Early Education and Development, in press

Assessing language, literacy, and mathematics skills with *Work Sampling for Head Start*

Samuel J. Meisels

Erikson Institute

Yange Xue

Mathematica Policy Research, Inc.

Melissa Shamblott

St. Paul Public Schools

Author note

We thank Lisa Gruenewald, Tom Watkins, Diane Reitter, and Marian Heinrichs of the St. Paul Public Schools, as well as the teachers and children who participated in this study. Pearson Early Learning, the publisher of the *Work Sampling System*™ and *Work Sampling for Head Start*™, provided funding for this study, although the authors alone are responsible for the findings and interpretations presented herein.

Correspondence concerning this article should be sent to Samuel J. Meisels, Erikson Institute, 420 N. Wabash, Chicago, IL 60614, (312) 893-7100, smeisels@erikson.edu.

Abstract

Research findings: We examined the reliability and validity of the language, literacy, and mathematics domains of *Work Sampling for Head Start* (WSHS), an observational assessment designed for three and four year olds. Participants included 112 children who were enrolled over a two year period in Head Start and a number of other programs sponsored by community-based organizations affiliated with a local school district. Teachers were trained to administer the WSHS checklist and to collect observational data about their children over the course of the children's year of enrollment. Outcome data were individually-administered tests of early reading and early mathematics. Cronbach alphas, correlations, regressions, and Receiver-Operating-Characteristic curves were computed. Results indicated very high reliability of WSHS subscales. Findings also demonstrated moderate correlations between WSHS and the outcomes and unique contributions to the assessments of reading and mathematics by WSHS, over and above demographic variables. The ROC curves showed that WSHS can be used accurately by teachers to predict children's early mathematics and reading performance.

Practice and policy: Discussion includes the role of observational vs. norm-referenced tests in early childhood classrooms. Also discussed are such issues as variance in methods of assessment and the impact of high-stakes tests on young children.

Assessing language, literacy, and mathematics skills with *Work Sampling for Head Start*

High-stakes testing—using student test scores to determine rewards or sanctions for children, teachers, administrators, and schools—has virtually overwhelmed educational practice in the first decade of the 21st Century. As a result of the *No Child Left Behind Act* (NCLB) of 2002 every child in Grades 3 – 8 is tested in reading and mathematics. Scores on these norm-referenced tests are used to decide whether students will be promoted or retained, whether teachers and administrators will receive congratulations or condemnation, and whether schools will be considered successful or not.

NCLB not only increased the stakes associated with testing in Grades 3 - 8, it brought increased attention to the issue of accountability beginning as early as preschool. The most dramatic example of high-stakes testing with young children was the *National Reporting System* (NRS; Administration for Children and Families, 2003), a test so flawed that it was suspended by Congress in 2007. It consisted of 40 – 50 items in language, literacy, and mathematics administered twice yearly to all English- and Spanish-speaking four and five year olds in Head Start. Although the NRS was never used for high-stakes purposes, its potential role in program closure was widely publicized (Administration for Children and Families, 2003). Between fall 2003 and spring 2007 the NRS was administered more than three and a half million times, despite questions raised about its psychometric properties by the General Accountability Office (2005) and others (Meisels & Atkins-Burnett, 2004). Several states (viz. Florida and Texas) have also adopted testing regimes that have high-stakes implications for preschools.

Alternatives to high-stakes tests in preschool are available, and have been used with young children for some time. Distinctively different in purpose and application from tests associated with the NRS and NCLB, these "low stakes" tests focus on data that facilitate instructional decision making and rely on observational methods for collection of information about children's performance. This paper presents psychometric data concerning one such instrument.

The three most-widely used early childhood observational assessments are the *Child Observation Record* (COR; High/Scope, 1992), the *Developmental Continuum* (Teaching Strategies, 2002), and the *Work Sampling System* (WSS; Meisels, Jablon, Dichtelmiller, Marsden, & Dorfman, 2001). Validity evidence for these three observational instruments is uneven, with no published data available for the *Developmental Continuum* as contrasted to fairly extensive information published about the COR (Fantuzzo, Hightower, Grim, & Montes, 2002; Sekino & Fantuzzo, 2005; Schweinhart, McNair, Barnes, & Larner, 1993) and WSS. This paper focuses on an adaptation of WSS for Head Start.

Psychometric data concerning WSS is available from a number of studies. In research conducted in 17 Title I classrooms (N = 345 students, K- 3) in a large urban school district, WSS ratings were compared with student scores on a nationally-normed, individually-administered psychoeducational battery in order to examine construct and predictive aspects of validity (see Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001; Meisels, Xue, Bickel, Nicholson, & Atkins-Burnett, 2001). Correlations between WSS checklist ratings in literacy and mathematical thinking and standardized test scores were moderate to high. Four-step hierarchical regressions showed that WSS ratings were a stronger predictor of test scores than demographic variables. Other analyses demonstrated that WSS discriminated between children who were at-risk and those not at-risk.

Another study, though not one dealing with validity per se, (Meisels et al., 2003) examined the trajectory of change in scores of WSS and non-WSS third and fourth graders on the *Iowa Tests of Basic Skills* (ITBS). The scores on the ITBS in third and fourth grades of WSS children were compared with those of students in a group of non-WSS contrast schools matched by demographic variables. A second comparison group consisted of all other students in the school district. Results indicated that students who were in WSS classrooms displayed growth in reading from one year to the next that far exceeded the demographically matched contrast group (25:1) as well as the average change shown by all other students in the district (8:1). The pattern of change was similar between mathematics and reading. Other studies of the reliability and validity of WSS with kindergarten-age children are also available (Meisels, Liaw, Dorfman, & Nelson, 1995).

However, no psychometric research about WSS with children younger than age five exists. This is due primarily to the limitations of criterion measures used in validity studies with children this young. LaParo and Pianta (2000), in a meta-analysis of more than 70 empirical studies designed to predict achievement in first or second grade from cognitive tests administered in preschool or kindergarten found that less than 25% of the variance in the outcomes could be accounted for by test-based predictors, even when differences in assessment method were minimized. Others have also noted problems of validity with early childhood measures of achievement (Kim & Suen, 2003; Meisels, 2007; Neisworth & Bagnato, 2004), although this remains an area of significant controversy (Meisels, 2007).

The present study represents the first research using WSS with preschoolers. It makes use of a modified version of the WSS checklist that was developed for use in Head Start in order to enhance the alignment between WSS and the Head Start Outcomes Framework (DHHS, 2001). Other than minor changes in format and organization, relatively few differences in content exist between *Work Sampling for Head Start* (WSHS; Dichtelmiller, Jablon, Meisels, & Marsden, 2001) and the standard WSS checklist for four year olds.

WSHS is used extensively with three and four year olds in Head Start and versions of WSHS or WSS are the mandated or preferred preschool and/or kindergarten assessments in a number of states (e.g., South Carolina, Georgia, Minnesota, Maryland, Arkansas, Colorado, Illinois, Pennsylvania). Because of this widescale use, and specifically because of the accountability culture in which all schools in the U.S. function today, it is important to understand the relationship of WSHS to normative measures of achievement. To accomplish this, we collaborated with a large, urban school district that uses WSHS to collect data on a sample of three- and four-year olds.

In this analysis, we investigate the reliability and validity of three domains of WSHS using data collected from children in the fall and spring of 2004 – 2005 and 2005 – 2006. The focus of this study is the relationship of WSHS to psycho-educational assessments of children's achievement in language development, early literacy, and mathematics. Its purpose is to establish the reliability and validity of this observational, performance-based assessment in relation to normative measures of achievement.

Methods

Participants

This analysis presents data from 112 children enrolled in the St. Paul Public Schools (SPPS) CHOICE program, an Early Reading First (ERF) federally-funded project. CHOICE was coordinated by SPPS and functioned as a community partnership between SPPS's School Readiness and Community Kindergarten program, the Ramsey County Head Start program, and a YMCA child care center. All programs were selected according to their relative proximity within the SPPS boundaries, their agreement to participate in and fulfill the requirements of the ERF grant (e.g., using a literacy curriculum, adopting specific instructional strategies, administering several assessments including WSHS, and participating in a comprehensive professional development plan), as well as the specifications of this study.

Participants were enrolled in three School Readiness classrooms, some of which were specifically designated "inclusion classrooms" that enrolled high proportions of children with special needs, 12 Head Start classrooms in a single center, and one community-based classroom operated by the YMCA. Due to design exclusions, most of the children in the study were enrolled in Head Start, but the study was conducted under the auspices of the public schools, rather than Head Start.

Children in the sample were required to have parental permission and to meet two selection criteria: (1) >3.6 years of age, and (2) able to communicate test responses in English and speak English in the classroom. Children with special needs whose IEPs indicated that they were in the mild to moderate range (most had speech or physical impairments) were included in the study. Children who had moderate to severe special needs were not eligible for the study. Using these criteria, our original sample included 71 children in 2004-2005 and 69 children in 2005-2006. Of these, 12 children in 2004-2005 and 16 children in 2005-2006 were excluded from the analyses due to incomplete data. Thus, our final sample included 59 participants in 2004-2005 (53% were enrolled in Head Start) and 53 in 2005-2006 (60% were enrolled in Head Start). Children from these two years were combined for the study. Missing data analyses comparing the 28 children who were excluded with the final sample reveal that the missing group did not differ from the final sample with respect to sex, race/ethnicity, and special education. However, the missing group was slightly younger than the final sample ($t_{[138]} = -3.22$, $p<.01$).

The age range of the children who remained in the study was 3.77 years - 4.98 years with a mean age of 53.84 months (SD = 3.97) (two children were enrolled for a second year in the program; there was no upper age limit). The sample consisted of slightly more boys (54.5%) than girls. The breakdown of the sample by race/ethnicity was primarily minority (80.4%) with 62.5% Black, 8.9% Hispanic, and 8.9% Other. Most of the children received free or reduced lunch (94.6%). Children with special needs constituted 11.6% of the sample. Sixteen teachers were involved in the study; seven of whom participated both years. Half of the teachers in the SPPS School Readiness program had Master's degrees; the others had BAs. The one teacher in the YMCA program had a BS degree. However, in the Head Start programs only one-fourth of the teachers had BAs (none had master's degrees). The others had early childhood AA degrees or CDA certificates.

Measures

Work Sampling for Head Start. WSHS is a curriculum-embedded, criterion-referenced performance assessment that is intended to document what children are learning and have begun to master by providing specific information about their academic, personal and social, and other accomplishments. This analysis used data from the language development, literacy, and mathematics domains of the WSHS checklist (see Appendix for a list of components and performance indicators for the domains). Two reasons explain why only these three domains were studied. First, as with other ERF projects nationwide, several specific language and literacy assessments were already required by that program. Additional testing to examine all WSHS domains was viewed by both federal and SPPS administrators as a potential overextension for the children, families, and programs. Because of this, we limited our focus to language, literacy, and mathematics. Second, even if it were possible to administer additional assessments, the outcome measures available for the domains not studied (social and emotional, approaches to learning, science, creative arts, and physical health and development) are either very limited or unreliable in the early childhood years. To reduce measurement error, they were not included here.

The WSHS checklist consists of 55 items that measure eight domains of development: social and emotional development, approaches to learning, language development, literacy, mathematics, science, creative arts, and physical health and development. Teachers in the study completed all domains with the exception of science and creative arts.

As shown in the Appendix, 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 child's performance. Accompanying every performance indicator are detailed developmental guidelines. These content standards present the rationale for each performance indicator and 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 across teachers, children, and schools.

Teachers rate children's performance on each item of the WSHS checklist three times per year (fall, winter, and spring) using an on-line record-keeping system. The rating scale includes three mastery levels: 1 (not yet), 2 (in process), and 3 (proficient). This report uses ratings in language development, literacy, and mathematics in the fall and spring. For purposes of analysis, language development and literacy were combined into one subscale. Subscale scores for language and literacy and mathematics were generated by computing the mean score for all items within the domain of language and literacy or mathematics.

Psychoeductional Assessments. The *Test of Early Reading Ability – 3d Edition* (TERA-3; Reid, Hresko, & Hammill, 2002) and the *Test of Early Mathematics Ability – 3d Edition* (TEMA-3; Ginsburg & Baroody, 2003) were administered in the fall and spring. The TERA-3 is an individually administered assessment of young children's reading achievement that was normed on a nationally representative sample of 875 children chosen in a random stratified sample procedure. It includes subtests in alphabet, conventions, and meaning and was designed for use with children aged 3 years, 6 months to 8 years, 6 months. It was administered to the children in this study in the fall and spring. The standard score for the composite of the three subtests has a mean of 100 and a standard deviation of 15 with the normative sample.

The TERA-3 was administered by two highly trained SPPS literacy coaches. In the second year a third master's level coach joined these two. To prepare for the TERA-3 administration the coaches studied the test materials, familiarized themselves with the assessment, and practiced administering the assessment to children not in CHOICE before giving the assessment to the students in the study.

The TEMA-3 is an individually administered test of early mathematical achievement that is appropriate for children aged 3 years through 8 years, 11 months. Available in two parallel forms, the test focuses on concepts of relative magnitude as well as knowledge of counting, calculation, conventions, and number facts. The TEMA-3 was normed on a nationally representative sample of 1228 children, of whom 673 took Form A and 591 took Form B. The standard scores for Forms A and B have a mean of 100 and a standard deviation of 15 with the normative sample. In this study, Form A was administered to the children in the fall and Form B was administered in the spring. In both years the TEMA-3 was administered by someone who was trained in the test administration by an SPPS school psychologist and supervised by someone familiar with the assessment.

Training and Fidelity

Training. Training was designed to help teachers learn how to use the WSHS *Developmental Guidelines and Checklists* (Dichtelmiller et al., 2001) for observing, documenting, and evaluating children's learning. The model implemented throughout the study included large and small group training and one-to-one consultation. Different training modalities were used to meet the needs of individual teachers as well as to accommodate the logistics of scheduling across three different agencies.

Prior to the onset of the study the teachers attended a one-day introduction to WSHS in which they were introduced to authentic performance-based assessment, the goals of the research study, as well as to the basics of observing, documenting, and recording children's skills, behaviors, learning, and achievements over time. After the initial training, teachers participated in 1 – 2 hours of WSHS follow-up training every month throughout the duration of the study period (excluding the summer months) and additional training when the checklists were being completed. The follow-up training took place in either large group, small group, or one-to-one settings and included an additional session focused solely on the Work Sampling on-line system that was used for data collection. All participating teachers were expected to attend the full complement of training.

Fidelity of implementation. The accuracy of the checklist procedures was monitored throughout the study by the WSHS trainer. The participating teachers' use of their collected observations as well as the WSHS guidelines and checklists were observed and monitored regularly. The participating teachers' written classroom observations were also reviewed periodically by both the WSHS trainer and the teachers' respective program supervisors.

Reliability of the checklist completion procedures was a major focus of the follow-up training. The training involved the review of simulated observation notes, work samples, and other collected documentation. The teachers were organized into small groups and asked to examine the documentation and determine which checklist rating best reflected the child's performance at various points throughout the year.

Once the small group work was complete, all teachers participated in a trainer-led discussion about the documentation and the checklist ratings. Each small group was asked to report their rating selections for the checklist indicators. The task continued until a unanimous rating selection had been made for each indicator. This training was repeated four times throughout the study period.

Analytic Approach

Four types of analyses were conducted using teachers' WSHS checklist ratings of children's achievement and children's test scores in the fall and spring: (1) reliability of each domain of the checklist; (2) correlations between children's test scores and the WSHS checklist ratings within the corresponding domain; (3) 3-step hierarchical regressions that examined factors accounting for children's test scores in the spring; and (4) Receiver-Operating-Characteristic (ROC) curves that determined whether a child who performed at average or below average level on the test scores was identified correctly based on the WSHS ratings. These methods were selected because they specifically enable us to answer our research questions about the reliability and validity of WSHS.

The reliability of the WSHS checklist ratings was examined by calculating Cronbach alphas for each domain in the fall and spring and estimating correlations between fall and spring within each domain. These statistics estimate the internal consistency or extent to which individual items within a specific subscale are correlated with the other items in that subscale. A Cronbach alpha value of at least .70 is considered sufficient but 0.80 to 0.90 is desirable (Nunnally & Bernstein, 1994). Correlations between fall and spring WSHS checklist ratings within each domain were obtained as well. The correlations indicate the reliability of WSHS ratings over time.

Traditional inter-rater reliability data were not collected. Doing so would violate the principle of independence required for establishing reliability between an observer and a "tester" (in this case the teacher). If two teachers were assigned to the same classroom neither would be "blind" to the other's decision making unless they did not confer about their pupils—an untenable situation. Moreover, if an external observer conducted reliability checks, the teacher and observer would have very different information from one another because the ratings reflect children's performance over time—not just on three specific occasions.

Using the TERA-3 and TEMA-3 as outcomes, we examined concurrent validity and predictive validity of the WSHS checklist. Correlations between WSHS ratings and test scores in the fall or spring indicate the shared variance between the two assessments and provide evidence of concurrent validity. Correlations of .70 - .75 are optimal because they show a substantial overlap between the two assessments as well as the uniqueness of each assessment. High correlations (≥ .80) suggest the predictor does not add enough new information to justify its use, while correlations that are low (≤ .30) suggest very little overlap between the two assessments. Concurrent validity was also examined through regression analyses relating spring WSHS to spring test scores.

The predictive validity of the WSHS checklist was examined by correlations between fall WSHS and spring test scores, hierarchical regression analyses, and ROC curve analyses. We performed 3-step regression analyses to examine if the WSHS checklist made a unique contribution to children's spring TERA-3 and TEMA-3 scores above and beyond demographic variables and before and after children's initial scores in the fall were controlled. Step 1 included demographic variables: age, sex (1 = boy, 0 = girl), race/ethnicity (two dummy indicators with white as the reference group), and free/reduced lunch (1 = yes, 0 = no). Step 2 added the WSHS checklist to the model; R square change from step 1 to step 2 demonstrates the contributions of WSHS in predicting test scores after controlling for demographics. Step 3 added test scores in the fall; this step indicates whether WSHS ratings make additional contributions in predicting spring test scores after adjustments were made for children's prior test scores. Separate models were run to determine whether the fall and spring checklist made different contributions. In addition, we included an indicator for the two cohorts of children in the regression analysis to examine whether there were any differences in relationships for the two cohorts.

ROC curve analysis, also called cost-matrix analysis, is a component of logistic regression. It is an effective method for evaluating two psychometric instruments that have a predictor-outcome relationship (Meisels, Henderson, Liaw, Browning, & TenHave, 1993). We used this method to examine whether two different assessments assign children to the same or different categories. An ROC figure provides a visual representation of a predictor’s accuracy while the area under the ROC curve gives a quantitative measure of its accuracy (Hanley & McNeil, 1982; Zhou et al., 2002). Area under the ROC curve that is ≥ .80 is considered excellent. An optimal cutpoint is defined statistically as the point at which the proportion of at-risk children who are correctly identified and the proportion of low-risk children who are correctly excluded from at-risk categories are maximized. We first established a cutoff for the test score (1SD below the mean for both language and literacy and mathematics, i.e., ≤ 85) to identify children at risk of learning difficulty. We then performed an ROC curve analysis to determine the probability that the WSHS ratings accurately assigned children to a high and low risk group.

Results

Descriptive statistics

Table 1 displays the descriptive statistics for the standard scores of the TERA-3 and TEMA-3 and the WSHS checklist ratings in language and literacy and mathematics in the fall and spring. The means of the test scores for the study sample were substantially lower than the national norms provided in the test manuals. The study sample scored more than one standard deviation below the test's norms in the fall and spring on the TEMA-3 and two-thirds of a standard deviation below the national norms on the TERA-3 in the fall and spring.

Reliability

Table 2 presents the Cronbach alphas, which indicate the degree of internal consistency among the items for the two WSHS checklist subscales in the fall and spring. Alphas ranged from .90 to .94, suggesting high internal reliability of the subscales in our sample. The correlations between the fall and spring WSHS scores were high: .71 for language and literacy; and .65 for mathematics (see Tables 3 and 4, respectively), suggesting that WSHS ratings are reliable over time.

Correlations

Tables 3 and 4 show the correlations between test scores and WSHS checklist ratings in language and literacy and mathematics, respectively. The results reveal that all of the correlations except those between fall WSHS Mathematics and TEMA-3 are within a moderate range (.35 - .44). The correlations in mathematics are lower than those in language and literacy. Although the correlations show substantial variance unaccounted for by WSHS, all relationships are statistically significant. The highest correlations are between the same measure across time. WSHS Language and Literacy correlations from fall to spring are somewhat higher than comparable correlations for the TERA-3, but the fall-spring TEMA-3 correlations are substantially greater than those for WSHS Mathematics over the year.

Regressions

Hierarchical regression results demonstrate that after controlling for sex, age, race/ethnicity, and SES, the spring or fall WSHS checklist is significantly associated with spring TERA-3 scores. This is true even after controlling for fall TERA-3 scores in the model (see Tables 5a and 5b). The checklist explained a substantial proportion (approximately one-fifth) of the variance in TERA-3 scores after controlling for demographics.

The regression results for mathematics demonstrate a similar pattern. Adjusting for sex, age, race/ethnicity, and SES, the spring or fall WSHS checklist alone is a significant predictor, and remains significant when controlling for fall TEMA-3 scores (see Tables 6a and 6b). The spring WSHS checklist explained approximately 16% of the variance in TEMA-3 scores; the fall WSHS checklist explained approximately 20% of the variance adjusting for demographics.

The analyses that included the indicator of cohort membership show no difference between the two cohorts of children. Therefore, we dropped this indicator in all regression analyses.

ROC

The ROC curve analysis used data from children who have both test scores and checklist ratings in the spring. Children were considered at risk if their scores were ≤ 85 in reading or mathematics. Using this cutoff, 38% and 63.7% of the children in our sample were at risk in reading and mathematics, respectively.

Figure 1 shows the ROC curve for language and literacy (N = 98). The area under the curve is .73, representing the probability of a child performing poorly or well on both the TERA-3 and the WSHS checklist. Figure 2 shows the ROC curve for mathematics (N = 110). The area under the curve is .74, representing a probability of a child being correctly identified on both the TEMA-3 and the WSHS checklist.

Discussion

This report presents the results of analyses of the reliability and validity of the language development, literacy, and mathematics domains of WSHS with a sample of 3- and 4-year olds. The findings provide evidence of its reliability and validity, suggesting that WSHS can be used with confidence in assessing these domains of children's learning. The results of reliability studies of the WSHS checklist demonstrated excellent internal consistency and high correlations within each domain across time. This shows that WSHS is highly reliable in assessing children's language and literacy and mathematics skills.

Most correlations between the WSHS checklist ratings and the test scores were moderate, ranging from .35 -. 44. The only exception was the correlation between fall WSHS and the TEMA-3, which was somewhat lower. (Interestingly, a study of the COR that used both the TERA-3 and the TEMA-3 also presented moderate correlations, albeit relationships that were weaker than those shown here for WSHS [Sekino & Fantuzzo, 2005].) Although these correlations provided only moderate support for the validity of WSHS, this study achieved these correlations despite the two conditions noted below that render these results quite conservative and that cannot be overlooked in interpreting the findings from this study.

First, WSHS and the TERA-3 and TEMA-3 measures represent a major difference in method. WSHS is a criterion-referenced, performance assessment based on indirect teacher report and teacher judgment; as used in this study, TERA-3 and TEMA-3 are direct, standardized, norm-referenced tests. Differences in results using the two types of assessments are highly probable simply because they measure different though overlapping parameters in different ways. All studies that incorporate significant method variance are left with the question of which assessment is more accurate, the normative test that takes place twice yearly or the performance assessment that collects continuous data across time. This question cannot be answered conclusively because of the incomparability of the measures and their differences in content. Indeed, some shared variance is the best that can be expected, since the two types of assessment do not measure the same things. Nevertheless, psychometricians aim to maximize the overlap between the two indicators of achievement as we have shown here.

Second, the means of the study children on the TERA-3 and TEMA-3 were substantially lower than the means of the normative sample on which the TERA/TEMA instruments were developed, despite the fact that both normative tools used a nationally representative sample in their standardizations**.** On the TEMA-3, the study sample scored a full standard deviation below the test's norms and on the TERA-3 this discrepancy was two-thirds of a standard deviation. Clearly, the results of the children in the study sample were lower overall than those in the TERA-3 and TEMA-3 normative samples. This is largely a reflection of the homogeneity of the study sample as compared with the normative samples, but its impact on the correlations that were obtained between the tests cannot be overlooked. (We do not believe that this finding suggests a "floor" problem with the tests, since both the TERA-3 and the TEMA-3 are normed for children more than six months younger than our participants.)

To better understand the inferences that we can draw based on WSHS we can turn to the results from the hierarchical regression analyses of fall and spring WSHS checklists in predicting children's performance in the spring. After controlling for demographic variables, WSHS accounts for about 20% of the variance in early reading and mathematics skills. (Analysis of the same outcomes with the COR shows that it accounts for less than 10% of the variance with the TERA-3 and TEMA-3 [Sekino & Fantuzzo, 2005].) These findings demonstrate that WSHS adds unique information to the predictive equation about children's early reading and mathematics achievement and support claims about the predictive aspects of WSHS's validity. It should be noted, however, that due to the clustered nature of our data (i.e., children nested within classrooms), our analyses might overestimate the statistical significance of findings.

The ROC curves contribute additional support to the validity argument for WSHS. These data indicate that teachers' WSHS ratings have substantial accuracy in identifying children at risk for learning difficulties in literacy and mathematics. The area under the curve in the ROC represents the proportion of correct identifications (both true positives and true negatives) between the outcomes (TERA/TEMA) and the predictor (WSHS). Nearly three-quarters of all such predictions were accurate, despite the method variance and sample differences noted above.

To explore the relationship among these outcomes still further, we examined the individual scores on WSHS and test scores by classroom. The findings uncovered no between-group differences in either the TERA-3 or TEMA-3 scores. However, there were significant differences ($p<.001$) between the Head Start ($N = 11$) and non-Head Start ($N = 5$) teachers on their WSHS ratings, with Head Start teachers indicating that their students performed substantially better than children in the non-Head Start classes. In particular, more than half of Head Start children were rated proficient on all WSHS items by their teachers in the spring, which suggests that Head Start teachers might overestimate their students' ability at the end of the school year. The between-group differences suggest that the context in which the assessments were administered, as well as potential differences in teachers' backgrounds and preparation, may account for some of the discrepancy between the normative measures and the observational assessment. On further examination, it appears that the greatest difference between the two groups may be familiarity with WSHS. The non-Head Start teachers had been using WSHS for 8 – 10 years, while the Head Start teachers were all new to WSHS. This, along with other factors in the teachers' and children's backgrounds and substantial differences in teachers' professional preparation, specifically regarding WSHS, could account for this within-sample variation. Howver, all teachers, regardless of program affiliation, viewed their children as performing at a more competent level on WSHS than would be expected from reviewing the TERA-3 and TEMA-3 scores in isolation and there were no between-group differences in these scores.

In short, within the limitations of this sample (mostly low-income, primarily minority, English-speaking, not randomly selected) and this group of teachers (wide range of qualifications and experience, working under diverse auspices), this study provides evidence for the psychometric validity of the inferences that can be drawn from WSHS with three and four year olds. To improve its accuracy, WSHS should be used as part of a systematic battery of instructional assessments in which its results could likely be strengthened by the addition of such supplementary data sources as, for example, evidence obtained from portfolios of children's work. In its original form, WSS incorporates data from both checklists and portfolios (Meisels et al., 2001).

As a low-stakes assessment, WSHS is not intended to supplant the normative tests that are used for accountability. However, because of the psychometric limitations of these instruments that have been documented elsewhere, the potential for iatrogenic effects of labeling and stigmatization, and the overall lability of development in the first five years of life, it is worth considering the advisability of assigning high-stakes to tests of achievement with young children under any circumstances (see Meisels, 2007). This study, and investigations of similar instruments, show that valuable and accurate information can be obtained from observational instruments that rely on teachers' judgments of children's performance. Such assessments focus on how children learn and how teachers can make instructional decisions that optimize development rather than on how children can be ranked and ordered and how tests can be used to allocate rewards and punishments.

References

Administration for Children and Families (June 6, 2003). Information Memorandum: Description of the NRS Child Assessment. Accessed online on September 28, 2003 at *http://www.headstartinfo.or/publications/im03_07.htm*

Department of Health and Human Services (2001). *The Head Start path to positive child outcomes*. Washington, DC: Author.

Dichtelmiller, M. L., Jablon, J. R., Meisels, S. J., & Marsden, D. B (2001). *The Work Sampling System for Head Start*. New York: Pearson Early Learning.

Fantuzzo, J., Hightower, D., Grim, S., & Montes, G. (2002). Generalization of the Child Observation Record: A validity study for diverse samples of urban, low-income preschool children. *Early Childhood Research Quarterly, 17(1)*, 106-125.

General Accountability Office (May 2005). *Head Start: Further development could allow results of new rest to be used for decision making*. Washington, D. C.: Author.

Ginsburg, H.P. & Baroody, A. J. (2003). *TEMA-3: Test of Early Mathematics Ability* (3rd ed.). Austin, TX: Pro-Ed.

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.

High/Scope Educational Research Foundation (1992). *High/Scope Child Observation Record for Ages 2 – 6.* Ypsilanti, MI: High/Scope Press.

Kim, J. & Suen, H. K. (2003). Predicting children's academic achievement from early assessment scores: A validity generalization study. *Early Childhood Research Quarterly, 18(4)*, 547-566.

LaParo, K. M. & Pianta, R. C. (2000). Predicting children's competence in the early school years. A meta-analytic review. *Review of Educational Research*, *70(4),* 443 – 484.

Meisels, S. J. (2007). Accountability in early childhood: No easy answers. In R. C. Pianta, M. J. Cox, & K. Snow (Eds.), *School readiness and the transition to kindergarten in the era of accountability* (pp. 31-47). Baltimore, MD: Paul H. Brookes.

Meisels, S.J. & Atkins-Burnett, S. (2004). The Head Start National Reporting System: A critique. *Young Children, 59*(1), 64-66.

Meisels, S. J., Atkins-Burnett, S., Xue, Y., Nicholson, J., Bickel, D. D., & Son, S. (2003). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement test scores. *Education Policy Analysis Archives*, *11(9).* http://epaa.asu.edu/epaa/v11n9/.

Meisels, S. J., Bickel, D. D., Nicholson, J., Xue, Y., & Atkins-Burnett, S. (2001). Trusting teachers' judgments: A validity study of a curriculum-embedded performance assessment in Kindergarten – Grade 3. *American Educational Research Journal*, *38(1),* 73–95.

Meisels, S. J., Henderson, L. W., Liaw, F., Browning, K., & TenHave, T. (1993). New evidence for the effectiveness of the Early Screening Inventory. *Early Childhood Research Quarterly, 8(3)*, 327-346.

Meisels, S. J., Jablon, J. R., Dichtelmiller, M. L., Marsden, D. B., & Dorfman, A. B. (2001). *The Work Sampling System* (4th ed.). New York: Pearson Early Learning.

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. D., Nicholson, J., & Atkins-Burnett, S. (2001). Parental reactions to authentic performance assessment. *Educational Assessment*, *7*(1), 61–85.

Neisworth, J. T., & Bagnato, S. J. (2004). The mismeasure of young children. *Infants & Young Children, 17*(3), 198-213.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed). New York: McGraw Hill.

Reid, D.K., Hresko, W.P., & Hammill, D. D. (2002). *Test of Early Reading Ability* (3d ed.). Austin, TX: Pro-Ed.

Schweinhart, L. J., McNair, S., Barnes, H., & Larner, M. (1993). Observing young children in action to assess their development: The High/Scope Child Observation Record Study. *Educational and Psychological Measurement*, *53(2)*, 445-455.

Sekino, Y. & Fantuzzo, J. (2005). Validity of the Child Observation Record: An investigation of the relationship between COR dimensions and social-emotional and cognitive outcomes for Head Start children. *Journal of Psychoeducational Assessment, 23(3)*, 242-260.

Teaching Strategies, Inc. (2002). *The Developmental Continuum Assessment System.* Washington, DC: Author.

Zhou, X. H., Obchowski, N. A., & McClish, D. K. (2002). *Statistical methods in diagnostic medicine*. New York: John Wiley & Sons.

Table 1. Descriptive statistics for test scores and WSHS checklist

	M Fall (SD)	M Spring (SD)
Test score		
TEMA-3	82.79 (13.42)[1]	83.77 (14.41)[2]
TERA-3	90.27 (11.43)[3]	90.67 (11.78)[4]
WSHS checklist		
Mathematics	2.25 (0.51)[5]	2.74 (0.36)[6]
Language and literacy	2.36 (0.50)[7]	2.78 (0.32)[8]

[1] N=107
[2] N=102
[3] N=101
[4] N=100
[5] N=110
[6] N=105
[7] N=110
[8] N=105

Table 2. Reliability of the WSHS checklist in the fall and spring

	Number of items	Fall	Spring
Language and literacy	12	.94	.94
Mathematics	8	.92	.90

Table 3. Correlations between TERA-3 and WSHS Language and Literacy

	TERA-3 (fall)	TERA-3 (spring)	WSHS Language and literacy (fall)
TERA-3 (fall)	--		
TERA-3 (spring)	.68***	--	
WSHS Language and Literacy (fall)	.44***	.39***	--
WSHS Language and Literacy (spring)	.41***	.41***	.71***

**p<.01; ***p<.001.

Table 4. Correlations between TEMA-3 and WSHS Mathematics

	TEMA-3 (fall)	TEMA-3 (spring)	WSHS Mathematics (fall)
TEMA-3 (fall)	--		
TEMA-3 (spring)	.77***	--	
WSHS Mathematics (fall)	.30**	.40**	--
WSHS Mathematics (spring)	.35***	.40**	.65***

**p<.01; ***p<.001.

Table 5a. Regression results (standardized coefficients) for spring WSHS Language and Literacy predicting TERA-3

	Model 1	Model 2	Model 3
Sex (male)	-.165	-.068	-.036
Age (months)	-.154	-.272**	-.247**
Race (Black)	-.060	-.044	.037
Race (other)	.011	-.025	.035
Reduced lunch	-.029	.015	.005
Spring WSHS Language and Literacy		.482***	.235**
Fall TERA-3 scores			.587***
R square	.057	.257	.532
R square change	--	.200	.275

N = 92. **p<.01; ***p<.001.

Table 5b. Regression results (standardized coefficients) for fall WSHS Language and Literacy predicting TERA-3

	Model 1	Model 2	Model 3
Sex (male)	-.171	-.024	-.014
Age (month)	-.147	-.333**	-.286**
Race (Black)	-.068	-.092	.033
Race (Other)	.011	-.058	.030
SES	-.030	-.063	-.031
Fall WSHS Language and Literacy		.503***	.204*
Fall TERA-3 scores			.624***
R square	.059	.250	.546
R square change	--	.191	.296

N = 93. *p<.05; **p<.01; ***p<.001.

Table 6a. Regression results (standardized coefficients) for spring WSHS Mathematics predicting TEMA-3

	Model 1	Model 2	Model 3
Sex (male)	.008	.072	.045
Age (month)	-.033	-.152	-.070
Race (Black)	-.147	-.084	-.048
Race (Other)	-.059	-.057	-.044
SES	-.211*	-.177	-.022
Spring WSHS Mathematics		.428***	.192*
Fall TEMA-3 scores			.688***
R square	.067	.225	.613
R square change	--	.158	.388

N = 99. *p<.05; ***p<.001.

Table 6b. Regression results (standardized coefficients) for fall WSHS Mathematics predicting TEMA-3

	Model 1	Model 2	Model 3
Sex (male)	-.024	.074	.048
Age (month)	.035	-.160	-.077
Race (Black)	-.180	-.171	-.094
Race (Other)	-.066	-.105	-.070
SES	-.211*	-.258**	-.068
Fall WSHS Mathematics		.487***	.261**
Fall TEMA-3 scores			.671***
R square	.083	.270	.642
R square change	--	.187	.372

N = 100. *p<.05; **p<.01; ***p<.001.

Figure 1. ROC curve for language and literacy

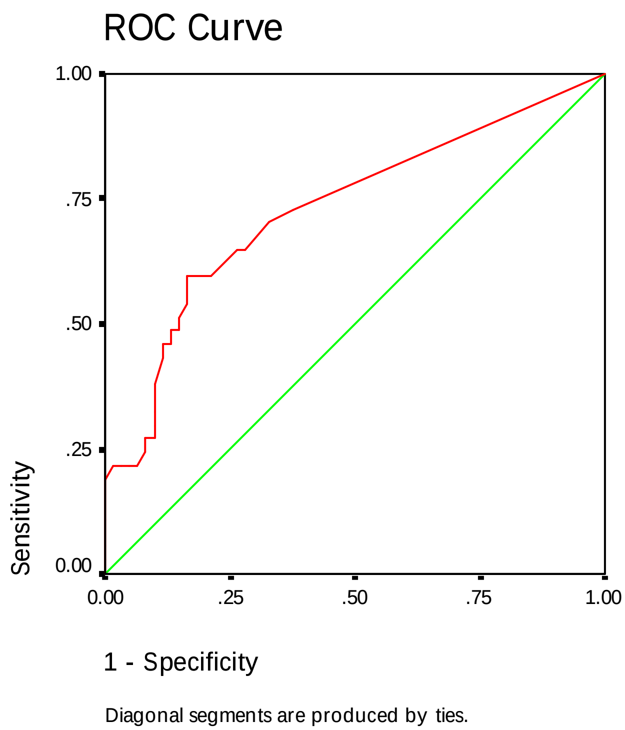
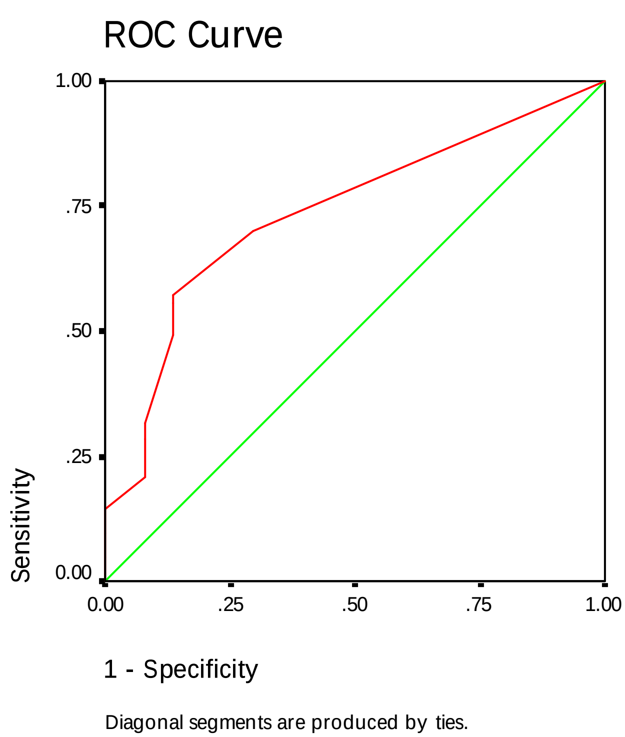


Figure 2. ROC curve for mathematics



Appendix

Components and performance indicators for three domains of

Work Sampling for Head Start (Dichtelmiller et al., 2001)

I. Language Development

A. Listening and understanding

1. Gains meaning by listening.
2. Follows two- or three-step directions.
3. Demonstrates phonological awareness.

B. Speaking and communicating

1. Speaks clearly enough to be understood without contextual clues.
2. Uses expanded vocabulary and language for a variety of purposes.

II. Literacy

A. Book knowledge and appreciation

1. Shows appreciation for books and reading.
2. Comprehends and responds to stories read aloud.

B. Print and alphabet awareness

1. Shows beginning understanding of concepts about print.
2. Begins to develop knowledge about letters.

C. Early writing

1. Represents ideas and stories through pictures, dictation, and play.
2. Understands purposes for writing.

3. Uses letter-like shapes, symbols, and letters to convey meaning.

III.. Mathematics

A. Problem solving

1. Begins to use simple strategies to solve mathematical problems.

B. Number and operations

1. Shows beginning understanding of number and quantity.

C. Geometry and spatial sense

1. Begins to recognize and describe the characteristics of shapes.
2. Shows understanding of and uses several positional words.

D. Patterns

1. Sorts objects into subgroups that vary by one or two characteristics.
2. Recognizes simple patterns and duplicates them.

E. Measurement

1. Orders, compares, and describes objects according to size, length, height, and weight.
2. Participates in measuring activities.