Education Policy Analysis Archives

Volume 11 Number 9 **February 28, 2003** **ISSN 1068-2341**

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the ERIC Clearinghouse on Assessment and Evaluation and are permanently archived in *Resources in Education*.

Creating a System of Accountability: The Impact of Instructional Assessment on Elementary Children's Achievement Test Scores

Samuel J. Meisels
Erikson Institute

Sally Atkins-Burnett
University of Michigan

Yange Xue
Julie Nicholson
Palo Alto, CA

Donna DiPrima Bickel
University of Pittsburgh

Seung-Hee Son
University of Michigan

Citation: Meisels, S. J., Atkins-Burnett, S., Xue, Y., Nicholson, J., Bickel, D. D., and Son, S-H. (2003, February 28). Creating a system of accountability: The impact of instructional assessment on elementary children's achievement test scores, *Education Policy Analysis Archives, 11*(9). Retrieved [Date] from http://epaa.asu.edu/epaa/v11n9/.

Abstract

This study examined the trajectory of change in scores on the Iowa Tests of Basic Skills (ITBS) of low-income, urban, third and fourth graders who had been enrolled in classrooms where the Work Sampling System (WSS), a curriculum-embedded performance assessment, was used for at least three years. The ITBS scores of children exposed to WSS were compared with those of students in a group of non-WSS contrast schools that were matched by race, income, mobility, school size, and number of parents in the home and to a comparison group 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 as well as the average change shown by all other students in the district. Children in WSS classrooms made greater gains in math than children in the other two groups, although the results were only marginally significant when compared with gains by the matched contrast group. The discussion concerns the complementarity of performance-based and normative tests in systems of accountability and the potential value of using a curriculum-embedded assessment to enhance teaching, improve learning, and increase scores on conventional accountability examinations.

More group-administered achievement testing is taking place in states and local school districts than ever before (Achieve, Inc., 2002). According to a survey published in *Education Week* (Olson, 2001), every state has adopted mandatory tests at one or more grades in elementary, middle, and high school, and 49 states have linked their academic standards to these tests. Moreover, legislation mandating annual testing in reading and mathematics for all children in grades 3 – 8 was recently enacted by Congress.

The tests in use nationwide are standards-based, but primarily norm-referenced; only 10 states report use of supplementary criterion-referenced tests (Olson, 2001). The principal purpose of these tests is to ascertain the current status of student achievement, rather than to identify students in need of intervention or to determine appropriate instructional strategies. Only seven states provide extra funding for low-performing schools and just nine states allocate funds for remediation of failing students. By the close of the year 2000, 18 states had made graduation contingent on student test performance and an additional five states were about to begin administering exit exams. As many as 27 states could withhold diplomas to students who fail to pass state accountability examinations by 2003 (Olson, 2001).

In addition to the increased prevalence of such tests and the escalation of consequences or "stakes" associated with them, the most notable change in the assessments themselves

concerns their alignment with curriculum standards. Never before in U. S. history have we witnessed such an explosion of attention to standards. The minimum competency testing movement of the 1980s that was inaugurated by the *Nation At Risk* report (National Commission on Excellence in Education, 1983) has been overtaken by a focus on high standards of achievement inspired by the nation's embrace of the national educational goals of the 1990s.

Ironically, one result of the standards-based reform movement has been a heightened emphasis on high-stakes, group-administered, decontextualized testing practices. This is due to the linkage between standards-based reform and the national political drive for accountability (Kohn, 2000). As Popham (2000) points out, these tests lend themselves to the "score-boosting game" in which educators devote most of their energy to raising students' scores on conventional achievement tests. Rather than trying to improve student performance by enhancing instruction, this approach views high stakes testing as a policy tool to leverage learning (Firestone & Mayorowetz, 2000). In short, funding, availability of other resources, and state and local prestige are all devoted to improving student test scores by changing the curriculum to match more closely the items or content standards of the assessments. This practice of "curriculum alignment" has the consequence of increasing test scores for some students, but often leaving general knowledge and mastery of curriculum domains virtually untouched (Amrein & Berliner, 2002; Corbett & Wilson, 1991; Haney, 2000; McNeil, 2001). High-stakes testing has been shown to impair the validity of the tests themselves as the test-taking experience becomes less a sampling of students' adaptive skills and higher-order thinking and more of an exercise in rote memory and mastery of basic skills (Orfield & Kornhaber, 2001). In a study of 18 states with high-stakes testing programs, the learning levels of the students in all but one of the states were at the same level as before the testing policies were implemented (Amrein & Berliner, 2002).

The alternative to teaching to the test, or measurement-driven instruction, is to transform instruction, but to do so in such a way that the standards that are intended to serve as the basis of the tests inform instructional decisions and are incorporated into new forms of assessment. Curriculum-embedded performance assessments (see Baron & Wolf, 1996; Darling-Hammond, 1992) represent an instructional-driven measurement in which students' actual classroom performance is evaluated in terms of standards-infused criteria. These criteria in turn suggest next steps in curriculum development which are consistent with advancing progress toward attainment of the defined standard. It is reasonable to assume that as students' learning improves, so will their scores on accountability examinations.

Unfortunately, the research literature contains few studies of the impact of curriculum-embedded performance assessments on group-administered achievement test scores (Borko, Flory, & Cumbo, 1993; Falk & Darling-Hammond, 1993). Rather, most studies of the impact of testing provide data about conventional results-driven accountability tests—tests that rely on public reporting of performance data and utilization of these data for reward or sanction. This approach to assessment (which can also be called "norm-referenced accountability") emphasizes the use of test data for instrumental purposes—purposes external to the classroom—rather than direct application of the data to improve educational practice.

In this paper we describe an alternative to typical conceptions of accountability. Instead of relying on either norm-referenced or performance-based assessments in isolation, we suggest a complementary approach that incorporates both types of assessment. In short, consistent with recent federal and state initiatives, we suggest the addition of performance assessments

to conventional norm-referenced testing.

This paper investigates whether students who are enrolled in classrooms in which a curriculum-embedded performance assessment is in use will show greater gains on a conventional test used for accountability than students who have not had exposure to the performance assessment. The research question this study poses is the following: Can ongoing, focused instructional assessments influence performance on group-administered achievement tests? A corollary to this question concerns whether instructional and high-stakes assessments can be linked to create an accountability system that relies on both classroom- and test-based information about student achievement.

Method

Procedures

This study is part of a larger investigation that evaluated the validity of a curriculum-embedded performance assessment—the Work Sampling System (WSS; Meisels, Jablon, Marsden, Dichtelmiller, & Dorfman, 1994, 2001)—and its influences on teacher practices and children's achievement in the Pittsburgh Public Schools (PPS). The overall study included data from teachers, parents, and children. This paper focuses on the impact of WSS on the trajectory of children's change in scores on a group-administered achievement test (the Iowa Tests of Basic Skills [ITBS]) from grade 3 to grade 4. Related studies focus on other aspects of the validity of WSS (Meisels, Bickel, Nicholson, Xue, & Atkins-Burnett, 2001), parental reactions to WSS (Meisels, Xue, Bickel, Nicholson, & Atkins-Burnett, 2001), and teachers' views of the consequences of using WSS (Nicholson, 2000).

The Work Sampling System (Meisels, 1997; Meisels et al., 1994; 2001) is a curriculum-embedded performance assessment designed for children from preschool through grade 5. WSS is comprised of developmental guidelines and checklists, portfolios, and summary reports. It uses teachers' perceptions of their students in actual classroom situations as the data of assessment while simultaneously informing, expanding, and structuring those perceptions. It involves students and parents in the learning and assessment process and it makes possible a systematic documentation of what children are learning and how teachers are teaching. This approach to performance assessment allows teachers the opportunity to learn about children's processes of learning by documenting children's interactions with materials, adults, and peers in the classroom environment and using this documentation as the basis for evaluating children's achievements and planning future educational interventions through comparisons with standards-based guidelines. Evidence of the reliability and validity of Work Sampling is available in Meisels, Bickel, Nicholson, Xue, and Atkins-Burnett (2001) and Meisels, Liaw, Dorfman, and Nelson (1995). Further descriptions of WSS are found in Meisels (1996, 1997) and Meisels, Dorfman, and Steele (1995).

The Iowa Test of Basic Skills (ITBS; University of Iowa and Riverside Publishing, 1994) is a group-administered achievement test designed to monitor year-to-year achievement differences in students from K – Grade 12. Norming for the 1993 edition (Form K) was completed on 136,934 individuals. KR-20 internal consistency ratings are reported at >.84 for all reading and mathematics subtests. The third and fourth graders in this study were administered the reading comprehension and the vocabulary subtests of the ITBS Survey

Battery (Levels 7 - 11). The Reading total score includes both the comprehension and vocabulary scores. The Mathematics total score includes items tapping computation, estimation, calculation, problem solving, and data interpretation (Levels 7 - 11). Developmental Standard Scores (DSS) were calculated using the raw score/DSS conversion tables provided by ITBS (Hoover et al., 1993). The DSS scales allow a comparison of both status and growth even when children take different level tests. The mathematics computation score is combined with the total score to determine the DSS for mathematics plus computation.

Design

This report describes the results of a natural experiment in which two groups of demographically matched students who were administered the ITBS were compared with one another and with all other students in their grade levels in the PPS who were also administered the ITBS. One of the target groups was composed of students who had been exposed to WSS for three years prior to being administered the ITBS; all other students had no experience with WSS. The ITBS was administered to all children by their teachers in the spring of 1997 (end of third grade) and the spring of 1998 (end of fourth grade). A total of 2708 students received the ITBS reading assessments in both 1997 and 1998 and 2564 students took the ITBS math assessments in both years. Data were coded and reported by school district personnel.

A longitudinal design was selected because the schools using WSS were among the lowest-performing schools on the ITBS in the district. It was evident to us that comparisons of absolute scores at the end of third or fourth grade would only confirm that the children in these low-scoring schools were still low scoring, despite potential improvements in comparison with children from the same schools but from different age cohorts. The longitudinal design focuses on the trajectory of change from third to fourth grade as a way of capturing growth over time *within* students. This intra-individual use of normative data enabled us to examine the relative change in student achievement without regard for comparisons of absolute differences in student scores.

WSS was adopted as part of the District's restructuring effort, but it was not the sole innovation taking place in the district. New reading, mathematics, and social studies curricula were introduced at the same time as WSS, however, not all schools and classrooms in the district implemented all of these practices. Since information was unavailable about which schools implemented which instructional innovations, it is not possible to test alternative explanations for our results by isolating WSS versus any of these interventions.

Sample

At the time this study took place WSS had been used in a sample of the Pittsburgh Schools for three years. All teachers in the WSS schools were voluntary participants. Selection criteria for participation of teachers in this study included use of WSS for at least two years and a determination of the fidelity of implementation of participating teachers. This was ensured by a review by external examiners conducted in the spring of 1996 of portfolios and of the teachers' 1996–97 WSS materials by the research staff. These selection criteria limit the generalizability of our results but provide a test of the full implementation of WSS.

The longitudinal data consist of 96 third grade students in the WSS schools, 116 students in

the non-WSS comparison schools, and 2922 students enrolled in all other PPS Grade 3 and 4 classrooms in 1996–98. For students in WSS schools, 71% were African-American and 90% received free or reduced lunch (see Table 1). There were more girls (58%) than boys in the sample. To form the comparison group, classrooms were chosen that matched those in the WSS schools as closely as possible on race, income, mobility, school size, and number of parents in the home. In other PPS schools, 70% of the students were African-American and 87% received free or reduced lunch (see Table 1).

Table 1
Demographic characteristics of WSS schools, comparison schools, and other PPS

Group	School Size (N)	F/R Lunch (%)	Other Parents in Household (%)	Mobility (%)	African-American (%)
WSS	400.4	90.0	70.6	9.8	70.6
Comparison	298.8	89.8	74.0	9.4	75.2
PPS Other	311.2	87.0	70.8	9.4	71.5

WSS = Work Sampling System
PPS = Pittsburgh Public Schools

Analysis

Comparisons of mean change in reading and math scores on the ITBS from Grades 3 to 4, as well as regression analyses, were conducted in order to study the average change in test scores from one year to the next among the three groups. As noted, analysis of previous school district results showed that the WSS schools scored at or near the bottom of the district's ITBS test scores. By controlling for initial ability we were able to study the trajectory of change in students' ITBS scores from one year to the next in order to determine if differential change on the ITBS took place despite the low levels of initial competence on this assessment shown by students in the WSS schools.

Our analytic approach began with a comparison of the change in ITBS scores of the WSS third graders in 1997 - 1998 as compared with a comparison group matched on key demographic characteristics and all other PPS third graders in that cohort. Differences between the 1997 and 1998 ITBS Developmental Standard Scores (DSS) were computed to create DSS change scores, which we used as one indicator of academic growth. In order to examine whether WSS had a differential impact on high or low achievers, the means of the 1997 DSS scores for the PPS district were used to divide all the PPS students into above average and below average groups. The gains for high and low achieving students within each group (WSS, comparison, and all other PPS) were then compared.

Next, three-step hierarchical regressions were performed to examine the relative effect of participating in WSS on the students' change in performance from third to fourth grade. Because change scores may be particularly sensitive to problems with floor and ceiling, we used a covariance model with the Grade 4 score as the outcome, controlling for initial ability

and the level of form administered. In the first step, children's 1997 DSSs were entered to control for differences in initial ability levels. Typically, fourth graders take a Level 10 form, however, PPS allows for administration of out-of-level tests. Although the publishers equated the forms, and the level of administration should have had minimal effect on the DSS, we entered the level form into the regression to ensure that results were not biased due to differences in forms or variations in administration. Children taking a below-grade form might be more apt to reach the ceiling of the measure and receive an inflated estimated ability. Only three children in math and two children in reading took a form above the grade level; we excluded these cases from analyses. For the remaining students we created a dummy variable that was entered in the second step of the regression to indicate if the student received a below-grade level form for the Grade 4 administration. (The percentage of students in below-grade level forms was similar in the WSS and all other PPS groups—10 and 11% respectively. Only 6% of students in the comparison schools took below-levels forms.) Finally, students' group membership (comparison and all other PPS with WSS as the referent group) was entered into the regression model. Since dependent variables (1998 DSS scores) and the children's initial scores (1997 DSS scores) were standardized. The regression coefficients can be interpreted as effect sizes.

Missing Data

In order to study the impact of missing data on our conclusions, the means of the students with missing subtests were compared by group membership. In 1997, 21.9% of the WSS students were missing the reading total score; 15.6% were missing the 1998 reading total score. The comparison group was missing 13.8% of its reading total scores in 1997 and 17.2% in 1998. Among the other PPS students, 3.7% were missing the 1997 reading total score and 11.2% were missing the 1998 reading total score. Due to a missing score in either of the years, 30.2% of the WSS students, 17.2% of the comparison students, and 13.4% of other PPS students were excluded from the reading change score analysis. In mathematics, 7.3% of the WSS group was missing the 1997 math total score compared with 13.8% of the comparison group and 5.7% of the other PPS. For the 1998 math total score 18.8% of the WSS group, 17.2% of the comparison group, and 16.1% of the other PPS students had missing data. Thus, 18.8% of the WSS group, 22.4% of the comparison group, and 19% of the other PPS group were not included in the analysis of the mathematics change score.

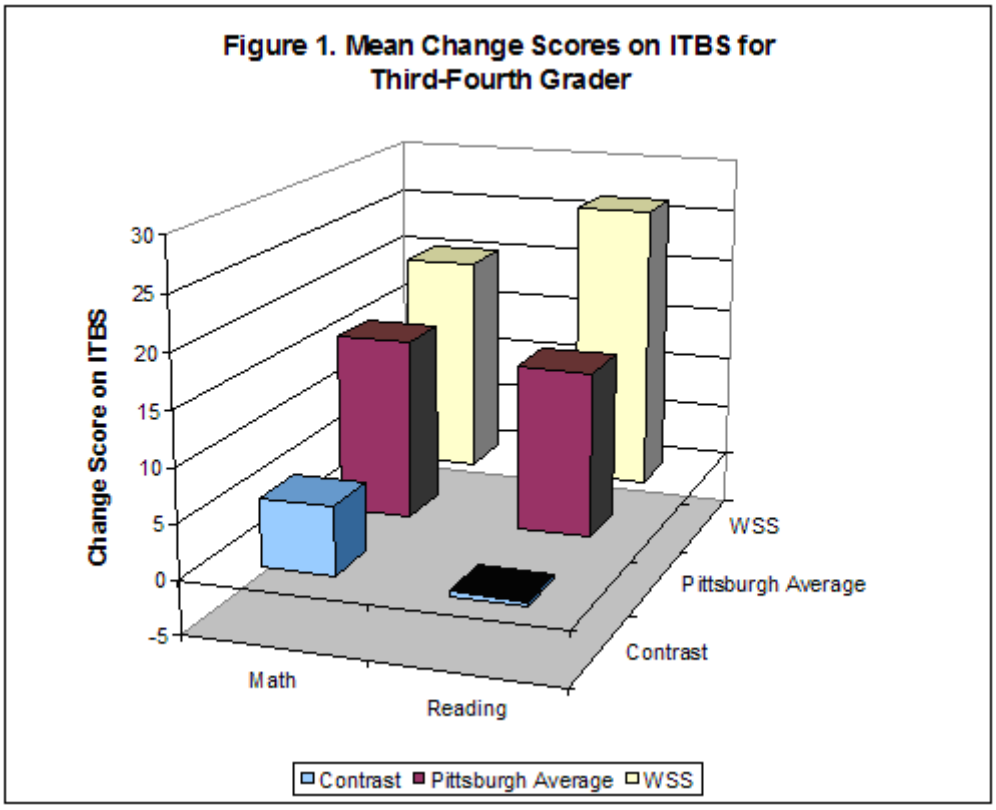
For the WSS group, we compared students' missing scores for one of those years with students who had scores for both years in terms of gender, ethnicity, age, SES, and Woodcock Johnson Psychoeducational Battery-Revised (WJ-R) broad reading, broad math, and broad writing scores in third grade. (The WJ-R was available only for the WSS group.) The results showed no significant differences across these variables except that fewer students who were missing at least one mathematics score received free or reduced lunch (58% vs. 92%, $p < .01$). There were no gender or race differences between missing and non-missing students for the comparison group.

The means for the ITBS subtests that were taken by the missing students were compared to the means of the other students in that group. For the WSS group, the means on the mathematics and reading total scores were not significantly different between the missing and non-missing groups. However, the 1997 reading DSS score was significantly lower for the comparison group students who were missing the reading subtests in one of those years (171 vs. 197, $p < .05$), but there were no significant differences in mathematics total score between the missing and non-missing students for the comparison group. For the other PPS

students, the reading total scores in both 1997 and 1998 were significantly lower among the missing group (173 vs. 181 in 1997, $p < .001$; 186 vs. 197 in 1998, $p < .01$), but the means on the mathematics total score were not significantly different for the missing and non-missing groups. Thus, the effects of the missing data on our findings were relatively limited.

Results

In this analysis we compare ITBS scores of third and fourth grade WSS and non-WSS students (both the matched comparison group and all other PPS students) in order to determine if differential achievement on this outcome is associated with participation in WSS classrooms. We begin by comparing the change in mean DSS in reading and math among the WSS students, the comparison group, and the remainder of the PPS students (see Figure 1). The mean change scores of the WSS group (27 and 20 points for reading and math, respectively) are substantially greater than those of the other groups (0 and 6 for the comparison and 15 and 17 points for all other PPS in reading and math, respectively). The differences in the groups are particularly strong in reading, with a mean change on the reading total DSS that is more than 11 points greater for the WSS students when compared with the other PPS students ($t = 4.33$; $p < .001$). The moderately large effect size of .68 indicates meaningful as well as significant differences in reading change scores (Cohen, 1988). When considered in relation to the comparison group, the discrepancy is even greater. The WSS group mean change score is more than 27 points higher than the comparison group's mean change score ($t = 8.86$; $p < .001$). The unusually large effect size ($d = 1.60$) indicates strong differences in the sample in measured change in reading.



In math, results follow a similar pattern, although they are not as dramatic. The mean change score of the WSS students is greater than that of the other PPS students by more than three

points, a marginally significant finding (t = 1.89; p = .059). The examination of the effect size (d = .20) suggests a small but nontrivial effect (Cohen, 1988). The mean change of DSS math score in WSS students is almost 14 points higher than that of the comparison group (t = 4.88; p < .001), indicating a large effect (d = .76).

To investigate whether WSS had a differential effect on high and low achieving students, we used a segmentation analysis. The mean 1997 DSS ratings of the entire sample were used to divide all students into above and below average groups (see Table 2). Before comparing differential changes of score in these groups, we examined the initial scores of the 1997 DSS. As expected, the initial scores of the WSS students were lower than the comparable group (above and below average) in the other samples (comparison and other PPS), although the mean differences were relatively small (2-4 point differences) and not highly significant, with the exception of the above average group of comparison students in reading. The mean of the 1997 reading DSS of the above average students differed by more than one SD between WSS and the comparison group and the above average performers represented a greater percentage of the comparison sample than was the case in the WSS sample.

Table 2
Mean of 1997 DSS in above and below average performers

Subtest	WSS			Comparison			Other PPS			WSS vs. Comparison	WSS vs. Other PPS
	M	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*t*	*t*
Reading total											
Above Average	191.40	11.05	10	205.36	17.15	77	199.91	15.90	1295	-3.49**	-1.69
Below Average	160.69	13.17	65	164.81	13.87	31	164.12	12.13	1519	-1.41	-2.22*
Math total											
Above Average	197.58	14.56	24	202.31	15.48	84	199.30	13.31	1534	-1.34	-.63
Below Average	163.57	9.95	65	168.61	9.99	23	167.19	9.42	1377	-2.09*	-3.17**

Note. t = t-score, DSS = Developmental Standard Score
* p <.05, ** p < .01

With one exception, above and below average WSS students in reading and math made gains that were greater than the comparison group and the other PPS students (see Table 3). With small differences in initial ability in all but one area, the effect sizes of the differences in change scores were all moderate to high in reading. In mathematics the results were more equivocal. Among below average math achievers, WSS had significantly higher change scores than other PPS students (t = 2.14; p < .05; d = .29). The change scores of low performing PPS students in math were greater than the change scores of either the WSS or the comparison group of low performers. However, these results were not significant and the effect size was negligible (d = .05).

Table 3
Mean of DSS change in above and below average performers (SD)

Subtest	WSS			Comparison			Other PPS			*WSS vs Comparison*		*WSS vs Other PPS*	
	M	*SD*	*n*	*M*	*SD*	*n*	*M*	*SD*	*n*	*t*	*d*	*t*	*d*
Reading total													
High	26.56	20.44	9	-4.01	17.42	72	12.29	16.78	1193	4.87***	1.77	2.54*	0.82
Low	26.67	21.67	58	10.50	12.71	24	17.93	15.44	1337	3.41***	1.02	3.04**	0.55
t (high vs. low)	.02			3.76***			8.76***						
d (high vs. low)	.00			.83			.35						
Math total													
High	8.32	23.31	19	2.77	6.26	70	12.91	23.31	1138	.94	.24	-.25	.05
Low	24.25	17.18	59	18.85	15.49	20	20.04	16.79	1230	1.21	.32	2.14*	.29
t (high vs. low)	3.21**			3.94***			10.31***						
d (high vs. low)	.80			.83			.39						

Note. t = t score, d = standardized estimates of effect size
*p < .05, ** p < .01, *** p < .001

The publishers of the ITBS contend that children with above average ability will show greater gains than other children from year to year. However, in each of the three samples, lower performing students made greater gains than higher performers. The difference in the mean reading change score between high and low performing students was greater in the comparison group (14.51) and the PPS group (5.64) than in the WSS group (0.12). The WSS group did not show substantial differences in reading score change between high and low performers (t = .02; p = N.S), although there were significant differences between high and low performing students in the comparison group (t = 3.76; p < .001; d = .83) and the other PPS group (t = 8.76; p < .001; d = .35). This suggests that high and low performing students profit equally from WSS in reading. In mathematics, the other PPS group demonstrated the least difference in change scores between high and low performing students (7.13) compared with the WSS group (15.94) and the comparison group (16.08). The difference in change scores between high and low performing students was significant in WSS (t = 3.21; p < .01; d = .80) and the comparison group (t = 3.94; p <. 001; d = .83), as well as the other PPS groups (t = 10.31; p <. 001, d = .39).

To examine whether group membership accounts for differences in achievement growth after controlling for differences in both initial achievement (i.e., 1997 DSS) and the level form taken in 1998, three step regressions were performed .The initial achievement and the 1998 level form (below-grade level form = 1, on-grade level form = 0) were entered in the first and the second step, respectively. Then, group membership was entered in the third step. All three models were statistically significant (see Table 4). Initial ability, level of form administered, and membership in WSS all predicted students' fourth grade reading DSS (see Table 4). Children who took on-grade level forms had fourth grade reading DSS more than one-half SD higher than those who took below-grade level forms, even after controlling for initial ability (p < .001). The children who were in WSS schools had higher fourth grade scores in reading by .17 SDs when compared with the other PPS students, after controlling for the effect of initial achievement and the level form taken in 1998 (p < .05). The average fourth grade score for the WSS group was .60 SDs greater than the average score for the comparison group after controlling for initial achievement and 1998 level form (p < .001)

Table 4

Covariance models of Grade 4 achievement

	Grade 4 DSS Reading Total (N=2,772)			Grade 4 DSS Mathematics Total (N=2828)		
Variable	Step 1	Step 2	Step 3	Step 1	Step 2	Step 3
Initial ability[a]	.74 ***	.67***	.68***	.56***	.56***	.56***
1998 Level Form (below-grade level)		-.55***	-.54***		-.10*	-.10*
Other PPS group			-.17*			-.06
Comparison group			-.60***			-.18
R^2	.55***	.58***	.58***	.32***	.32***	.32***
R^2 change		.03***	.01***		.00*	.00

Note. All regression coefficients are shown in effect size using standardized scores.
DSS = Developmental Standard Score
* $p < .05$, *** $p < .001$

[a]Initial ability is the third grade DSS on the respective measures (reading and mathematics)

The covariance model for the mathematics total fourth grade DSS showed a somewhat different pattern. Group membership was not a significant predictor of the fourth grade mathematics DSS after controlling for initial ability and level form taken. Students who had higher mathematics DSS on the 1997 ITBS ($p < .001$) and took an on-grade level form in 1998 ($p < .05$) had higher fourth grade scores than other students. However, the variance explained by this model was unusually low for a covariance model that uses the identical assessment to control for initial ability ($R^2 = .32$).

Discussion

This paper examined whether students enrolled in classrooms that use a curriculum-embedded performance assessment will show greater gains on a conventional test of academic accountability than students without exposure to such a performance assessment. Our results indicate that students who were in WSS classrooms display growth in reading from one year to the next that far outstrips a demographically matched comparison group and that also exceeds the average change shown by all other students in the district. Further, by examining the results of above and below average students separately, we were able to demonstrate that the impact of the curriculum-embedded performance assessment is not limited solely to those who start with either low or high skills. Rather, the impact appears to be across the board with high and low performing students making comparable gains in reading. This analysis effectively dismisses the objection that these results are attributable to regression to the mean, since students with higher scores in *both* groups performed better than low scoring students in either group. The three-step regressions further demonstrate that participation in WSS classrooms accounts for these differences even after taking into account initial achievement ranking and level of test form administered.

The pattern of change is similar though not as strong in mathematics, although it does not appear to benefit high and low performing students equally. Students who were in the WSS group had higher mathematics scores after controlling for initial ability and level of form administered, but these findings were not statistically significant with this size sample (the effect size with the other PPS group was .18). In addition, the segmentation analysis indicated that the mean change score was lower for high performing students in all groups than for below-average students.

Examination of the construction of the DSS on this edition of the ITBS provides a clue to potential reasons for this finding. In this version of the ITBS, the mathematics DSS appears to be heavily weighted by number concepts and operations and, in particular, by computation (procedural knowledge of operations). In contrast, national standards in effect at the time (NCTM, 1989) and WSS (which is based on these and other standards) address multiple strands of mathematical thinking, including geometry, measurement and spatial sense, data analysis, statistics and probability, patterns, functions, and algebra, as well as number concepts and operations. The findings in mathematics may be a reflection of a mismatch between the broad standards WSS reflects and the test specifications of this edition of the ITBS. In addition, the covariance model for mathematics suggests that there is a problem with this measure of mathematics skills. Only 32% of the variance in the fourth grade scores could be explained by the third grade scores.

The findings on the reading assessment are robust and pervasive and make an important contribution to discussions of accountability. Whether looking at high or low performing students, examining gain scores or using a covariance model, students enrolled in WSS classrooms made greater gains in reading than students who did not have this exposure. These findings, as well as those for math, though not definitive because of our inability to disentangle the impact of WSS from the other innovations co-occurring in the district, suggest a new way to approach accountability testing. For too long, accountability examinations have been assumed to be of a particular kind with an unambiguous focus: normative assessments intended to rank students numerically and compare them to the performance of a specified group. As noted at the outset of this paper, it is likely that the U.S. is spending more money on tests at this time than at any previous point in its history. However, children are not faring well on these assessments. Media reports of large numbers of failures in numerous states are interpreted either in terms of students' lack of skills or teachers' inability to align curricula to the standards that are used to design high stakes tests (Manzo, 2001).

However, another explanation is also possible, and this perspective provides a link to our second research question: Is it possible to design an accountability system that relies on both classroom- and test-based information about student achievement? The alternative view of the accountability debate is that we not only need high standards and tests that reflect these standards, we also need curricula that will enable students to be successful on these assessments—but that are not simply instances of measurement-driven instruction.. By implementing an instructional assessment such as Work Sampling, teachers obtain information about their students on a continuous basis across multiple curriculum domains and from several assessment sources. They compare student performance with standards-based guidelines. They collect multiple sources of information from checklists, portfolios, and student and parent reports. They engage in curriculum analysis in order to evaluate artifacts included in portfolios. And they participate in processes of planning, review, and analysis with their colleagues. Students also have a meaningful role in the

assessment process and thus become active participants in the evaluation process by becoming more familiar with the standards and how to progress towards those standards. This appears to enhance teaching and improve learning.

Perhaps the most important lesson that can be garnered from this study is that accountability should not be viewed as a test, but as a *system*. When well-constructed, normative assessments of accountability are linked to well-designed, curriculum-embedded instructional assessments, children perform better on accountability exams, but they do this not because instruction has been narrowed to the specific content of the test. They do better on the high stakes tests because instruction can be targeted to the skills and needs of the learner using standards-based information the teacher gains from ongoing assessment and shares with the learner. “Will this be on the test?” ceases to be the question that drives learning. Instead, “What should I learn next?” becomes the focus.

When accountability is seen as a system that incorporates both instructional assessment and on-demand tests, both teaching and learning can be affected positively. Moreover, this methodology provides policy makers with clear documentation not only of summative accomplishments, but also of the process of teaching and learning. The approach described in this study places emphasis where it belongs: on teaching and learning, rather than on testing. And it does so without sacrificing either the student or the teacher on the altar of accountability.

Acknowledgement

We wish to thank Jack Garrow for assisting us in obtaining and interpreting school district data. We are also deeply grateful to the principals, teachers, parents, and children who participated in this study, and to the staff and administrators of the Pittsburgh Public Schools. This study was supported by a grant from the School Restructuring Evaluation Project, University of Pittsburgh, the Heinz Endowments, and the Grable and Mellon Foundations. The views expressed in this paper are those of the authors and do not necessarily represent the positions of these organizations.

References

Achieve, Inc. (2002). *Achieve + McRel Standards Database*. http://frodo.mindseye.com/achieve/achieved.nsf

Amrein, A. L. & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, *10*(18). Retrieved 4/1/02 from http://epaa.asu.edu/epaa/v10n18/.

Baron, J. B. & Wolf, D. P. (Eds.), *Performance-based student assessment: Challenges and possibilities* (95th Yearbook of the Society for the Study of Education, pp. 166-191). Chicago, IL: Society for the Study of Education.

Borko, H., Flory, M., & Cumbo, K. (October, 1993). *Teachers' ideas and practices about assessment and instruction. A case study of the effects of alternative assessment in instruction, student learning, and accountability practice*. CSE Technical Report 366. Los Angeles: Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York: Wiley

Corbett, H. D. & Wilson, B. L. (1991). *Testing, reform, and rebellion*. Norwood, NJ: Ablex.

Darling-Hammond, L. (1992). *Standards of practice for learner-centered schools*. New York: Columbia University, Teachers College, National Center for Restructuring Education, Schools, and Teaching.

Falk, B., & Darling-Hammond, L. (March, 1993). *The primary language record at P.S. 261: How assessment transforms teaching and learning*. New York: National Center for Restructuring Education, Schools, and Teaching.

Hoover, H. D., Hieronymus, A. N., Frisbie, D. A., Dunbar, S. B., Oberley, K. R., Cantor, N. K., Bray, J. C., Lewis, J. C., & Qualls-Payne, A. L. (1993), *Iowa Tests of Basic Skills Norms and scores conversions Form K Survey Battery*. Chicago: Riverside.

Jennings, J. (1998). *Why national standards and tests? Politics and the quest for better schools.* Thousand Oaks, CA: Sage Publications, Inc.

Kohn, A. (2000). *The case against standardized testing: Raising the scores, ruining the schools*. Portsmouth, NH: Heinemann.

McNeil, L. M. (2000). *Contradictions of school reform: Educational costs of standardized testing*. New York: Routledge.

Manzo, K. K. (2001, June 20). More than half of California 9th graders flunk exit exam. *Education Week*, p. 19.

Meisels, S. (1996). Performance in context: Assessing children's achievement at the outset of school. In A. Sameroff and M. Haith (Eds.), *The five to seven year shift: The age of reason and responsibility* (pp. 407-431). Chicago: The University of Chicago Press.

Meisels, S. (1997). Using Work Sampling in authentic assessments. *Educational Leadership, 54* (4), 60-65.

Meisels, S., Dorfman, A., & Steele, D. (1994). Equity and excellence in group-administered and performance-based assessments. In M. Nettles and A. Nettles (Eds.), *Equity in educational assessment and testing* (pp. 195-211). Boston: Kluwer.

Meisels, S.J., Jablon, J., Marsden, D.B., Dichtelmiller, M.L., & Dorfman, A. (1994). *The Work Sampling System* (3d ed.). Ann Arbor, MI: Rebus Inc.

Meisels, S.J., Jablon, J., Marsden, D.B., Dichtelmiller, M.L., & Dorfman, A. (2001). *The Work Sampling System* (4th ed.). Ann Arbor, MI: Rebus Inc.

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., Liaw, F-r, Dorfman, A., Nelson, R. F. (1995). The Work Sampling System: Reliability and validity of a performance assessment for young children. *Early*

Childhood Research Quarterly, 10, 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.

National Council of Teachers of Mathematics (1989). *Curriculum and evaluation standards for school mathematics, grades K-6*. Reston, VA: Author.

Nicholson, J. M. (2000). *Examining aspects of consequential validity in a curriculum-embedded performance assessment*. Doctoral dissertation, Ann Arbor, MI: University of Michigan.

National Commission on Excellence in Education. (1983). *A nation at risk: The imperative for educational reform: A report to the nation and Secretary of Education, United States Department of Education.* Washington, D.C.: The Commission [Supt. Of Docs., U.S. G.P.O. distributor].

Orfield, G. & Kornhaber, M. L. (Eds.) (2001). *Raising standards or raising barriers? Inequality and high-stakes testing in public education*. New York: The Century Foundation Press.

Popham, W. J. (2000). *Testing! Testing! What every parent should know about school tests.* Boston: Allyn & Bacon.

Olson, L. (2001, January 11). Finding the right mix. *Education Week, Quality Counts Special Issue*, p. 12 - 20.

University of Iowa and Riverside Publishing Co. (1994). *Riverside 2000 Integrated Assessment Program: Technical Summary I*. Chicago: The Riverside Publishing Company.

About the Authors

Samuel J. Meisels
Erikson Institute
420 N. Wabash
Chicago, IL 60614
Email: smeisels@erikson.edu

Samuel J. Meisels is president of the Erikson Insititute. An emeritus professor of education at the University of Michigan, Dr. Meisels's research concerns the development of alternative assessments from birth through grade 5, policy studies of early intervention, and research regarding the developmental consequences of high risk birth. He has co-authored *The Handbook of Early Intervention*, the “Work Sampling System,” and the “Early Screening Inventory-Revised,” among many other publications.

Sally Atkins-Burnett is an assistant professor at the University of Toledo. She earned her Ph.D. from the University of Michigan. She has been a major contributor to the NCES-sponsored Early Childhood Longitudinal Study and has also worked on the development of instruments used in the Study of Instructional Improvement, a national longitudinal study of the impact of school improvement efforts on instruction and student

performance. Her research interests include assessment in early childhood, social development, literacy instruction, and children with special needs.

Yange Xue is a Research Scientist at the National Center for Children and Families, Teachers College, Columbia University, New York, NY. She holds a doctorate in early childhood education from the University of Michigan. Her areas of specialization include early development and education and quantitative methodology.

Julie Nicholson received her PhD in Early Childhood Education from the University of Michigan. Her research interests include emergent literacy, the use of technology in early childhood classrooms, and the positive and negative consequences associated with different methods of early childhood assessment.

Donna DiPrima Bickel, is a Resident Fellow at the University of Pittsburgh's, Institute For Learning (IFL). Since 1999, Dr. Bickel has coordinated the IFL program in Content-Focused Coaching in elementary literacy. She has co-developed the professional development video and print materials used in preparing school district coaches to work with teachers using this model of professional development.

Seung-Hee Son is a doctoral student in Early Childhood Education at the University of Michigan. Her research interests include parenting and preschool influences on early language and literacy development, and policy issues in child development.

The World Wide Web address for the *Education Policy Analysis Archives* is **epaa.asu.edu**

Editor: Gene V Glass, Arizona State University

Production Assistant: Chris Murrell, Arizona State University

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Editorial Board