Early Childhood Research Quarterly, 10, 277–296 (1995)

The Work Sampling System: Reliability and Validity of a Performance Assessment for Young Children

Samuel J. Meisels
University of Michigan
Fong-ruey Liaw
National Taiwan Normal University
Aviva Dorfman
University of Michigan
Regena Fails Nelson
Western Michigan University

Performance assessment, an alternative approach to assessing students' achievements in school, refers to assessment methods that allow students to demonstrate their skills, knowledge, behavior, and accomplishments across a wide variety of classroom domains on multiple occasions. This article presents data concerning the reliability and validity of the Work Sampling System with 100 kindergarten-age children. A psychometric design was implemented in which children were enrolled in classrooms where the Work Sampling System was used and were also given individually-administered norm-referenced assessments in the fall and spring; in addition, their teachers completed a behavior rating scale in the spring. Results show that the Work Sampling checklist and summary report have very high internal and moderately high interrater reliability. The Work Sampling System accurately predicts performance on the norm-referenced achievement battery, even when the potential effects of gender, maturation (age), and initial

This article is based on research supported by the John D. and Catherine T. MacArthur Foundation. It was presented at the annual meeting of the American Educational Research Association, Atlanta, April 13, 1993. Grateful acknowledgment is made to the teachers and children who participated in this study. We also wish to thank our colleagues, Margo Dichtelmiller, Judy Jablon, Dorothea Marsden, Dorothy Steele, and Carolyn Burns for their advice regarding this study and their contribution to the development of the Work Sampling System®. Only the authors are responsible for the opinions expressed.

Correspondence and requests for reprints should be sent to Samuel J. Meisels, School of Education, University of Michigan, Ann Arbor, MI 48109.

ability are controlled. These data provide empirical support for the reliability and criterion validity of this performance assessment system as a measure of children's overall school achievement in kindergarten. The discussion covers issues raised by the study's design and by the use of performance assessment in general.

Performance assessment, representing an alternative approach to assessing achievement, is gaining widespread use in schools throughout the nation. It refers to assessment methods that enable students to demonstrate their knowledge or skills by solving problems, doing mathematical computations, writing journal entries or essays, conducting experiments, presenting oral reports, or assembling a portfolio of representative work. Nearly every state in the nation has begun to experiment with some form of performance assessment for obtaining achievement data in high school (U.S. Congress, Office of Technology Assessment, 1992), and some states (e.g., Michigan and Vermont) have mandated that some performance data be collected on all students at various points in their school careers.

Performance assessment may be most easily understood in contrast to the prevailing paradigm of group-administered achievement testing. In this approach to determining students' current knowledge or skill, students are confronted with a series of objectively evaluated, multiple-choice, norm-referenced, computer-scored questions that generally test their knowledge in an objectified, decomposed, and decontextualized manner (see Stallman & Pearson, 1990). These tests commonly entail reading a short-answer, multiple-choice question, and filling in bubbles, ovals, or circles. They are usually given on an annual basis and consume an estimated 20 million school days and the equivalent of $700 to $900 million in direct and indirect expenditures every year (National Commission on Testing and Public Policy, 1990). Findings from these tests are often used for "high stakes" purposes (Madaus, 1988; McGill-Franzen & Allington, 1993): to determine whether students are eligible to be promoted, receive their diplomas, or enter academically gifted classes as well as to determine the allocation of resources and, often, the actual content of curriculum. High-stakes group-administered achievement tests have been analyzed and criticized extensively (Calfee, 1987; Darling-Hammond & Wise, 1985; Fredriksen, 1984; Frederiksen & Collins, 1989; Haladyna, Nolen, & Haas, 1991; Haney, 1991; Koretz, 1988; Linn, 1987; Meisels, 1989a; Neill & Medina, 1991; Nickerson, 1989; Smith, Edelsky, Draper, Rottenberg, & Cherland, 1989). The new paradigm of performance assessment has emerged both in response to these criticisms and as a means of capturing elements of the teaching/learning process that are inaccessible within the constraints of the norm-referenced, group-administered framework (Calfee, 1987; Fredriksen, 1984; Shavelson, Baxter, & Pine, 1992; Stiggins, 1991; Wolf, Bixby, Glenn, & Gardner, 1991; Wolf, LeMahieu, & Eresh, 1992).

Authentic performance assessments are methods of documenting children's skills, knowledge, and behaviors using actual classroom-based experiences, activities, and products. (A variant that is not discussed in this article, known as "on-demand" performance assessment, calls for students to engage in performance tasks that are not part of the regular classroom routine.) Although no "canon" exists, and uniformity concerning the principles of performance assessment is unavailable, several features or criteria are common to authentic performance assessments (cf. Calfee, 1992; Herman, Aschbacher, & Winters, 1992; Shepard, 1991; Wiggins, 1989). First, they document children's daily activities; they do not simply provide a "snapshot," or a discontinuous view of children's accomplishments. Second, they provide an integrated method for evaluating the quality of children's work. This work is collected in a manner that bridges the broad range of curriculum areas and engages children in the metacognitive task of reviewing and evaluating their own learning. Third, they are flexible enough to reflect an individualized approach to academic achievement. Although performance assessments should be based on well-conceived values and systematic standards of knowledge and curriculum development, the actual implementation of these values and standards can be adjusted in relation to a specific classroom, teacher, and child. Finally, performance assessments are intended to evaluate those elements of learning and development that group-administered achievement tests do not capture very well: analysis, synthesis, evaluation, and interpretation of facts and ideas—the so-called "higher order thinking skills"—as well as student initiative and creativity.

Group-administered tests have been part of the educational scene for nearly a century. Performance assessment can make no such boast. Indeed, relatively little evidence is available to demonstrate the effectiveness of performance assessment (see Baker, O'Neil, & Linn, 1993; Shavelson et al., 1992). The data that are available are confined to children in late elementary school grades and above. Studies concerning performance assessment with young children are not available. However, it is unreasonable to expect that major changes in assessment and instructional policy can be undertaken without an empirical foundation for guiding those changes. In this article we describe a preliminary study of the reliability and validity of an authentic performance assessment for young children, the Work Sampling System (WSS; Meisels, Jablon, Marsden, Dichtelmiller, & Dorfman, 1994).

THE WORK SAMPLING SYSTEM

Work Sampling is a curriculum-embedded, continuous progress performance assessment system that offers an alternative to product-oriented, group-administered achievement tests in preschool through Grade 5. The purpose of Work Sampling is to assess and document children's knowledge, skills, behavior, and accomplishments on multiple occasions across a wide variety

of classroom domains. Work Sampling systematizes teacher observations by guiding those observations with specific criteria and well-defined procedures. It consists of three complementary elements: (a) developmental guidelines and checklists, (b) portfolios, and (c) summary reports. These elements are all classroom-focused and instructionally relevant. They comprise an ongoing evaluation process that reflects the goals and objectives of the classroom teacher and keeps track of children's continuous progress. By serving as a record of instruction as well as a summary of classroom achievement and accomplishment, these procedures are intended to have a positive effect on both instructional practices and children's learning. Here, each element of the WSS is described briefly.

Developmental Guidelines and Checklists

The guidelines and checklists are designed to assist teachers in observing and documenting individual children's growth and progress. Each of the 8 checklists, from age 3 to Grade 5 is completed 3 times per year and covers 7 domains: personal and social development, language and literacy, mathematical thinking, scientific thinking, social studies, the arts, and physical development.

The checklists are intended to reflect common activities and expectations in classrooms that are structured around developmentally appropriate activities. They and their associated guidelines are based on state and national curriculum standards (e.g., American Association for the Advancement of Science, 1993; Center for the Study of Reading, 1993; National Council of Teachers of Mathematics, 1993), and guidelines for developmentally appropriate practice (Bredekamp, 1987). Teachers are instructed to complete the checklists without actually "testing" their children. Rather, the checklists are used to create a profile of children's individualized progress in developing skills, acquiring knowledge, and mastering important behaviors across developmental domains as demonstrated in curriculum-embedded tasks. Variability in development across groups of children, and even across different domains within individuals, is expected. However, the information in the guidelines provides teachers with explicit rationales and examples for every performance indicator included on the checklists. In this way teachers can complete the checklists reliably and interpret the indicators in a consistent manner in their classrooms.

Portfolios

The second major element of the WSS consists of portfolios of children's work. Portfolios are a purposeful collection of students' work that illustrate their efforts, progress, and achievements, and potentially provide a rich documentation of each child's experience throughout the year. Portfolios also make it possible for children to become involved with the process of selecting and judging the quality of their own work. Portfolio collection in

the WSS is an activity in which teachers and children make instructional decisions as they compile the portfolio and as they discuss its contents. Portfolio contents parallels classroom activities and leads to the development of new activities based on joint teacher–student assessment of the child's progress and interests.

Work Sampling advocates a relatively structured approach to portfolio collection. Two different types of work are identified and collected: core items and individualized items. Core items are designed to show child growth and progress over time by documenting a particular area of learning within a curriculum domain. Individualized items are intended to capture the unique characteristics of individual children and to reflect activities that integrate multiple domains across the curriculum. In Work Sampling, the portfolio is a tool for documenting, analyzing, and summarizing the child's growth and development across the entire school year. The core items in particular permit analysis of changes both within and between children. Examples of core items include the following: a record of a child's use of language to obtain information, a child's record of how he or she solved a problem involving estimating, calculating, and/or measuring, or a child's collection of data over time about changing phenomena (e.g., weather, life cycles, plant growth).

Summary Reports

The final element of the WSS is the summary report, completed on each child three times per year. This report consists of a brief summary of each child's classroom performance. It is based on teacher observations and on the checklists and portfolios that are kept as part of the WSS. The report contains specific criteria for evaluating children's performance and progress in each of the domains of learning and behavior that are emphasized in the classroom. In completing this form, teachers carefully review the checklist and portfolio, and then make overall judgments using well-defined categories and criteria in order to report to parents, administrators, and others about each child's activities and progress.

The three elements of the WSS form an integrated whole. Checklists record a student's growth in relationship to teacher expectations and national standards. Portfolios graphically display the texture and quality of the child's work as well as his or her progress over time. Summary reports integrate this information into a concise record that the student's family can understand and that administrators can use.

Work Sampling draws upon teachers' perceptions of students while informing, expanding, and structuring their perceptions. It assesses students' development and accomplishments—rather than test-taking skills—in meaningful, curriculum-based activities. It enables teachers to recognize and nurture children's strengths and weaknesses, instead of rigidly classifying students as high or low achievers based on one-dimensional assessments.

Through interaction with the teacher around the portfolio and summary report, it enables families to become actively involved in the assessment process. By objectively documenting what children know and can do as well as how teachers teach, the WSS makes possible meaningful evaluation and authenthic assessment of achievement.

DESIGN

The program of research on the Work Sampling System that is presented in this study explores classical psychometric parameters (see American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1985). Its purpose is to examine the effectiveness of the WSS using traditional methods of reliability and validity. This article incorporates a method of quantitative analysis that can be applied to Work Sampling data for research purposes, but which is not recommended for use by classroom teachers in their individual classrooms.

Another approach to the use and interpretation of measurement information that has gained prominence recently in studies of performance assessment is known as "consequential validity." Building on the work of Messick (1989) and going beyond the traditional validity categories of construct-content-criterion (Moss, 1992), the consequential basis of test use is concerned with the reactions of the participants to the assessment program (cf. Miller & Legg, 1993; Miller & Seraphine, 1993). Consequences can be positive, as in the improvement of instruction or the enhancement of students' sense of control over their learning, or negative, as in the narrowing of curricula and the reduction in teacher autonomy.

Consensus is growing among researchers that alternative criteria of validation must be considered in order to understand fully the meaning and effectiveness of performance assessments (Messick, 1994; Moss, 1992). Linn, Baker, and Dunbar (1991) suggested that such criteria should include intended and unintended consequences of the assessment, the degree to which performance on specific assessment tasks transfers, fairness, content quality, comprehensiveness, cost and effectiveness, cognitive complexity demonstrated by students, and "the meaningfulness of the problems for students and teachers" (p. 20). Clearly, a program of research that explores the parameters of consequential validity is equally comprehensive to one that investigates psychometric issues.

This study does not explore the consequential basis of the Work Sampling System. Rather, it seeks to establish a psychometric foundation for this assessment through the comparison of the WSS with norm-referenced, individually administered rather than group-administered assessments. The choice of an individually administered assessment as a criterior for validity raises an important question: If the Work Sampling System can be validated through compari-

sons to individually administered assessments, why not simply use individually administered tests of achievement instead of performance-based measures? After all, individually administered assessments avoid many of the problems of group-administered, high-stakes tests noted earlier (e.g, narrowing of the curriculum, teaching to the test, encouragement of student guessing, decontextualized test items, one-dimensional responses) as well as the practical problems associated with demonstrating the "consequential validity" of performance assessments. The answer to this question relates primarily to the differences in information obtained from these two types of assessment. Individually administered assessments (especially if more than one is used) are excellent indicators of students' learning, but they are not designed to provide information that is as encompassing or differentiated as that included on performance assessments. They are not modifiable to meet differences in learning style, cultural background, or language usage of individual students. Moreover, once we adopt a summative, individually administered assessment as a high-stakes criterion, we open the door to a narrowing of curriculum, a focus on priorities that lie outside of the classroom, and other abuses that potentially accompany group-administered, high-stakes tests (see Fredriksen, 1984; Haladyna et al., 1991; Haney, 1991; Koretz, 1988; Madaus, 1988; Meisels, 1989a; Neill & Medina, 1991). Nevertheless, it is important to demonstrate a strong association between Work Sampling and more conventional assessments of achievement in order to provide a foundation for making important policy decisions.

Both psychometric and consequential data are extremely important for fully understanding performance assessment. But in light of the psychometric tradition of previous assessment research and the need to establish a strong empirical baseline to support potential innovations in this area, we believe that studies of the consequential validity of Work Sampling should follow more traditional investigations. These studies may eventually demonstrate that nothing is lost through the use of performance assessment; and a great deal more may be gained by this approach in terms of improving the conditions of learning and teaching.

METHOD

Subjects

A total of 100 children from 10 different classrooms in three Michigan school districts was selected for this study. Although the Work Sampling System covers the age range of 3 to 11 years, this study focused only on the filed trial edition of the kindergarten version. Children in this study entered kindergarten in September 1991. They ranged in age from 4 years, 11 months to 6 years, 6 months in the fall (M age = 5;7). None had been retained or previously enrolled in kindergarten. The children were selected by their teachers, who

were all voluntary participants in a field trial of the WSS. The sample was obtained by asking the teachers at the outset of the school year to select 3 students whom they thought would do well in kindergarten, 3 whom they thought might have some problems, and 4 other children at random. These selections were all "hunches" and were not based on any independent assessment. No records were kept by the researchers concerning which children were included in these groups and no further categorization in these terms was used by the teachers, the project, or its staff.

The three school districts that were represented included one that was primarily middle class and White (40 students), one that was both working class and middle class White (30 students), and one that was urban, poor, and racially mixed (30 students). No effect for race (comparing White vs. Black vs. Hispanic on every measure) was detected. Also, no differences were found in gender, school, district, race, or child age when children with missing data were compared with those on whom data were complete.

Procedures

Throughout the year the teachers implemented the WSS by completing the checklists three times (fall, winter, and spring), collecting material for the portfolios, and preparing a summary report at the end of the year.[1] The teachers participated in training sessions regarding Work Sampling on three occasions. In addition, project staff visited them in their classrooms approximately once per month. These visits combined observations of the classroom and the children, and general consultation about the WSS.

Measures

At the time of this study, the development checklist consisted of 69 items tapping 5 domains of development: art and fine motor (12 items: e.g., writing, drawing, cutting, pasting, working with clay, block building, working with puzzles), movement and gross motor (11 items: e.g., movement control, jumping, balancing, riding a wheeled toy, throwing and catching a ball), concept and number (15 items: e.g., time and space, classification and seriation, number, measurement, nonliving and living things), language and literacy (17 items: e.g., articulation and structure, usage and expression, listening, literature, writing, composition, reading), and personal/social development (14 items: e.g., making choices, conflict resolution, persistence, interaction, games with rules). The teacher rated the child's performance on each item in the fall, winter, and spring as (1) *not yet,* (2) *sometimes* (3) *often.*

---

[1] During the period of this data collection, the Work Sampling System was still in a pilot phase. Changes that were introduced following this study included relabeling the domains and expanding them from five to seven; revising the performance indicators and guidelines; modifying the rating scales in the checklist and summary report; clarifying the methodology for portfolio collection; increasing the collection periods for the summary report from one to three times per year; and increasing the age range to Grade 5.

At the end of the school year, teachers summarized each child's performance on the five domains of development as (1) *not yet accomplished,* (2) *accomplished,* or (3) *highly accomplished* based on the information gathered in the checklists, materials collected for the portfolio, and teachers' judgment about the child's progress. Subscores for checklist, portfolio, and progress, respectively, were constructed by summing the child's ratings on the five domains for each element. A total summary score for the summary report was then computed as the sum of the three subscale scores. Although this approach to aggregating data is widely used in the research literature, some loss of information occurs whenever multiple dimensions of functioning are collapsed into a single score. This limitation is ameliorated somewhat by the actual review of checklists and portfolios by independent raters (a description of the reliability procedure for summary reports is provided in the Results section).

In addition to the WSS, two individually-administered norm-referenced assessments were given to the children in the fall and spring. One was derived from the Kindergarten Achievement Battery of the Woodcock–Johnson Psychoeducational Battery—Revised (WJ-R; Woodcock & Johnson, 1989). Six subtests were administered: letter word identification, applied problems, dictation, science, social studies, and humanities. A total score for the WJ-R by grade was derived by summing the standard scores of the six subtests.

The second assessment consisted of the Motor Scale of the McCarthy Scales of Children's Abilities (MSCA; McCarthy, 1972). The index score was used to indicate children's motoric development. These assessments were administered by five trained examiners who were blind to the study's purposes. The order of the assessments in the fall and spring was counterbalanced to avoid order effects.

In the spring, the teachers also rated the children's social behavior on a 32-item Child Behavior Rating Scale (CBRS; Bronson & Love, 1987). This scale has a dual focus on prosocial behavior and on cognitively oriented on-task behavior. Examples of prosocial behavior are "plays with other children," or "cooperative with playmates when participating in a group play activity; willing to give and take in the group, to listen or to help others." Examples of cognitively oriented items include "completes learning tasks involving two or more steps (e.g., cutting and pasting) in an organized way," or "finds and organizes materials and works in an appropriate place, when activities are initiated." With two exceptions, all of the items used a 5-point Likert-type scale ranging from (1) *never* to (5) *always.* The two items that measured a child's hostility to other children were reverse-coded. Two additional items were excluded from the analyses due to missing data. A total score was computed by adding the remaining 30 items, with high values indicating more social or more on-task behavior.

A total of 96 children who had data for the fall, winter, and spring developmental checklists was available for study. Among the sample, 52

Table 1. Descriptive Statistics of the Measures

Assessments	*M*	*SD*	Range	Cut-Off (-1.5 *SD*)	*N*(%) ≤ Cut-Off[a]
Developmental Checklist Total					
Fall	148	25	87–193	109	11 (11.7%)
Winter	167	24	87–201	130	10 (10.6%)
Spring	179	21	98–201	147	6 (6.4%)
Summary Report					
Teacher	34	6	15–45		
Rater 1	35	3	22–39		
Rater 2	35	3	28–40		
Woodcock-Johnson Psycho-Educational Battery (Revised)					
Fall	630	72	448–784	522	8 (8.5%)
Spring	633	73	441–785	524	7 (7.4%)
McCarthy Scales of Children's Abilities (Motor Scale)					
Fall	48	10	22–73	33	6 (6.4%)
Spring	48	10	21–77	33	5 (5.3%)
Child Behavior Rating Scale					
Spring	119	17	69–148	93	8 (8.5%)

[a] Cut-off = 1.5 *SD* below *M*.

(54.2%) were girls, 83 were White, 95 had scores for the WJ-R in the fall and spring, and 86 had data on the CBRS. Table 1 presents the descriptive stastistics for these measures.

RESULTS

Reliability of the Checklist

The reliability of the developmental checklist was examined by Cronbach alphas and correlations. A subscale score was created for each of the five domains by summing the individual items of that domain. A total score for the developmental checklist in the fall, winter, and spring, respectively, was then computed by adding the five subscale scores. The correlations between the fall, winter, and spring checklists were high: .89 between the fall and winter checklists; .69 between the fall and spring checklists; and .89 between the winter and spring checklists. These correlations indicate a moderate to high level of reliability of measurement across the school year.

Table 2 presents the Cronbach alphas indicating the degree of internal consistency among items for the five domains of the checklist at all three time points. Alphas ranged from .87 to .94 demonstrating the high internal reliability of the checklist.

Table 2. Reliability of Developmental Checklist: Cronbach Alphas[a]

Domain	Fall	Winter	Spring
I. Art and fine motor (12 items)	.89	.90	.87
II. Movement & gross motor (11 items)	.88	.89	.91
III. Concept & number (15 items)	.90	.92	.91
IV. Language & literacy (17 items)	.93	.94	.94
V. Personal/social development (14 items)	.92	.93	.93

Criterion Validity of the Checklist

Concurrent validity was investigated by comparing the fall checklist to the fall WJ-R and MSCA scores, and the spring checklist to the spring WJ-R, MSCA, and CBRS scores. Predictive validity was examined by relating the fall and winter checklists to the spring assessments.

Concurrent Validity. The concurrent validity of the checklist was examined by zero-order correlations between the checklist and the other assessments. Moderate to high correlations were obtained between the checklist and the WJ-R ($rs = .75$ for the fall and .66 for the spring), and between the spring checklist and the spring CBRS scores ($r = .80$; see Table 3). In contrast, the correlations between the checklist and the MSCA were low ($rs = .39$ for the fall and .28 for the spring).

Predictive Validity. The predictive validity of the checklist was examined by means of correlation, regression, and computation of sensitivity and specificity, relating the fall and winter checklist to the spring individually administered assessments. Again, high correlations were obtained between the fall and winter checklists and the spring WJ-R and the CBRS scores (rs ranged from .67 to .76; see Table 3). Moderate to low correlations were found between the checklist and the spring MSCA ($rs = .43$ and .34).

Two-step hierarchical regressions were conducted to determine the unique contribution of the checklist to children's performance on the WJ-R (by grade), the MSCA, and the CBRS, over and above the effects of children's gender, age, and initial ability. In the first step, the fall checklist total score and the covariates (i.e., gender, age, and fall test score) were entered. The results (see Table 4) are presented as standardized regression coefficients, allowing us to determine the correlations between the predictors and the outcomes as well as the relative power of each predictor after controlling for other variables (see Cohen & Cohen, 1983). The winter checklist total score was entered in the second step, and the increment in the variance was noted to determine the contribution of the winter checklist above and beyond the fall checklist.

Table 3. Correlations Between the Developmental Checklist and Other Assessments[a]

	Fall		Spring		
Developmental Checklist Subscale	**WJ-R Total by Grade**	**MSCA Index**	**WJ-R by Grade**	**MSCA Index**	**CBRS[b]**
Fall					
Art & fine motor	.59	.32	.60	.40	.50
Movement & gross motor	.38	.34	.38	.32	.39
Concept & number	.75	.36	.75	.38	.64
Language & literacy	.75	.233	.78	.34	.67
Personal/social development	.60	.33	.60	.37	.63
Total Score	**.75***	**.39***	**.76***	**.43***	**.67***
Winter					
Art & fine motor			.65	.36	.62
Movement & gross motor			.39	.29	.43
Concept & number			.78	.30	.72
Language & literacy			.78	.23	.72
Personal/social development			.58	.34	.81
Total Score			**.76***	**.34***	**.76***
Spring					
Art & fine motor			.58	.29	.68
Movement & gross motor			.25	.22	.40
Concept & number			.67	.21	.74
Language & literacy			.72	.20	.78
Personal/social development			.51	.31	.84
Total Score			**.66***	**.28***	**.80***

[a] $N = .96$. [b] $N = 86$ for this scale. * $p < .001$.

Table 4. Standardized Regression Coefficients of Gender, Age, Initial Test Score, and Developmental Checklist Scores Predicting Spring Test Scores

Regression Step	**WJ-R Total**	**MSCA Motor**	**CBRS**
I. Gender (Female)	−.05	−.06	.07
Age at fall test[a]	−.04	−.20*	−.06
Initial test score[b]	.80***	.32**	.43***
Fall checklist total	.17**	.32**	.33**
Subtotal R^2	.80***	.31***	.52***
II. Winter checklist total			
ΔR^2	.02**	.01	.12***

[a] "Age at fall test" was the child's age when tests were given in the fall.
[b] Variable used for "initial test score" varied depending on the outcome, for example, fall test for WJ-R was used for spring WJ-R, and so forth.
* $p < .05$. ** $p < .01$. *** $p < .001$.

The results of the first-step regression indicate significant associations between the fall checklist and all spring outcomes even when the potential effects of gender, maturation (age), and initial ability (i.e., fall test scores) were controlled. It should be noted that, except on the WJ-R, the predictive power of the fall checklist was similar to that of the initial test scores. The results of the second-step regression indicate that the increment in the variance of the outcomes due to the entry of the winter checklist was minimal for the WJ-R and the MSCA (.02 and .01), but the substantial (.12) for the CBRS. These findings suggest that the fall checklist makes a significant contribution to predictions of the children's performance in the spring, and the winter checklist has additional importance in predicting children's prosocial behavior in the spring.

Sensitivity and Specificity Analyses. To this point we have presented correlational results that indicate the extent to which the rank of children's scores on one measure is retained on another measure. Although this information is useful, it does not tell us whether a particular child's developmental status has changed over time, for example, from above average to below average or vice versa. However, this limitation can largely be resolved by analysis of sensitivity and specificity ratios.

To compute sensitivity and specificity, the performance of each child on the assessments was first classified as above or below the cut-off score. A cut-off score is a value below which poor school performance may be suspected. In this study, we set the cut-off score 1.5 standard deviations below the means of each assessment. This corresponds to a standard score of 78 on the WJ-R, which is classified as *low ability* (Woodcock & Mather, 1989). A similar status is obtained for this cut-off on the MSCA (McCarthy, 1972).

In Table 1, we presented the cut-off scores for each of the assessments and the number of children whose scores fell below the cut-off. In the fall, a substantial proportion of the children (11.7%) were rated by their teachers as not performing well. This proportion fell to 6.4% on the spring checklist. The proportion of children whose scores fell below the cut-off decreased from 8.5% to 7.4% on the WJ-R, and 6.4% to 5.3% on the MSCA, respectively, from fall to spring. On the CBRS, 85% of children were rated below the cut-off.

Sensitivity and specificity were calculated using the cut-offs shown in Table 1. Sensitivity refers to the proportion of students who were below the cut-off on both the predictor and the outcome, shown as a function of all children who were below the outcome cut-off. Specificity is the converse: It reflects the proportion of students above the cut-off on both the predictor and the outcome in relationship to all those above the outcome cut-off. These proportions are extremely useful because they allow the number of true and false negatives to be computed, and they serve to evaluate the accuracy of the

Table 5. Sensitivity and Specificity of Fall and Winter Developmental Checklist and Fall Woodcock-Johnson—Revised[a]

Predictor	Outcome	Sensitivity	Specificity
Fall checklist	Spring checklist	.83	.93
Fall checklist	Spring WJ-R	1.00	.95
Fall checklist	Spring MSCA	.60	.91
Fall checklist	Spring CBRS	.75	.93[b]
Winter checklist	Spring checklist	1.00	.96
Winter checklist	Spring WJ-R	1.00	.97
Winter checklist	Spring MSCA	.60	.92
Winter checklist	Spring CBRS	.88	.97[b]
Fall WJ-R (by grade)	Spring WJ-R	.71	.97
Fall WJ-R (by grade)	Spring MSCA	.40	.93
Fall WJ-R (by grade)	Spring CBRS	.38	.95[b]
Fall MSCA motor	Spring MSCA	.40	.96

[a] $N = 96$. [b] $N = 84$.

predictors. Sensitivity and specificity ratios above .80 are considered acceptable (Meisels, 1989b).

Table 5 presents the sensitivity and specificity results. It shows that the fall and winter checklists have high sensitivity and specificity in predicting the spring checklist. In contrast, the fall WJ-R and especially the fall MSCA have low sensitivity in predicting the spring WJ-R and MSCA, indicating underidentification. In addition, the prediction of the fall and winter checklist to the spring WJ-R is substantially more accurate than the prediction of the fall WJ-R to itself in the spring. The checklist is a poor predictor of only the spring MSCA, for which the fall MSCA was even a poorer predictor than the checklist.

Reliability and Validity of the Summary Report

The reliability of the summary report was examined by interrater correlations. Validity was investigated by comparing the summary report to the spring assessments.

Reliability of the Summary Report. The reliability procedure consisted of two raters each coding 75 of the summary reports. Each rater had been present in half of the classrooms on a monthly basis. For this study, they were familiar with two thirds of the children who were coded ($N = 50$); one third ($N = 25$) were not known to either of them. Forty-nine children were jointly coded by the raters, consisting of 24 familiar and 25 unfamiliar children. Both raters were blind to the ratings assigned by the teachers. Only the children who were rated by both their teachers and the raters and had complete data on the checklist, portfolio, and summary report were included in the following analyses ($N = 33$). This subsample of children did not differ from the rest

Table 6. Correlations Between the Summary Report and Total Scores of Other Assessments by Raters[a]

	WJ-R Total by Grade	MSCA Motor	CBRS
Teachers	.65***	.37*	.61***
Rater 1	.69***	.54**	.80***
Rater 2	.69***	.58***	.66***

[a] $N = 33$. * $p < .05$. ** $p < .01$. *** $p < .001$.

of the sample ($N = 63$) in terms of test scores on the individually administered assessments or the fall checklist.

Interrater reliability was calculated by means of zero-order correlations between teachers and the two raters on the spring total scores. High interrater correlations between the two raters of the summary report ($r = .88$ for the total score, $p < .001$) were obtained, although the correlations between each of the raters and the teachers were lower (.73 and .68 for the total score, $p < .001$).

Criterion Validity of the Summary Report. To examine the concurrent validity of the summary report, the summary reports of the teachers and the raters were correlated with the WJ-R, MSCA, and the CBRS in the spring (see Table 6). Overall, the summary reports of both the teachers and the two external raters correlated moderately to highly with the spring WJ-R and the CBRS (*r*s ranging .61–.80), but lower with the MSCA.

DISCUSSION

These data provide initial evidence for the reliability and criterion validity of the WSS. Although the sample was relatively small, and an earlier version of the Work Sampling materials was used for this study, these findings provide a justification for the continued use of the WSS and a basis for designing more extensive studies of this and other performance assessments. In this section we focus primarily on the validity data. Both the internal and the interrater reliability data demonstrate that the Work Sampling checklist and summary report are highly dependable given the sample size and design we employed.

Most of the validity data are correlational. Although correlational data are relatively limited in terms of the conclusions that can be drawn about individual subjects, these data are instructive. The correlations between the WSS and the other assessments that are shown in Tables 3 and 6 provide strong support for the checklist's relation to the WJ-R and the CBRS, but not to the motor scale of the MSCA.

Why were such differences obtained between the findings for the WJ-R and the MSCA? Primarily, these differences can be attributed to the differences in the measures. Although both are individually administered, the WJ-R is a general achievement test that assesses skills that overlap with several of the checklist domains; in contrast, the MSCA is a full-scale developmental assessment. The motor scale, which is one of the six MSCA scales, evaluates a child's fine and gross motor performance in depth. It is likely that the teachers in the study did not have sufficient information to draw valid conclusions about their children's overall motor development because motor performance in these districts is left to the physical education instructor. However, the sensitivity of the MSCA from fall to spring is low in this study (.40). In other words, even when the MSCA is used as a predictor of itself, it overlooks 60% of the children who were at-risk in the spring. Because the total number of children who scored low on the MSCA at either time point was very small (4 of 97), few conclusions can be drawn from this other than the need to find some other measure of motor performance with better sensitivity to verify the perceptions of the classroom teacher. It is also worth noting that the motor scale is not included in computation of the General Cognitive Index (GCI) of the MSCA. Thus, even within the framework of the MSCA, this scale is considered to contain highly specialized information that is independent of general cognitive ability.

Specificity is extremely high for all fall assessments (including the fall checklist) as predictors of any of the spring outcomes, meaning that very few children who were assessed as low achieving in the fall were later found to be high achieving on any of the spring measures. But, the sensitivity of the fall and winter checklist was higher than that of the fall WJ-R in predicting the spring outcomes. Only 1 of 7 children who were low on the spring WJ-R was not low on the checklist in the fall. These findings suggest that the checklist is a better tool for evaluating children's general knowledge, skills, behavior, and achievements than the WJ-R.

The sensitivity/specificity data also provide a perspective on the potential for selection bias. Due to the lack of data about which children were categorized by their teachers as those whom would do well or poorly, or those whom were chosen at random, we cannot rule out the potential effect of initial selection procedures on teachers' perceptions of children's performance. However, if this were the case, we would expect the checklist to have extremely high sensitivity and specificity in predicting other teacher ratings (i.e., the CBRS). This does not obtain in our analysis, suggesting that the selection design may not be a threat to the conclusions drawn by this study.

The question remains, then, as to whether the checklist provides any significant information above and beyond initial ability on the WJ-R and such other predictors as age and gender. The regression analysis shown in Table 4 responds to the question. These data show that the checklist contributes to

the explanation of outcome even when age, gender, and fall test scores are controlled. Moreover, in other analyses (not shown) when the change score in performance on the checklists (i.e., the difference between fall and winter scores, and between winter and spring scores) is used to predict the child's performance in the spring, the change from fall to winter shows a significant association above and beyond the effects of gender, age, and fall test scores, indicating that the improvement over time that is tracked on the checklist is meaningful.

The summary report also provides very important information. Overall, using the WJ-R total score by grade as the outcome, the correlations with the summary report are moderately high. When the scores of teachers and raters are compared, it is clear that the raters were more highly correlated with themselves (.88) than they were with the teachers (.68 and .73). Because each rater scored 75 summary reports and went through pilot work to achieve reliability—as compared with each teacher's scoring of only 10 summary reports with relatively brief training—it is not surprising that the raters were more accurate on the summary report. This suggests that the raters, who scored more than seven times as many summary reports as did the teachers, may be more accurate in integrating different sources of information and summarizing children's performance. If so, this finding would suggest that with more training and familiarity with the WSS, more accuracy with the summary report may be achieved.

These findings led to the revision of the summary report and to the formulation of new staff development procedures. For this study, the summary report was to be completed only once per year using three response levels (expected, above expectations, and below expectations). The revised format is intended to be used on three occasions, corresponding to the three collection periods of the checklist and portfolio. Only two levels are now used, and the instructions and training concerning how these levels are to be applied have been clarified. We anticipate that these changes will result in higher reliability between teachers and observers in future studies.

The data presented here show that Work Sampling yields extremely stable results that are no less accurate than those obtained by a standardized assessment administered by trained examiners. Indeed, our results show that by virtue of the increased information available to teachers, the WSS provides more accurate predictions from fall to spring than are available from the norm-referenced, individually administered assessment itself.

But, can the characteristics of this particular performance assessment also be impediments to widescale adoption? For example, how can we know that the information obtained in one classroom, by one teacher, is comparable to the information obtained elsewhere, given the diversity of school curricula, teacher preparation, and child characteristics? In other words, how do we know that a common metric is being used when performance assessment is

implemented in different contexts? The data presented here are intended to address this issue by providing information about internal reliability, interrater reliability, and comparisons to norm-referenced measures. However, these data must be replicated with larger, more diverse samples and with more sensitive and comprehensive outcome criteria in order to be assured of the generalizability of this method and its results.

Another question concerns accountability. If the WSS were used to report school accountability, much as is done with group-administered tests today, would we not simply recreate the problems that are associated with high stakes tests? What prevents teachers from responding to pressure for better student outcomes by indicating that their students are performing more optimally than is actually the case? Any assessment can be misused, and performance assessment has no claim on immunity from such abuse. Although using Work Sampling for accountability is not desirable, several safeguards can be installed to determine whether a performance assessment such as the WSS is being distorted. For example, a random sample of students can be administered an individually administered assessment, as was done in this study. These students can also have their Work Sampling materials reviewed by independent raters. These steps will show whether or not Work Sampling is being used in an expected fashion or in a manner that suggests that the assessment is invalid. The individual assessment provides a normative reference point for comparison with the performance assessment, and the external/independent raters can evaluate the consistency and plausibility of the materials as a whole. Performance assessment, and Work Sampling in particular, consists of a web or net of interlocking data that must display consistency, continuity, and purpose, using both children's performances and teachers' judgments. This interlocking net of data is very difficult to invent because it is based on the continuous collection of evidence from and about children.

In sum, these findings demonstrate that the WSS is a reliable and valid approach for assessing the achievement of kindergarten-age children. These findings are still preliminary, and they reflect several shortcomings. For example, no attempt to collect interrater reliability data for the checklist was made, the portfolio was not analyzed in the same depth as the other elements of the system, and the checklist items may have included within-domain redundancy. Nevertheless, this study gives a good indication of the Work Sampling System's potential. Since the study was completed, refinements have taken place in all aspects and all elements of the WSS. The checklists have been rewritten to enhance their clarity, remove redundancy, increase the number of domains, change their scaling, and present a unified perspective from ages 3 through 11. The guidelines have similarly been revised to reflect the changes in the checklists and to specify more clearly what teachers should be evaluating. The portfolio materials have also been modified, particularly to clarify the selection of core items. Finally, the summary report has under-

gone major changes, as described earlier. In conjunction with the ongoing revision of staff development methods and materials, these changes should result in the WSS becoming an extremely useful performance assessment for preschool and the early elementary grades.

REFERENCES

American Association for the Advancement of Science. (1993). *Project 2061: Benchmarks for science literacy.* New York: Oxford University Press.

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

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

Bredekamp, S. (Ed). (1987). *Developmentally appropriate practice in early childhood programs serving children from birth through age 8.* Washington, DC: National Association for the Education of Young Children.

Bronson, M.B., & Love, J. (1987). Child behavior rating scale. Boston: Abt Associates.

Calfee, R. (1987). The school as a context for assessment of literacy. *The Reading Teacher, 40,* 738–743.

Calfee, R. (1992). Authentic assessment of reading and writing in the elementary classroom. In M.J. Dreher & W.H. Slater (Eds.), *Elementary school literacy: Critical issues* (pp. 211–226). Norwood, MA: Christopher-Gordon.

Center for the Study of Reading at The University of Illinois, The International Reading Association, & The National Council of Teachers of English. (1993). *Standards project for English language arts: A collection of documents for the English language arts profession.* Draft document.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Darling-Hammond, L., & Wise, A. (1985). Beyond standardization: State standards and school improvement. *The Elementary School Journal, 85,* 315–336.

Fredriksen, N. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist, 39,* 193–202.

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

Haladyna, T.M., Nolen, S.B., & Haas, N.S. (1991). Raising standardized achievement test scores and the origins of test score pollution. *Educational Researcher, 20,* 2–7.

Haney, W. (1991). We must take care: Fitting assessment to functions. In V. Perrone (Ed.), *Expanding student assessment* (pp. 142–163). Alexandria, VA: Association for Supervision and Curriculum Development.

Herman, J.L., Aschbacher, P.R., & Winters, L. (1992). *A practical guide to alternative assessment.* Alexandria, VA: Association for Supervision and Curriculum Development.

Koretz, D. (1988). Arriving in Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? *American Educator, 12,* 8–15, 46–52.

Linn, R.L. (1987). Accountability: The comparison of educational systems and the quality of test results. *Educational Policy, 1,* 181–198.

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

Madaus, G.F. (1988). The influence of testing on the curriculum. In N. Tanner & K.J. Rehage (Eds.), *Critical issues in curriculum: Eighty-seventh yearbook of the national society for the study of education* (pp. 83–121). Chicago: University of Chicago Press.

McCarthy, D. (1972). *McCarthy scales of children's abilities.* New York: Psychological Corporation.

McGill-Franzen, A., & Allington, R.L. (1993). Flunk'em or get them classified: The contamination of primary grade accountability data. *Educational Researcher, 22,* 19–22.

Meisels, S.J. (1989a). High-stakes testing in kindergarten. *Educational Leadership, 46,* 16–22.

Meisels, S.J. (1989b). Can developmental screening tests identify children who are developmentally at risk? *Pediatrics, 83,* 578–585.

Meisels, S.J., Jablon, J., Marsden, D.B., Dichtelmiller, M.L., & Dorfman, A.B. (1994). *The work sampling system: An overview* (3rd ed.). Ann Arbor, MI: Rebus Planning Associates.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–104). New York: Macmillan.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23,* 12–23.

Miller, M.D., & Legg, S.M. (1993). Alternative assessment in a high-stakes environment. *Educational Measurement: Issues and Practice, 12,* 9–15.

Miller, M.D., & Seraphine, A.E. (1993). Can test scores remain authentic when teaching to the test? *Educational Assessment 1,* 119–129.

Moss, P.A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Research, 62,* 229–258.

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

National Council of Teachers of Mathematics. (1993). *Curriculum and evaluation standards for school mathematics.* Reston, VA: Author.

Neill, D.M., & Medina, N.J. (1991). Standardized testing: Harmful to educational health. *Phi Delta Kappan, 73,* 688–697.

Nickerson, R.E. (1989). New directions in educational assessment. *Educational Researcher, 18,* 3–7.

Shavelson, R.J., Baxter, G.P., & Pine, J. (1992). Performance assessments: Political rhetoric and measurement issues. *Educational Researcher, 21,* 22–27.

Shepard, L.A. (1991). Interview on assessment issues. *Educational Researcher, 20,* 21–23, 27.

Smith, M.L., Edelsky, C., Draper, K., Rottenberg, C., & Cherland, M. (1989). *The role of testing in elementary schools.* Los Angeles: Center for Research on Educational Standards and Student Tests, Graduate School of Education, UCLA.

Stallman, A.C., & Pearson, P.D. (1990). Formal measures of early literacy. In L.M. Morrow & J.K. Smith (Eds.), *Assessment for instruction in early literacy* (pp. 7–44). Englewood Cliffs, NJ: Prentice Hall.

Stiggins, R.J. (1991). Assessment literacy. *Phi Delta Kappan, 72,* 534–539.

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

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

Wolf, D.P., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigating new forms of student assessment. In G. Grant (Ed.), *Review of research in education* (Vol. 17, pp. 31–74). Washington, DC: American Educational Research Association.

Wolf, D.P., LeMahieu, P.G., & Eresh, J. (1992). Good measure: Assessment as a tool for educational reform. *Educational Leadership, 49,* 8–13.

Woodcock, R.W., & Johnson, M.B. (1989). *Woodcock-Johnson psychoeducational battery—Revised.* Allen, TX: DLM Teaching Resources.

Woodcock, R.W., & Mather, N. (1989). Tests of achievement: Examiner's manual. In R.W. Woodcock & M.B. Johnson (Eds.), *Woodcock-Johnson psychoeducational battery—Revised.* Allen, TX: DLM Teaching Resources.