard scores in letter word identification and dictation for kindergartners, and broad reading and broad written language standard scores for first through third graders). All WJ-R scores discussed in this report represent standard scores (versus raw scores) and were computed on software supplied by the test manufacturer (Compuscore) using grade level norms. Because the WJ-R is a very different type of assessment from WSS it introduces method variance into all analyses. However, it was selected because no other performance assessment comparable to WSS exists that could be used as an external criterion (completing two different performance assessments would be impractical in any event). The WJ-R is comprehensive, well-researched, and

covers the two principal areas of academic achievement focused on by this study. Moreover, as an individually-administered assessment, it is clinically more sensitive than conventional group-administered tests.

Procedures and Analyses

The 17 study teachers implemented WSS throughout the 1996-1997 school year by completing checklists on three occasions (fall, winter, and spring), continuously collecting material for the portfolios, and preparing a summary report for the fall, winter, and spring reporting periods. The WJ-R was administered twice—in October/November and in April/May. All examiners received training on the administration of the WJ-R prior to the fall testing period and a follow-up review of administration procedures before the spring testing dates. Examiners were blind to the study's purposes.

Three analyses were conducted with the cross-sectional data using teachers' WSS ratings of student achievement and students' WJ-R standard scores: a) correlations comparing the students' standard scores on the various subtests of the WJ-R and the WSS checklist and summary report ratings of student achievement within the corresponding WSS domains, b) four-step hierarchical regressions examining the different factors that accounted for the variance in students' spring WJ-R scores, and c) Receiver-Operating-Characteristic (ROC) curves, which make possible a determination of whether a random pair of average and below-average scores on the WJ-R would be ranked correctly in terms of performance on the WSS. Descriptions of each of these analyses follow.

Evidence of concurrent aspects of WSS's validity was examined by computing correlations between WSS subscale scores and students' WJ-R standard subtest and broad scores to show the amount of shared variance between the two assessments. Correlations of .70 to .75 are considered optimal because they indicate a substantial overlap between the two assessments, yet also recognize that each instrument contributes independently to the assessment of students' learning. If correlations are high, that is, $\geq .80$, more than half of the variance between WSS and the WJ-R is shared and an argument can be made that the predictor (in this case, WSS) does not add sufficient new information to justify its use. Conversely, low correlations ($\leq .30$) suggest very little overlap between WSS and more conventional achievement tests, thus raising the question of what exactly is measured by the predictor.

Four-step hierarchical regression analyses were used to determine whether the WSS checklist and summary report ratings made a unique contribution to the children's performance on the WJ-R over and above the effects of children's gender, age, socioeconomic status (as represented by free and reduced lunch versus regular lunch status), ethnicity, and initial performance level on the WJ-R. The demographic variables (gender, age, socioeconomic status, ethnicity) were entered in the first step of the four-step model. The WSS checklist was entered in the second step and the summary report

was added third. In the final step, children's initial performance level (fall WJ-R standard scores) was entered. The increment in the variance explained was noted for each step in order to assess the contribution of WSS and initial performance level above and beyond the demographic factors.

Receiver-Operating-Characteristic curve (ROC curve) analysis was conducted in order to study the utility of using WSS to classify students in need of supportive educational services (e.g., Title I). ROC data enable investigators to examine whether two different assessments will assign students to the same or different categories (ROC percentages $\geq .80$ are considered excellent). To accomplish this we established cutoffs for the WJ-R and performed a cost matrix analysis to obtain optimal cut-offs for WSS. The WJ-R is commonly used in clinical applications with children suspected of having learning disabilities or other problems that might affect their academic success. This analysis enabled us to determine the probability that WSS ratings can be used accurately to assign students to a high risk or low risk group.

Missing Data

Sample sizes vary in the cross-sectional study, ranging from 75 to 94 per grade, due to several factors. These include children whose families changed residences between fall and spring, incomplete WSS records (both fall and spring checklists and summary reports were required in order for a child to be included in the analyses), and examiner variability in the administration of the WJ-R. A small number of examiners did not obtain a ceiling score for all children administered the WJ-R. In order to standardize the administration, the ceiling rule was lowered by one point and all test protocols were rescored and rechecked. This modification had the effect of foreshortening the range of student responses, thus making the WJ-R results in this study a more conservative estimate of performance than if the standard six item ceiling rule were in use. Students were dropped from the analyses when a ceiling rule of five could not be obtained from the rescoring.

In order to study the impact of the missing data on our conclusions we combined the missing data into two groups: a) students whose WJ-R data were excluded from analysis due to examiner variability (Group 1), and b) students who moved and/or had missing WSS or WJ-R data (Group 2). Analyses were completed to determine whether there were systematic differences between Groups 1 and 2 and the final total sample which is described in Table 1. For second and third graders, boys were over-represented in Group 2; otherwise, there were no gender differences between the groups. In a few cases, the small number of children included in Group 1 prevented the use of statistical procedures to compare this group with the larger sample. For all analyses, no systematic differences were found between the sample of children whose data were dropped due to variability in test administration (Group 1) and the final total sample. Therefore, the relatively small numbers in Group 1, and the lack of differences between Group 1 and the total sample suggest that the study's findings were not affected when we dropped some

students due to differences in WJ-R administration. No differences were found in kindergarten, first, or second grade between Group 2 (those students who moved and/or had missing WSS data) and the final sample. However, for third graders, Group 2 had lower WJ-R scores on all literacy subtests, and on calculation and broad math. Thus, except for third grade, where the final sample performed above Group 2, there are no effects on the findings due to the missing data.

RESULTS

This study was designed to describe the cross-sectional academic achievements of four separate grade level samples of children throughout the course of one school year. Although comparisons are useful, it is important to recognize that these four grade level samples may differ from each other in systematic ways that are not captured by our analyses (e.g., retention history, age of entry into school, curriculum exposure).

Correlations Between WSS Ratings and WJ-R Standard Scores

Table 2 displays correlations for all WJ-R subtests and cluster scores with WSS checklist and summary report ratings across the four grade levels. This table enables us to begin to examine the concurrent aspects of WSS's validity, that is, how WSS teacher ratings correlate with students' standard scores using grade norms on an individually administered standardized achievement test. Over three-quarters of the correlations listed in Table 2 are within the range of .50 to .75. Further, 48 of the 52 correlations (92%) between WSS and the comprehensive scores of children's achievement (broad reading, broad writing, language and literacy, and broad math) fall within this moderate to high range. Only four of these correlations fall below .50. Overall, this table presents strong prima facie evidence for the concurrent aspects of WSS's validity.

Predictors of WJ-R Test Scores for Each of the Four Grade Level Samples

Concurrent aspects of validity were also examined by means of four-step hierarchical regression analyses. These regressions enabled us to establish whether the WSS ratings made a unique contribution to children's performance on the WJ-R over and above the influence of demographic factors and children's initial performance level on the WJ-R. Tables 3 and 4 show the predictors of spring WJ-R language and literacy and broad mathematics scores respectively, kindergarten—Grade 3. Results of the four-step regressions indicate that significant associations between WSS spring ratings and WJ-R spring outcomes remained even after controlling for the potential effects of age, SES, ethnicity, and students' initial performance level on the WJ-R in literacy (K—2) and in math (K and 1). Because children's performance on standardized achievement tests generally improves over time, we expected that as children progressed in grade, the fall to spring reliability of their WJ-R standard scores would increase significantly and a larger amount of the variance in students' spring WJ-R standard scores would be explained by

their fall WJ-R standard scores (their initial performance level). As anticipated, the stability of the second and third grade WJ-R standard scores was so high that initial performance on the fall WJ-R explained most of the variance in the spring WJ-R scores (Table 5).

When examined across grades several patterns are evident in the regression results. In the first step of the regressions only the demographic variables were entered. This model was significant only in kindergarten and second grade for language and literacy and in kindergarten for math. The checklist was significant at all grade levels for both math and literacy when entered into the second step of the regressions with the demographic variables; it explained more than half of the variance in literacy scores in grades 1 and 3. When the summary report was entered in the third step, both the summary report and the checklist contributed significantly in explaining the variance in the spring WJ-R literacy scores for kindergarten—second grade. In the third grade, the checklist alone was a significant predictor of the language and literacy score. In math, the WSS variables (either checklist or summary report) were significant predictors in step 3 of the regressions for kindergarten—grade 3. In brief, these results provide further support for the concurrent aspect of WSS's validity, particularly in the area of literacy.

Receiver-Operating-Characteristic Curves

To determine whether WSS can assist districts in identifying children who are in need of Title I programs or other supportive services, and in order to test whether children were classified in the same way by both WSS and WJ-R, a cost-matrix analysis was conducted. Cost-matrix analysis is a component of logistic regression. It consists of a statistical method for evaluating a cost, or for weighting differential outcomes, and then evaluating the weighted outcome distributions at a number of cut-off points. It is particularly useful for comparing two psychometric instruments that have a predictor—outcome relationship (see Meisels, Henderson, Liaw, Browning, & Ten Have, 1993). An optimal cutpoint is defined statistically as the point at which the loss value is minimized. In other words, when used, for example, with a screening instrument, an optimal cutpoint will produce a favorable ratio of overreferrals to underreferrals while maximizing correct identifications. Relying on the concepts of sensitivity (the proportion of at-risk children who are correctly identified) and specificity (the proportion of low-risk children who are correctly excluded from at-risk categories), this type of cost-matrix analysis is also called Receiver-Operating-Characteristic (ROC) curve analysis (Hasselblad & Hedges, 1995; Sackett, Haynes, & Tugwell, 1985; Toteson & Begg, 1988).

In this analysis we used data from students who had spring WJ-R Broad Reading and Broad Math scores and those who had WSS checklist ratings in language and literacy and in mathematical thinking. Because the WJ-R does not generate broad scores in reading and math in kindergarten, kindergartners were excluded from this analysis.

TABLE 2. CORRELATIONS OF WJ-R STANDARD SCORES USING GRADE NORMS WITH WSS CHECKLIST AND SUMMARY REPORT RATINGS*

WJR Subtest/ WSS Domain	Kindergarten		First Grade		Second Grade		Third Grade	
	Fall	Spring	Fall	Spring	Fall	Spring	Fall	Spring
WJR Letter Word Knowledge/WSS Language and Literacy Checklist	.36 (.005) N = 66	.45 N = 66	.65 N = 73	.68 N = 73	.61 N = 78	.63 N = 78	.78 N = 74	.76 N = 74
WJR Letter Word Knowledge/WSS Language and Literacy Summary Report	.53 N = 66	.56 N = 66	.55 N = 74	.68 N = 74	.70 N = 80	.72 N = 80	.65 N = 74	.45 N = 74
WJR Passage Comprehension/WSS Language and Literacy Checklist	NA	NA	.30 (.01) N = 74	.63 N = 74	.55 N = 75	.60 N = 75	.74 N = 83	.71 N = 83
WJR Passage Comprehension/WSS Language and Literacy Summary Report	NA	NA	.32 (.005) N = 76	.70 N = 76	.62 N = 77	.74 N = 77	.57 N = 83	.40 N = 83
WJR Dictation/WSS Language and Literacy Checklist	.44 N = 66	.47 N = 66	.68 N = 74	.68 N = 74	.53 N = 59	.48 N = 59	.74 N = 66	.72 N = 66
WJR Dictation/WSS Language and Literacy Summary Report	.59 N = 66	.62 N = 66	.57 N = 76	.58 N = 76	.68 N = 59	.69 N = 59	.55 N = 65	.47 N = 65
WJR Writing /WSS Language and Literacy Checklist	NA	NA	.49 N = 67	.59 N = 67	.59 N = 66	.61 N = 66	.55 N = 71	.54 N = 71
WJR Writing /WSS Language and Literacy Summary Report	NA	NA	.46 N = 69	.67 N = 69	.55 N = 67	.66 N = 67	.50 N = 72	.42 N = 72
WJR Broad Reading/WSS Language and Literacy Checklist	NA	NA	.55 N = 67	.73 N = 67	.60 N = 73	.62 N = 73	.80 N = 71	.79 N = 71

WJR Broad Reading/WSS Language and Literacy Summary Report	z	NA	.48 N = 68	.74 N = 68	.67 N = 75	.74 N = 75	.64 N = 71	.45 N = 71
WJR Broad Writing /WSS Language and Literacy Checklist	NA	NA	.63 N = 62	.68 N = 62	.56 N = 52	.57 N = 52	.70 N = 54	.70 N = 54
WJR Broad Writing /WSS Language and Literacy Summary Report	NA	NA	.56 N = 64	.67 N = 64	.66 N = 52	.73 N = 52	.57 N = 54	.53 N = 54
WJR Language and Literacy /WSS Language and Literacy Checklist	.44 N = 66	.50 N = 66	.68 N = 55	.79 N = 55	.63 N = 48	.60 N = 48	.78 N = 45	.82 N = 45
WJR Language and Literacy /WSS Language and Literacy Summary Report	.61 N = 66	.65 N = 66	.54 N = 56	.73 N = 56	.77 N = 48	.74 N = 48	.62 N = 45	.53 N = 45
WJR Applied Problems/WSS Mathematical Thinking Checklist	.52 N = 66	.57 N = 66	.44 N = 79	.56 N = 79	.40 N = 78	.36 (.005) N = 78	.67 N = 80	.65 N = 80
WJR Applied Problems/WSS Mathematical Thinking Summary Report	.59 N = 66	.65 N = 66	.52 N = 79	.51 N = 79	.46 N = 80	.56 N = 80	.54 N = 79	.56 N = 79
WJR Calculation/WSS Mathematical Thinking Checklist	NA	NA	.50 N = 80	.73 N = 80	.59 N = 83	.40 N = 83	.66 N = 80	.63 N = 80
WJR Calculation/WSS Mathematical Thinking Summary Report	NA	NA	.57 N = 80	.52 N = 80	.58 N = 85	.52 N = 85	.48 N = 80	.56 N = 80
WJR Broad Math/WSS Mathematical Thinking Checklist	NA	NA	.54 N = 78	.71 N = 78	.56 N = 78	.39 N = 78	.76 N = 76	.72 N = 76
WJR Broad Math/WSS Mathematical Thinking Summary Report	NA	NA	.64 N = 78	.56 N = 78	.57 N = 80	.58 N = 80	.58 N = 75	.60 N = 75

*All significance levels are based on 2 tail tests; $p < .001$ unless noted otherwise in parentheses.

NA = Not applicable

TABLE 3. PREDICTORS OF STUDENTS' SPRING WJ-R LANGUAGE AND LITERACY SCORES (K—3)

Variables Entered into Regression Model	Regression Coefficients for Model 1				Regression Coefficients for Model 2				Regression Coefficients for Model 3				Regression Coefficients for Model 4			
	K	1	2	3	K	1	2	3	K	1	2	3	K	1	2	3
Sex of student (Female)	-.159	.171	.243	-.004	-.205*	.018	.113	-.026	-.156	-.046	.098	-.023	-.184	-.105	.044	-.060
Student age (in months)	.241*	-.000	-.268*	-.181	.105	.092	-.181	-.128	-.022	.079	-.165	-.123	-.127	.053	-.040	-.002
Ethnicity (African American)	-.207	.017	.186	.022	-.072	-.008	.240*	-.012	-.102	-.048	.150	-.020	.032	.023	.008	-.010
Socioeconomic status (free or reduced lunch)	-.410*	-.203	-.344**	-.098	-.340**	-.035	-.262*	-.004	-.306**	.009	-.156	.001	-.146	.079	-.019	-.039
WSS Checklist Mean Scores (spring)					.413**	.800***	.528***	.812***	.231*	.538***	.245*	.866***	.205*	.388**	-.007	.111
WSS Summary Report Mean Scores (spring)									.483***	.378**	.487***	-.075	.262*	.308*	.158*	.034
WJ-R Standard Scores (fall)													.525***	.309*	.799***	.828***
R ²	.273	.057	.305	.040	.409	.639	.548	.689	.507	.694	.677	.692	.705	.734	.896	.885
R ² change	—	—	—	—	.136	.582	.243	.649	.158	.055	.129	.003	.138	.040	.219	.193

*p ≤ .05, **p ≤ .01, ***p ≤ .001

TABLE 4. PREDICTORS OF STUDENTS' SPRING WJ-R BROAD MATHEMATICS SCORES (K — 3)

Variables Entered into Regression Model	Regression Coefficients for Model 1				Regression Coefficients for Model 2				Regression Coefficients for Model 3				Regression Coefficients for Model 4			
	K	1	2	3	K	1	2	3	K	1	2	3	K	1	2	3
Sex of student (female)	-.012	.106	.079	.026	.035	.041	.051	-.020	.076	.000	.035	-.007	.044	.025	.088	.025
Student age in months	.229	.015	-.139	.061	.112	-.031	-.112	.091	.022	-.042	-.137	.091	-.089	-.032	-.011	.077
Ethnicity (African American)	-.021	.033	-.089	-.115	-.063	-.014	-.100	-.074	-.124	-.071	-.153	-.056	-.037	-.043	-.080	.013
Socioeconomic status (free or reduced lunch)	-.248**	-.279*	-.164	-.093	-.144	-.152	-.163	.040	-.101	-.121	-.065	.036	-.045	-.102	.010	-.029
WSS Checklist Mean Scores (spring)					.509**	.705***	.376***	.725***	.186	.579***	-.011	.598***	.171	.435**	-.046	.172
WSS Summary Report Mean Scores (spring)									.531***	.211	.591***	.182	.254*	.070*	.155*	.096
WJ-R Standard Scores (fall)													.494***	.370*	.770***	.695***
R ²	.137	.083	.064	.023	.363	.556	.205	.530	.487	.575	.391	.547	.609	.646	.739	.789
R ² change	—	—	—	—	.226	.473	.141	.507	.124	.019	.186	.017	.122	.071	.348	.242

*p ≤ .05, **p ≤ .01, ***p ≤ .001

TABLE 5. KINDERGARTEN—THIRD GRADE STUDENTS' ACHIEVEMENT GROWTH ON WJ-R STANDARD SCORES IN COMPARISON TO A NATIONALLY REPRESENTATIVE SAMPLE OF STUDENTS IN THE SAME GRADE LEVEL

WJ-R Subtest	Kindergarten	First Grade	Second Grade	Third Grade
Letter-word identification	**	*	**	*
Passage Comprehension	NA		**	*
Dictation	**		**	*
Writing Sample	NA		**	*
Broad Reading	NA	*	**	*
Broad Writing	NA		**	
Applied problems	**		**	*
Calculation	NA	*	*	**
Broad Math	NA	*	**	**
Math Skills	**	NA	NA	NA

* Students' growth as measured by WJ-R standard scores from fall to spring *meets* expected academic growth patterns for a nationally representative sample of students in the same grade level.

** Students' growth as measured by WJ-R standard scores from fall to spring *exceeds* expected academic growth patterns for a nationally representative sample of students in the same grade level.

NA = Not Applicable for subtests in kindergarten. Not Available for subtests first through third grade.

The remaining sample included all the children in Grades 1—3 who had been administered both the WJ-R and the WSS (N= 237 for Broad Reading and N=241 for Broad Math). Children were considered at-risk for academic difficulties if their score on the WJ-R was one or more standard deviations below the mean (i.e., WJ-R standard score ≤ 85). Analyses were conducted separately for Broad Reading and Broad Math. Children were considered not at risk if their scores were >85 . Using this cutoff, 42.2% (100/237) and 23.2% (56/241) of the children in this low-income, urban sample were at-risk in reading and math respectively. Using logistic regression cost matrices, optimal WSS cut-offs were derived for each domain with the dichotomous WJ-R categories as outcomes. The cutoff scores were a mean rating of 1.4 on the WSS Language and Literacy checklist, and a mean score of 1.2 on the Mathematical Thinking checklist.

Figures 1 and 2 show the area under the curve for the Language and Literacy Checklist and the area under the curve for the Mathematical Thinking Checklist. The area under the ROC curve represents the probability of a student performing poorly or well on both the WJ-R and the WSS. For Language and Literacy the probability represented by this area was 84%; for Mathematical Thinking it was 80%. These findings are very favorable because they show that a student in academic difficulty in either reading or math on WSS who is chosen randomly has a much higher probability of being ranked lower on the WJ-R than a randomly chosen student who is performing at or above average.

DISCUSSION

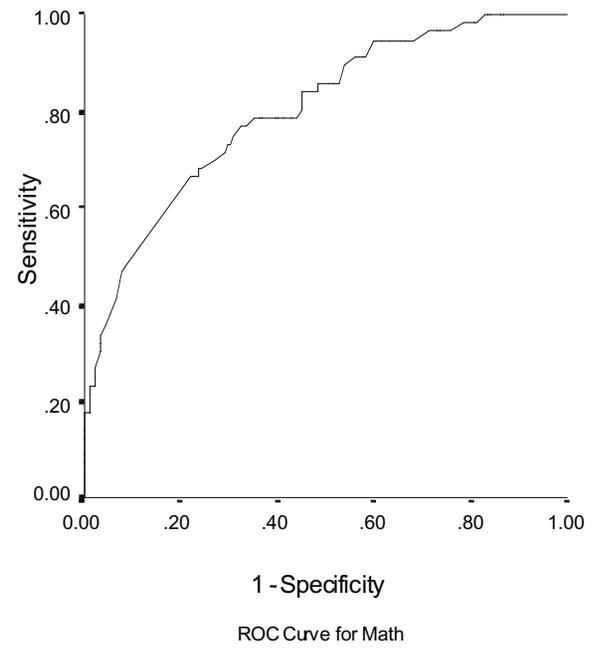
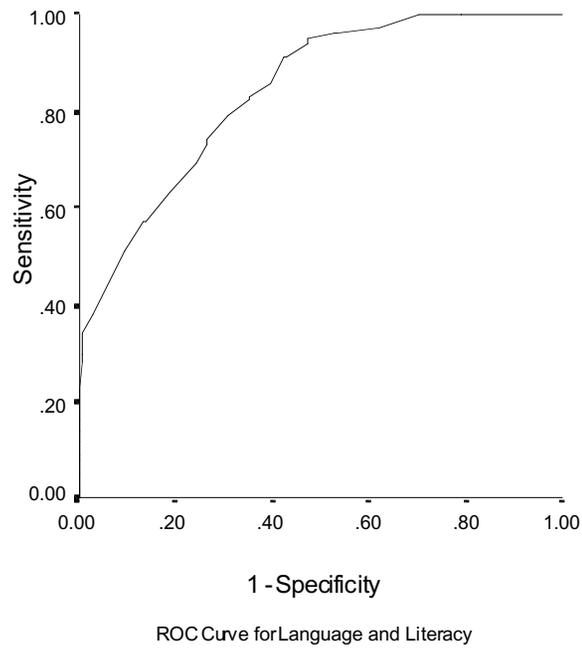
This study examined the question of whether teachers' judgments about student achievement are accurate when they are based on evidence from a curriculum-embedded performance assessment. We approached this question by examining psychometric aspects of the validity of the Work Sampling System. Overall, the results reported are very encouraging and support teachers' use of WSS to assess children's achievement in the domains of literacy and mathematical thinking in kindergarten—grade 3.

Aspects of WSS's validity were examined by comparing WSS checklist and summary report ratings with a nationally-normed, individually-administered, standardized assessment—the Woodcock-Johnson Psychoeducational Battery-Revised. Results of these correlational analyses provided evidence for these aspects of the validity of WSS. WSS demonstrates overlap with a standardized criterion measure while also making a unique contribution to the measurement of students' achievement beyond that captured through reporting WJ-R test scores. The majority of the correlations between WSS and the comprehensive scores of children's achievement (broad reading, broad writing, language and literacy, and broad math) are similar to correlations between the WJ-R and other standardized tests. For example, the WJ-R manual reports correlations between the WJ-R and other reading measures of .63 to .86; the majority of correlations between WJ-R comprehensive scores in literacy and WSS range from .50 to .80. Correlations between the WJ-R and other math measures range from .41 to .83; the range for the majority of

correlations between WSS and WJ-R broad math was .54 to .76 (Woodcock & Johnson, 1989).

Although most correlations reported were moderate to strong, a few of the correlations were $<.50$ in each of the grade levels. The lower correlations in kindergarten and the fall of first grade can be understood by considering the contrast between the limited content represented on the WJ-R literacy items in comparison to the full range of emergent and conventional literacy skills considered by WSS teachers as they rate young students literacy achievement. Cohort differences particularly in first grade also may have contributed to this variability. As students make the transition to conventional literacy—the focus of the WJ-R test items—correlations generally increase between the two measures. The lower correlations in Grade 3 are seen only with WSS summary report ratings and WJ-R spring scores. It is possible that teachers were influenced by factors other than the information normally considered when completing a summary report. For example, third graders' spring ITBS achievement scores, retention histories, or age for grade status may have strongly influenced teachers' judgments about whether students were performing by the end of the year in ways that met the expected levels of achievement for third graders. Analysis of mean WSS scores in third grade indicates that teachers overestimated student ability on the summary report in comparison to the WJ-R. Some teachers may have been trying, intentionally or not, to avoid retaining children—a high-stakes decision that was to be made by the District based on third grade performance. WSS is not intended to be used for high-stakes purposes and may lose its effectiveness when so applied. Nevertheless, despite the decrease in correlations at the end of third grade, the absolute correlations in third grade are very robust, especially between the checklist and WJ-R.

Aspects of validity for WSS were also investigated through four-step hierarchical regressions. Results of these analyses were very supportive of WSS. WSS ratings were more significant predictors of students' spring WJ-R standard scores than any of the demographic variables. Further, for kindergarten through second grade, WSS literacy ratings continued to show statistical significance in the regression models after controlling for the effects of students' initial performance level (fall standard scores). It is important to recognize that the increasing stability over time in students' WJ-R standard scores proved to be a significant factor in our design for examining the validity of WSS beyond second grade. That is, because children's standard scores begin to stabilize as they spend more time in school, by third grade the majority of the variance in children's spring standard scores was explained by their initial performance level. Thus, the fact that WSS ratings no longer emerged as significant predictors for third graders' spring standard scores was not necessarily a statement about the validity of WSS, but instead, reflected the increasing stability of standardized assessments with students in Grade 3 and beyond. Overall, the regression results provide evidence that WSS ratings demonstrate strong evidence for concurrent aspects of validity, especially regarding students' literacy achievement.



FIGURES 1 AND 2. ROC CURVES FOR LANGUAGE LITERACY AND MATH

The information provided by the ROC curve enables us to go beyond correlations to investigate whether individual students who score low or high on the WJ-R are also rated low or high on WSS. Correlations cannot fit individual subjects into a binary classification—that is, positive or negative, disabled or non-disabled, at-risk or not at-risk. ROC analysis focuses on the probability of correctly classifying individuals, thereby providing information about the utility of the predictions made from WSS to WJ-R scores.

The ROC curve has been utilized largely in epidemiological and clinical studies. The area under the ROC curve represents the probability that a random pair of normal and abnormal classifications will be ranked correctly as to their actual status (Hanley & McNeil, 1982). In its application to this study we targeted for identification those students who were above and below a standard score of 85 on the WJ-R, using the broad scores for reading and math. Students in need of educational intervention (i.e., those in academic difficulty) scored one or more standard deviations below the mean on the WJ-R. Students with standard scores >85 on the WJ-R were considered to be developing normally compared to a nationally representative sample.

These data showed us that if a student with reading difficulty (i.e., performing more than one SD below the mean on the WJ-R) and another student without reading difficulty are chosen randomly, the student in academic difficulty has an 84% chance of being ranked lower on the WSS Language and Literacy checklist than the student who is developing normally. Similarly, a randomly chosen student having difficulty in math has an 80% chance of being ranked lower on the WSS Mathematical Thinking checklist than a student who is developing normally. Although we are not suggesting that WSS be used to classify students into tracks or learning groups, the ROC analysis demonstrates that WSS teacher ratings have substantial accuracy and therefore significant utility in practice—particularly for programs that target at-risk learners, such as Title I.

Taken as a whole, this study's findings demonstrate the accuracy of the Work Sampling System when compared with a standardized, individually-administered psychoeducational battery. WSS avoids many of the criticisms of performance assessment noted earlier and it is a dependable predictor of achievement ratings in kindergarten—Grade 3. Moreover, the data obtained from WSS have significant utility for discriminating accurately between children who are at-risk and those not at-risk. As an instructional assessment, WSS complements conventional accountability systems that focus almost exclusively on norm-referenced data obtained in on-demand testing situations. In short, the question raised at the outset of this paper can be answered in the affirmative. When teachers rely on such assessments as the Work Sampling System we can trust their judgments about what and how well children are learning.

REFERENCES

- Almasi, J., Afflerbach, P., Guthrie, J., & Schafer, W. (1995). *Effects of a statewide performance assessment program on classroom instructional practice in literacy* (Reading Research Rep. No. 32). University of Georgia, National Reading Research Center.
- Aschbacher, P.R. (1993). *Issues in innovative assessment for classroom practice: Barriers and facilitators* (Tech. Rep. No. 359). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.
- Baker, E., O Neil, H., & Linn, R. (1993). Policy and validity prospects for performance-based assessment. *American Psychologist*, 48 (12), 1210—1218.
- Baron, J. B. & Wolf, D. P. (Eds.) (1996). *Performance-based student assessment: Challenges and possibilities* (Ninety-fifth Yearbook of the National Society for the Study of Education, Part I). Chicago: University of Chicago Press.
- Borko, H., Flory, M., & Cumbo, K. (October, 1993). *Teachers ideas and practices about assessment and instruction. A case study of the effects of alternative assessment in instruction, student learning, and accountability practice*. CSE Technical Report 366. Los Angeles: CRESST.
- Calfee, R., & Hiebert, E. (1991). Teacher assessment of achievement. *Advances in Program Evaluation* (vol. 1, pp. 103—131). JAI Press.
- Cizek, G. (1991). Innovation or enervation? Performance assessment in perspective. *Phi Delta Kappan*, 72 (9), 695—699.
- Corbett, H. D. & Wilson, B. L. (1991). *Testing, reform, and rebellion*. Norwood, NJ: Ablex Publishing.
- Darling-Hammond, L. (1994). Performance-based assessment and educational equity. *Harvard Educational Review*, 64 (1), 5—30.
- Darling-Hammond, L. Aness, J. (1996). Authentic assessment and school development. In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities*. Ninety-fifth yearbook of the National Society for the Study of Education (Part 1, pp. 52—83). Chicago: University of Chicago Press.
- Dichtelmiller, M. L., Jablon, J. R., Dorfman, A. B., Marsden, D. B., & Meisels, S. J. (1997). *Work sampling in the classroom: A teacher s manual*. Ann Arbor, MI: Rebus Inc.
- Falk, B., & Darling-Hammond, L. (March, 1993). *The primary language record at P.S. 261: How assessment transforms teaching and learning*. New York: National Center for Restructuring Education, Schools, and Teaching.

- Frederiksen, J., & Collins, A. (1989). A systems approach to educational testing. *Educational Researcher, 18* (9), 27—32.
- Gardner, H. (1993). Assessment in context: The alternative to standardized testing. In H. Gardner, *Multiple intelligences: The theory in practice* (pp. 161—183). New York: Basic Books.
- Gearhart, M., Herman, J., Baker, E., & Whittaker, A. (July 1993). *Whose work is it? A question for the validity of large-scale portfolio assessment*. CSE Technical Report 363. Los Angeles: CRESST.
- Green, D. R. (1998). Consequential aspects of the validity of achievement tests: A publisher's point of view. *Educational Measurement: Issues and Practice, 17* (2), 16—19, 34.
- Hanley, J. A. & McNeil, B. J. (1982). The meaning and use of the area under a Receiver Operating Characteristic (ROC) Curve. *Diagnostic Radiology, 143* (1), 29—36.
- Hasselblad, V. & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests. *Psychological Bulletin, 117*, 167—178.
- Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Hoge, R. D. (1983). Psychometric properties of teacher-judgment measures of pupil aptitudes, classroom behaviors, and achievement levels. *Journal of Special Education, 17*, 401—429.
- Hoge, R. D. (1984). The definition and measurement of teacher expectations: Problems and prospects. *Canadian Journal of Education, 9*, 213—228.
- Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. *Journal of Educational Psychology, 76*, 777—781.
- Hoge, R. D., & Coladarci, T. (1989). Teacher-based judgments of academic achievement: A review of the literature. *Review of Educational Research, 59*, 297—313.
- Hopkins, K. D., George, C. A., & Williams, D. D. (1985). The concurrent validity of standardized achievement tests by content area using teachers' ratings as criteria. *Journal of Educational Measurement, 22*, 177—182.
- Kenny, D. T., & Chekaluk, E. (1993). Early reading performance: A comparison of teacher-based and test-based assessments. *Journal of Learning Disabilities, 26*, 227—236.
- Kentucky Institute for Education Research. (January 1995). *An independent evaluation of the Kentucky Instructional Results Information System (KIRIS). Executive summary*. Frankfort, KY: The Kentucky Institute for Education Research.

- Khatti, N., Kane, M., & Reeve, A. (1995). How performance assessments affect teaching and learning. *Educational Leadership*, 53 (3), 80—83.
- Koretz, D., Mitchell, K., Barron, S., & Keith, S. (1996). *Final report: Perceived effects of the Maryland School Performance Assessment Program* (Tech. Rep. No 409). Los Angeles: CRESST.
- Koretz, D., Stecher, B., Klein, S., & McCafrey, D. (1994). *The evolution of a portfolio program: The impact and quality of the Vermont program in its second year (1992—1993)* (CSE Tech. Rep. No. 385). Los Angeles: CRESST.
- Linn, R. (1993). Educational assessment: Expanded expectations and challenges. *Educational Evaluation and Policy Analysis*, 15 (1), 1—16.
- Linn, R. (1994). Performance assessment: Policy promises and technical measurement standards. *Educational Researcher*, 23 (9), 4—14.
- Linn, R. (2000). Assessments and accountability. *Educational Researcher*, 29 (2), 4—15.
- Linn, R., Baker, E., & Dunbar, S. (1991). Complex, performance-based assessment: Expectations and validation criteria. *Educational Researcher*, 20 (8), 15—21.
- McTighe, J. & Ferrara, S. (1998). *Assessing learning in the classroom*. Washington, DC: National Education Association.
- Mehrens, W. (1998). Consequences of assessment: What is the evidence? *Educational Policy Analysis Archives*, 6 (13). Available: <http://olam.ed.asu.edu/epaa/v6n13.htm>.
- Meisels, S. J. (1996). Performance in context: Assessing children's achievement at the outset of school. In A. J. Sameroff & M. M. Haith (Eds.), *The five to seven year shift: The age of reason and responsibility* (pp. 407—431). Chicago: The University of Chicago Press.
- Meisels, S. J. (1997). Using Work Sampling in authentic assessments. *Educational Leadership*, 54 (4), 60—65.
- Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y., Atkins-Burnett, S. (1998). *Pittsburgh Work Sampling Achievement Validation Study*. Ann Arbor: University of Michigan School of Education.
- Meisels, S., Dorfman, A., & Steele, D. (1994). Equity and excellence in group-administered and performance-based assessments. In M. Nettles & A. Nettles (Eds.), *Equity in educational assessment and testing* (pp. 195—211). Boston: Kluwer Academic Publishers.
- Meisels, S.J., Henderson, L.W., Liaw, F., Browning, K., & Ten Have, T. (1993). New evidence for the effectiveness of the Early Screening Inventory. *Early Childhood Research Quarterly*, 8, 327—346.

- Meisels, S. J. , Jablon, J., Marsden, D. B., Dichtelmiller, M. L., & Dorfman, A. (1994). *The Work Sampling System*. Ann Arbor, MI: Rebus Inc.
- Meisels, S. J., Liaw, F-R., Dorfman, A., & Nelson, R. (1995) The Work Sampling System: Reliability and validity of a performance assessment for young children. *Early Childhood Research Quarterly*, 10 (3), 277—296.
- Meisels, S. J., Xue, Y., Bickel, D. P., Nicholson, J. & Atkins-Burnett, S. (in press). Parental reactions to authentic performance assessment. *Educational Assessment Journal*.
- Mills, R. P. (1996). Statewide portfolio assessment: The Vermont experience. In J. B. Baron & D. P. Wolf (Eds.), *Performance-based student assessment: Challenges and possibilities* (Ninety-fifth Yearbook of the National Society for the Study of Education, Part I, pp. 192—214). Chicago: University of Chicago Press.
- Moss, P. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research*, 62 (3), 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.
- Murphy, S., Bergamini, J., & Rooney, P. (1997). The impact of large-scale portfolio assessment programs on classroom practice: Case studies of the New Standards Field-Trial Portfolio. *Educational Assessment*, 4 (4), 297—333.
- Perry, N. E. & Meisels, S. J. (1996). *Teachers judgments of students academic performance*. Working Paper #96-08, National Center for Education Statistics. Washington, D. C.: U. S. Department of Education, OERI.
- Popham, W. J. (1996). *Classroom assessment: What teachers need to know*. Needham, MA: Allyn & Bacon.
- Resnick, L.B., & Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. In B. Gifford & M. C. O Connor (Eds.), *Cognitive approaches to assessment* (pp. 37—75). Boston: Kluwer-Nijhoff.
- Salvesen, K. A., & Undheim, J. O. (1994). Screening for learning disabilities. *Journal of Learning Disabilities*, 27, 60—66.
- Sackett, D. L., Haynes, R. B., & Tugwell, P. (1985). *Clinical epidemiology: A basic science for clinical medicine*. Boston: Little, Brown.

- Sharpley, C. F., & Edgar, E. (1986). Teachers' ratings vs. standardized tests: An empirical investigation of agreement between two indices of achievement. *Psychology in the Schools, 23*, 106—111.
- Shavelson, R. J., Baxter, G.P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement reality. *Educational Researcher, 21*, 22—27.
- Shepard, L. A. (1991). Interview on assessment issues. *Educational Researcher, 20*, 21—23, 27.
- Silverstein, A. B., Brownlee, L., Legutki, G., & MacMillan, D. L. (1983). Convergent and discriminant validation of two methods of assessing three academic traits. *Journal of Special Education, 17*, 63—68.
- Smith, M., Noble, A., Cabay, M., Heinecke, W., Junker, M., & Saffron, Y. (July 1994). *What happens when the test mandate changes? Results of a multiple case study*. CSE Technical Report 380. Los Angeles: CRESST.
- Stecher, B.M. & Mitchell, K.J. (1995, April). *Portfolio-driven reform: Vermont teachers understanding of mathematical problem solving and related changes in classroom practice*. CSE Technical report 400. Los Angeles: CRESST.
- Sternberg, R. J. (1996). *Successful intelligence: How practical and creative intelligence determine success in life*. New York: Simon & Schuster.
- Stiggins, R. J. (1997). *Student-centered classroom assessment* (2d ed.). Columbus: Merrill.
- Stiggins, R. J. (1998). *Classroom assessment for student success*. Washington, DC: National Education Association.
- Sykes, G. & Elmore, R. (1989). Making schools manageable. In J. Hannaway & R. Crowson (Eds.). *The politics of reforming school administration*. Philadelphia: Falmer.
- Taylor, C. (1994). Assessment for measurement or standards: The peril and promise of large-scale assessment reform. *American Educational Research Journal, 31* (2), 231—262.
- Toteson, A. N. A., & Begg, C. B. (1988). A general regression methodology for ROC curve estimation. *Medical Decision Making, 8*, 204—215.
- United States General Accounting Office. (1993). *Student testing: Current extent and expenditures, with cost estimates for a national examination*. (GAO/PEMD Publication No. 93-8). Washington, DC: Author.
- Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70* (9), 703—713.
- Wiggins, G. (1993). *Assessing student performance: Exploring the purpose and limits of testing*. San Francisco: Jossey-Bass.

- Wolf, D., Bixby, J., Glenn III, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. *Review of Research in Education, 17*, 31—74.
- Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Psychoeducational Battery-Revised*. Allen, TX: DLM Teaching Resources.