Using Observational Assessment to Evaluate Young Children's Learning: The Technical Quality of the *Work Sampling System*[1]

Samuel J. Meisels
Erikson Institute

June 2011

Propelled by dissatisfaction with conventional testing in early childhood and by a growing interest in performance assessments across all levels of instruction, formalized early childhood observational assessments began to appear during the 1990s and have now become an accepted methodology for assessing young children's learning (Meisels, 1996). This paper focuses on evidence concerning the validity and utility of the Work Sampling System (WSS; Meisels, Jablon, Dichtelmiller, Marsden, & Dorfman, 2001), an observational assessment for children from preschool (age 3) - Grade 6. More evidence is available about the psychometric properties of WSS than any other performance assessment used with young children. According to the publisher, it is in use annually with more than 850,000 children (mostly in Pre-K and Kindergarten) enrolled in WSS classrooms in Maryland, Minnesota, Pennsylvania, Georgia, South Carolina, Colorado, Arkansas, and Illinois, as well as in New York City and many other locations. A parallel observational assessment for birth - 42-month olds, called the Ounce Scale (Meisels, Dombro, Marsden, Weston, & Jewkes, 2003), is also in use nationally but will not be discussed here.

General Description

WSS is a curriculum-embedded, criterion-referenced performance assessment that is intended to document and evaluate what children are learning and have begun to master by providing information to teachers about individual students' academic, personal and social, and other cognitive and non-cognitive achievements. WSS is highly systematic in structure. It enables teachers to collect extensive information from multiple sources and use this information to evaluate what children know and can do. In its reliance on observing, recording, and evaluating, WSS organizes the assessment process so that it is both comprehensive in scope and manageable for teachers and students.

WSS is composed of three components: (1) Checklists and Guidelines/Standards; (2) Portfolios; and (3) Summary Reports. These elements are all classroom-focused and instructionally relevant, reflecting the objectives of the classroom teacher. Instead of providing only a snapshot of academic skills at a single point in time, WSS creates a continuous evaluation process designed to improve both the student's learning and the teacher's instructional practice. Although multiple customized adaptations of WSS have been created by the publisher for SEAs, LEAs, and Head Start, this memo will refer primarily to the original version of the assessment.

[1] Dr. Meisels is an author of two of the assessments mentioned in this memo—the Work Sampling System and the Ounce Scale—and a consultant to Pearson, the company that publishes them. These assessments are cited solely for illustrative purposes, rather than to promote them or imply an endorsement by a particular institution.

Checklists for each age level (preschool-sixth) consist of items that measure seven domains of development:

- *Personal and Social* (self concept, self control, approach to learning, interactions with others, conflict resolution),
- *Language and Literacy* (listening, speaking, literature and reading, writing, spelling),
- *Mathematical Thinking* (patterns, number concepts and operations, geometry and spatial relations, measurement, probability and statistics),
- *Scientific Thinking* (observing, investigating, questioning, predicting, explaining, forming conclusions),
- *Social Studies* (self, family, community, interdependence, rights and responsibilities, environment, the past),
- *The Arts* (expression and representation, appreciation), and
- *Physical Development and Health* (gross and fine motor, health and safety).

Each 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")* that is designed to help teachers document each student's performance. Accompanying the checklists are detailed developmental guidelines or standards. These content and performance standards present the rationale for each performance indicator and briefly outline reasonable expectations for children of that age. Examples show several ways children might demonstrate the skill or accomplishment represented by the indicator. The guidelines promote consistency of interpretation and evaluation among different teachers, children, and schools.

Portfolios are purposeful collections of children's work that illustrate children's efforts, progress, and achievements. These structured collections are intended to display the nature and quality of individual children's work and their progress over time. Both the child and teacher are involved in the design, selection, and evaluation of portfolios.

The Summary Report records evaluations of student progress and achievement for parents, teachers, and administrators three times per year. The summary report ratings are based on information about the child's progress and accomplishments across all domains. The report is available in a number of hard copy and web-based versions. The purpose of the report is to summarize student performance and progress and permit this evidence to be analyzed, aggregated, and reported to parents, administrators, policy makers, and others. The Summary Report typically takes the place of conventional report cards.

Teachers rate students' performance on each item of the checklist in comparison with national standards for children of the same grade in the fall, winter, and spring using a modified mastery scale: Not Yet, In Process, or Proficient. In the fall, winter, and spring, teachers also rate the portfolios and complete the hand-written or electronic summary report on which they summarize each child's performance in the seven domains. Teachers are asked to rate students' progress separately from performance on the Summary Report. All materials for families are available in Spanish and English and WSS has been used successfully with children whose first language is not English as well as with children with special needs.

Technical Support for Work Sampling

WSS has the strongest research base of any instrument of its type for young children. Although recent papers have explored the validity of the Ounce Scale with birth - 3 ½ year olds (Meisels, Wen, & Beachy-Quick, 2010) and WSS with 3- and 4-year olds (Meisels, Xue, & Shamblott, 2008), the psychometric qualities of WSS have been studied most extensively with children in Kindergarten – Grade 3 and are the focus of this brief report.

In order to answer the overall question of "Can teachers of young children use observational assessments accurately?" (in other words, Can we trust teachers' judgments?) a cross-sectional study was conducted in 17 Title I classrooms (N = 345 students, K- 3) in the Pittsburgh Public Schools. Most of the children (70%) were African-American, 80% qualified for Free and Reduced Lunch, half were male, and 8% had IEPs. WSS ratings were compared with student scores on the Woodcock Johnson Psychoeducational Battery-Revised™ (WJ-R) in order to examine construct and predictive aspects of validity (see Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001; Meisels, Atkins-Burnett, Xue, Bickel, Nicholson, & Son, 2003). In addition, levels of teacher implementation, understanding, and satisfaction, and evaluations of family understanding and satisfaction were investigated by means of survey questionnaires (see Meisels, Xue, Bickel, Nicholson, & Atkins-Burnett, 2001).

Major research questions and findings are as follows:

1. *Is WSS a valid means of evaluating student achievement and progress?* Correlations between WSS Checklist ratings in literacy and mathematical thinking and standardized test scores were moderate to high, demonstrating the concurrent and predictive validity of WSS and suggesting that it is an effective marker of student learning. Over three-quarters of the correlations were between .50 – .75. Four-step hierarchical regressions show that WSS ratings were a stronger predictor of test scores than demographic variables.

2. *Can WSS scores discriminate at-risk from not at-risk students?* Data from a Receiver-Operating-Characteristic Curve (ROC) analysis show that WSS can discriminate between children who are at-risk and those not at-risk. ROC analysis allows us to examine the probability of a student performing similarly on both WSS and WJ-R. The findings show that more than 80% of those scoring low on either the math or reading portions of the WJ-R performed poorly on WSS as well.

3. *How well do teachers understand and implement WSS and how satisfied are teachers with it?* Teachers reported a high level of understanding and implementation of WSS on surveys. The majority were positive about WSS and satisfaction increased with experience with the WSS.

4. *How well do families understand WSS and how satisfied are they?* More than 240 parents participated in this survey (70% return rate). The majority (> 66%) reported very positive ratings of WSS; satisfaction was linked to understanding of WSS. Almost two-thirds preferred WSS to traditional report cards with letter grades and the majority wanted to continue receiving a WSS Summary Report.

5. *How do families describe the major benefits of WSS?* Families reported that WSS helped them understand their children's school work and learning. They also said that WSS helped children understand their own learning and achievement. Structural equation modeling was used to examine the direct and indirect effect of parents' perceptions of teachers' willingness to use WSS and other factors on parents' overall satisfaction. Results demonstrated that parents hold positive attitudes towards WSS and believe that WSS benefits their children.

6. *How do students in WSS classrooms perform on standardized achievement tests?* A study of the trajectory of change in scores of WSS and non-WSS third and fourth graders on the Iowa Tests of Basic Skills (ITBS) compared students exposed to WSS since Kindergarten with those in a group of non-WSS contrast schools that were matched by race, income, mobility, school size, and number of parents in the home. A second comparison group consisted of all other students in the school district. Comparisons of mean change in reading and math scores on the ITBS and regression analyses were conducted in order to study variance in test scores from one year to the next across the three groups. Results indicated that students who were in WSS classrooms displayed growth in reading from one year to the next that significantly 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 for math, though less robust. The impact of WSS was not limited solely to those students who started with either low or high skills but was generalized across the entire sample.

An earlier study of the reliability and validity of WSS presented data concerning the reliability and validity of the field-trial version of WSS with 100 kindergarten-age children who were administered the WJ-R in the fall and spring (Meisels, Liaw, Dorfman, & Nelson, 1995). Results showed high internal and moderately high inter-rater reliability for the Checklist and Portfolio (alphas = .84 – .95). WSS accurately predicted performance on the WJ-R, even when the potential effects of gender, maturation (age), and initial ability were controlled. Predictive validity evidence from correlations between teacher ratings and student academic performance ranged from .67 - .76.

Other Studies

Additional studies of WSS have been conducted by Gallant (2009) and Gallant and Moore (2008a; 2008b). These studies are important because they were performed by independent investigators, their sample sizes were large, and they explored issues of sensitivity and potential bias. In particular, they examine the predictive nature of WSS in relation to a third grade high-stakes state mandated test (Gallant, 2009); the extent to which WSS items may function differently for White and African-American male students (Gallant & Moore, 2008a); and whether ethnic-based differences exist in teacher ratings of White and African-American students (Gallant & Moore, 2008b).

Predictive Validity. This study included 1281 students (81.3% non-white, 71% eligible for reduced price lunch). WSS Language and Literacy and Mathematical Thinking scores in first grade were compared with performance on a third grade criterion-referenced high-

stakes test in math and literacy using a multilevel modeling approach. Findings showed positive moderate associations across the two year span for both domains. The study also showed that when student demographic variables were controlled, WSS scores were significant, positive predictors of later achievement. In the words of the study, "we can expect approximately a one point increase in third grade achievement scores for each one point increase in first grade readiness assessment [WSS] domain scores" (p. 140).

Impact of Ethnicity on Item Functioning. Focusing on 732 urban first-grade African-American males and 120 urban White males, this study considered whether ethnicity had an impact on teachers' ratings of male students. An ordinal logistic regression procedure was used to investigate differential item functioning [DIF] on performance of these students on WSS language and literacy and mathematics domains. Differences in ratings between ethnic groups were detected, but these differences were small. When students were matched on ability level, the indicators did not function differently for ethnicity groups.

Ethnic-Based Equity in Teacher Judgment. Focused strictly on the language and literacy domain of WSS, this study included 1442 African-American and 319 White, urban Grade 1 students. Its purpose was to explore the extent to which ethnic-based differences exist in WSS teacher ratings. Using an ordinal logistic regression procedure, three findings emerged. First, teachers demonstrated a high level of consistency in rating students (Cronbach alpha = .96). Second, a discrepancy between White and African-American students was detected, with a greater proportion of African-American students scoring lower than White students. This finding is consistent with other research about gaps in achievement between White and African-American students and is further clarified by the third finding: the absence of DIF in teacher ratings on the indicators. The language and literacy domain score accounted for 70 – 86% of the total variation in teacher ratings; student ethnicity explained only a small amount of variation in teacher ratings.

These three interrelated studies demonstrate that WSS functions well across ethnic groups and, in particular, is not problematic when teachers rate African-American males. One of the issues raised by critics of observational assessment is that it will heighten racial and ethnic disparities because this approach is so dependent on teacher judgment. These studies are encouraging in that they suggest that an observational assessment that is as highly structured and connected to standards as is WSS is less likely to demonstrate these problems in equity.

State Studies of WSS

Since so many states have adapted WSS for their own use, it is critical that data be collected to evaluate the accuracy of the judgments made with these customized variations. One such study was performed for the state of Minnesota by Arthur Reynolds and associates (Human Capital Research Collaborative, 2011). The Minnesota Department of Education has been using a shortened version (32 items representing five domains) of the P-4 checklist to evaluate children's school readiness at the outset of Kindergarten. The checklist is completed annually on a random sample of 10% of the state's schools, stratified by region. Based on data from Kindergarten cohorts in 2003, 2004, and 2006 (total N = 12,552), the researchers found that Minnesota WSS checklist predicted third grade Minnesota

Comprehensive Assessment results significantly and consistently in reading and math. Other findings were as follows:

- A factor analysis showed that the checklist items were best represented by one overall school readiness dimension, rather than the five domains from which the items were selected.
- The internal reliability was high at .98.
- Students who had higher WSS Kindergarten scores also had higher scores on the third grade criterion-referenced state test.
- Holding constant gender, race/ethnicity, parent education, income, and IEP status in Kindergarten, those children who were proficient on Language and Literacy and on Mathematical Thinking in Kindergarten were two to three times as likely to meet or exceed state reading and math proficiency in Grade 3 as were Kindergarteners who were not proficient on WSS.

Reliability

One of the major issues confronting WSS is reliability. An assessment such as WSS does not lend itself to the collection of conventional inter-rater reliability data. This is because inter-rater reliability requires observations by two or more raters who are independent of one another. Classical definitions of reliability describe it in terms of independent measurements of the same phenomenon. But in a curriculum-embedded assessment, such independence is virtually impossible to obtain. WSS relies on multiple observations made by an individual who has close contact with a child over a substantial period of time. In those cases where there may be co-teachers or two teachers working in a head teacher/assistant teacher relationship, it is highly unlikely that their judgments would be independent of one another. Moreover, we cannot avoid this problem simply by introducing an "independent" observer into the classrooms solely for reliability purposes, even if we make the improbable assumption that such an individual will have the same information as the person who works with the child on a continuing basis. If there is a discrepancy between the teacher and the observer, it is impossible to say whether that discrepancy is due to a problem in the construction of the assessment or is a result of differences in role and perspective of the teacher and observer.

Studies of internal reliability can be done and have been conducted. One such study occurred in Maryland, which has been using a customized version of 30 P-4 WSS items for more than a decade in their Maryland Model of School Readiness (MMSR) program (see Maryland State Department of Education, 2009). The state recently reported reliability data for 57,775 children who were administered the MMSR in fall, 2008. Among the findings were the following:

- A preliminary correlation analysis of students' composite scores with WSS's seven domain scores showed that individual student ratings are relatively independent of their school or LEA. In other words, this analysis allows us to conclude that no systematic bias in assessment was detected in this large sample.
- The coefficient alpha, or internal reliability, was shown to be very high: .966 ($p < .05$).
- Since inter-rater reliability is not practical or feasible, as discussed above, split-half reliability coefficients were computed. In this procedure the scale is split into two parts

and the correlation between the parts is examined. The alpha for part 1 was .944 and for part 2 was .938. Two related statistics were also determined. The Spearman-Brown reliability coefficient was .918 and the Guttman split half was .914. These results indicate a high reliability for WSS.

- A Guttman item-scale analysis was also performed. This procedure records the average values of students' total scores if a particular item is removed from the assessment. This information is useful in determining the relative influence of each item on WSS. Results show that the internal reliability, as measured by Cronbach's Alpha is virtually unchanged regardless of whether any given item is deleted from WSS, thus demonstrating excellent stability of the assessment.

Other Considerations and Future Research

A number of issues remain unresolved by the studies reported above. For example, there is need for replication. The studies described above should be repeated with larger samples drawn from a larger universe of teachers and conducted by those independent of the authors.

Other areas in need of research include the validity—that is, the construct representation—of domains other than language and literacy and mathematical thinking. This work has not taken place principally because the outcome measures in early childhood for other domains are not well established.

Another aspect that has not been fully explored is the clinical utility of WSS, or the relationship between WSS and teacher planning and implementation of curriculum. Using a qualitative interview methodology, as well as observations, much can be learned about how teachers translate formative information from WSS into data about teaching.

Finally, we know very little about the longitudinal impact of WSS. It would be of interest to explore whether student performance on WSS is in some way influenced by length of time in WSS classrooms so that children who may have spent several years in WSS classrooms differ from those who are the same age and have similar demographic characteristics, but are only in their first year with WSS. The issue in need of study and clarification is whether the impact of WSS varies not only with teacher familiarity with WSS but also with length of time that children have been exposed to it.

References

Gallant, D. J. (2009). Predictive validity evidence for an assessment program based on the Work Sampling System in mathematics and language and literacy. *Early Childhood Research Quarterly, 24*, 133-141.

Gallant, D. J. & Moore, J. L. (2008a). Assessing ethnicity: Equity for first-grade male students on a curriculum-embedded performance assessment. *Urban Education, 43* (2), 172-188.

Gallant, D. J. & Moore, J. L. (2008b). Ethnic-based equity in teacher judgment of student achievement on a language and literacy curriculum-embedded performance assessment for children in Grade One. *Educational Foundations, 22*(1-2), 63-77.

Maryland State Department of Education (2009). *Reliability Analysis: Maryland Kindergarten students' MMSR Kindergarten assessment data.* Available at http://www.marylandpublicschools.org/NR/rdonlyres/264017F7-E1A1-469B-8C9D-F824AAF21159/23294/Reliability_0809.pdf.

Meisels, S. J. (1996). Performance in context: Assessing children's achievement at the outset of school. In A. J. Sameroff & M. M. Haith (Eds.), *The five to seven year shift: The age of reason and responsibility* (pp. 410 - 431). Chicago: University of Chicago Press.

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., Dombro, A. L., Marsden, B. B., Weston, D. & Jewkes, A. (2003). *The Ounce Scale: An Observational Assessment for Infants, Toddlers, and Families*. New York: Pearson Early Learning.

Meisels, S. J., Jablon, J. R., Marsden, D. B., Dichtelmiller, M. L., 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. Wen, X., & Beachy-Quick, K. (2010) Authentic Assessment for Infants and Toddlers: Exploring the Reliability and Validity of the Ounce Scale. *Applied Developmental Science, 14* (2),1 – 17.

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.

Meisels, S. J., Xue, Y., & Shamblott, M. (2008). Assessing language, literacy, and mathematics skills with Work Sampling for Head Start with preschoolers. *Early Education and Development, 19* (6), 963-981.

Reynolds, A.T., et al. (2011). *Assessing the validity of Minnesota school readiness indicators.* Minneapolis, MN: Human Capital Research Collaborative. Available at http://humancapitalrc.org/mn_school_readiness_indicators.pdf.