MINIMIZING ASSESSMENT BURDEN ON PRESCHOOL CHILDREN: BALANCING BURDEN AND RELIABILITY

Susan Sprachman, Sally Atkins-Burnett, Steven Glazerman, Sarah Avellar, Miriam Loewenberg
Mathematica Policy Research, Inc., PO Box 2393, Princeton, NJ 08543-2393

Key Words: methodology, assessment, early childhood

A. Introduction

Using educational and clinical assessments to measure the developmental functioning of infants, toddlers, and young children has become more common in large surveys in recent years. About 25 years ago, the National Longitudinal Survey of Youth (NLSY79) broke new ground by adding developmental testing of the children of female respondents to its data collection. Since then, in measuring program impacts, many studies of families in poverty, women on or leaving welfare, and pregnant teenagers have included direct assessments of children. With the current policy emphasis on early childhood education (Barnett et al. 2007), an increasing number of studies are describing or evaluating early childhood environments, curricula, or programs. Child outcomes are an important part of these evaluations and descriptive studies. Most of these studies include direct testing of children, using a variety of commercially developed assessments.

Administering developmental assessments to children, especially very young ones, creates unique challenges for large surveys. One such challenge is the length of the assessment batteries. Researchers are interested in examining many domains of a child's functioning, with each domain requiring a different test. Because young children can tire when they have to attend for a long time to the unfamiliar tasks in standardized assessments, their attention to the task may be compromised if the procedures are too long. Yet an assessment battery of half an hour is not unusual. A recent federally funded large scale study, the Preschool Curriculum Evaluation Research (PCER) project, had a battery that averaged close to an hour, so that the battery could capture information on all domains important for the research.[1]

Researchers need to balance the desire to assess many domains of child development against the potential threats to measurement posed by long administrations. Minimizing child burden while maintaining high reliability of estimates of achievement is an ongoing objective. As survey researchers, we have an ethical responsibility to use tools that do not compromise scientific integrity or erode our confidence in measures, while still addressing the needs of young children. We also have an obligation to our clients, who want us to collect the maximum amount of high-quality data possible from those children at a reasonable cost. How can the needs of the children and researchers be balanced?

This paper uses two child assessment measures from the PCER 2003 study as a test case.[2] The PCER battery included three test from the Woodcock-Johnson III (WJ-III) Tests of Achievement, a battery of tests used on many survey research projects of young children. The WJ III tests use stop rules that often take children into questions that are well beyond their ability, which can result in frustration for both the child and the assessor.

Although these rules add just a few minutes to the assessment on any one test, the extra minutes have a cumulative effect.[3] For PCER, the published basal and ceiling rules were used when administering the assessments. To examine what the differences would be if different stop were used, we simulated the effects of different stop rules on the following two tests from the WJ-III:

[1] This research was funded by the U.S. Department of Education's Institute of Education Sciences (IES).through contract award number ED-01-CO-0039 0005 to Mathematica Policy Research, Inc. The authors would like to thank John M. Love, the MPR Project Director and Louisa Tarullo, one of the Principal Investigators for their support.

Corresponding author: Susan Sprachman; ssprachman@mathematica-mpr.com

[2] PCER included two cohorts of researchers. PCER 2002 was awarded to RTI International and PCER 2003 was awarded to Mathematica Policy Research. This paper is based on an analysis of the PCER 2003 data.

[3] The publisher's stop rule can also impact the quality of the data when the tests are administered using paper and pencil versions as it is difficult for field staff to keep track of both the number of errors and the rule to finish the page, resulting in test administrations that do not meet the ceiling rule.

1. ***Letter-Word Identification,*** which examines children's letter knowledge and word recognition. The test begins with naming letters, progresses to two- and three-letter words, and then moves to less familiar words of increasing difficulty.

2. ***Applied Problems,*** which examines the ability to reason mathematically. At the preschool level, the items start by testing a child's beginning number concepts, move to problem solving that requires addition and subtraction, then present word problems of increasing difficulty.

To determine whether one could reduce the burden on children without reducing reliability, we compare the standard scores using the publisher's stop rules with those that would have resulted from administering fewer items. We examine the measurement error, the stability of the relationship between these scores and other measures, and the effect of different stop rules in our analyses.

B. Adaptive Testing

A method for reducing the burden of a lengthy test administration is to use adaptive tests, which tailor the difficulty of items to the ability of the test-taker. Two such adaptive tests that are commonly used on survey projects are the Peabody Picture Vocabulary Test (PPVT III or 4) and the Woodcock Johnson Tests of Achievement III. The PPVT contains over 200 items and the WJ-III individual tests that make up the battery can include as many as 50 items of varying difficulty per test. In an adaptive test the items are grouped in the order of difficulty and the examiner starts the test at a specific point based on a child's age or grade in school, according to an initial basal rule set by the test publisher. To meet the starting or basal rule, the child must be successful on a designated number of items. If the child does not answer the initial items correctly, the examiner administers earlier items, moving backward through the test until the basal rule is met before moving forward. In scoring, when the basal rule is met, it is then assumed that the child would have gotten all earlier items correct. Similarly, the examiner would not administer items beyond the point where a child is consistently missing items. Ceiling rules instruct the examiner when to stop. On adaptive assessments, the basal and ceiling rules can be from three to eight consecutive items, depending on how gradual the increase in difficulty is between items and the probability cut point used by the developer. These rules for starting and stopping were typically developed for individual diagnosis and may require the administration of more items than is necessary for research purposes.

C. Ceiling Rules Were Developed for Precision of Individual Data

The basal and ceiling rules in many commercially available tests used in research projects are developed for a specific purpose. As noted above, these rules are developed using precise statistical techniques so that they are reliable at the child level. This is important when, based on the test, a child might be referred for special services. The Woodcock-Johnson tests were designed for situations where the reliability of the individual child's score is paramount—determining eligibility for services. Therefore, conservative stop rules were developed. However, on survey research projects, where data will be aggregated and the group means and standard deviations are more important, this level of precision can increase the cost of the data collection and create additional burden, perhaps without adding much to the precision of the data.

D. When the Formal Test Rules Do Not Fit a Research Project—What Can You Do?

Reducing the length of a test can be done in several ways. Researchers can use Item Response Theory (IRT) to create a shorter standard set of items targeted to the age range of the children in a study, thus eliminating the traditional procedures of establishing an individual basal and ceiling as part of the administration. Using this method, all children receive the same set of items. A variation on this model includes an additional basal and an additional ceiling set for children who score below or above a certain score. Some studies have used home-grown adapted versions of the Peabody Picture Vocabulary Test[4] to shorten its administration time. However, creating an abbreviated test using IRT analysis requires access to a considerable amount of field testing and analysis of item-level data with a representative sample so that an appropriate mix of items can be selected. This is an expensive process and typically would not be budgeted into studies. Nor would most research projects have access to enough item-level data to allow this type of analysis. An alternative technique for shortening an adaptive test is adjusting stop rules so that a child exits an assessment after making fewer consecutive errors.

[4] The PPVT was adapted using IRT analysis for the National Reporting System. See, for example, Meisels, 2004.

This type of change requires careful analysis of the impact of these changes on the subgroups you are studying.

Using data from the PCER 2003 cohort, we explored whether a shortened adaptive test could provide results with acceptable validity. We simulated the effects of further shortening assessments by adjusting the stop rules on two scales from the WJ-III. To determine whether one could reduce the burden on children without reducing reliability of the group estimates, we compare the standard scores using the publisher's stop rules with those that would have resulted from administering fewer items. We (1) examine the difference in means, range, and variance; (2) look at the stability of the relationship between these scores and other similar measures (that used the standard scoring and administration); and (3) estimate covariate-adjusted models with the standard scores from the published stop rules and with the standard scores generated from a stop rule of 3. For school-age children, a screening version similar to the WJ tests called the Mini-Battery of Achievement uses a stop rule of 3, so we selected a stop rule of 3 as our starting point in comparing results with the standard scoring.

E. Analysis

The sample of children ranged in age from 3 to 6 years old in the PCER study. Approximately half the sample was African-American, about one-third was white, and the rest included children from diverse ethnic backgrounds (see Table 1). Most of the children were low income (although not as disadvantaged as a sample of children from Head Start would be), and all attended an early childhood program. Assessments were administered in the preschool year fall 2003 and spring 2004.[5]

TABLE 1: RACE AND GENDER OF CHILDREN IN THE STUDY

Total N Fall	**1,224**[a]
Male	621
Female	582
White	422
African-American	600
Hispanic	57
Asian or Pacific Islander	**
Native American	**
Multiple Other	63

[a] Not all children had complete test data.
** Values suppressed to protect student confidentiality

To examine the impact of different stop rules on the reliability of the children's scores, we calculated standard scores for the data using a 3, 4, 5, and 6 consecutive error stop rule. The application of different stop rules did not affect the range of scores, except in the spring of the year for the Letter-Word Identification test (see Table 2). The highest score using the standard stop rule (6 errors and finish an easel page) was 168, while the rule of 3 to 6 truncated the range somewhat, to 51 to 161. The mean was greater with the standard scoring, but greater variance was evident with the stop rule of 3 or 4. (see Table 3).

TABLE 2: RANGE WITH STANDARD AND ADJUSTED SCORING

	Standard Scoring		Rules of 3,4,5 or 6	
	Min	Max	Min	Max
Letter-Word Identification, Fall (N = 1,149)	64	160	64	160
Letter-Word, Spring (N = 1,128)	51	168	51	161
Applied Problems, Fall (N = 1,134)	46	132	42	132
Applied Problems Spring (N = 1,125)	29	135	29	135

[5] This analysis covers the preschool year only. Children were also assessed at the end of kindergarten. The PCER study included two cohorts. This analysis includes data from the PCER 2003 cohort only, collected by MPR. RTI collected data from a similar cohort.

TABLE 3: WOODCOCK-JOHNSON III LETTER-WORD IDENTIFICATION SCORE USING DIFFERENT STOP RULES

Time Point	Stop Rule	Mean	Standard Deviation
Fall	3	98.01	16.20
	4	98.91	16.15
	5	99.57	15.81
	6	99.86	15.67
	Standard	99.94	15.64
Spring	3	101.61	14.45
	4	102.24	14.22
	5	102.61	13.89
	6	102.77	13.78
	Standard	102.85	13.94

The first time the children took the test was probably the first experience that many of these children had had with a testing situation. In the fall, the variance on the Letter-Word Identification score increased with the adjusted scores. In the spring, the mean with the standard scoring was again greater than the mean with the alternative scoring rules, but the variance was less with stop rules of 5 or 6, and greater with stop rules of 3 or 4.

When we compare the least extreme change, ending Letter-Word after six consecutive errors but without requiring the tester to finish an easel page, for 99 percent of the sample the standard scores were unchanged for the fall and spring with both methods. Similarly, the rule of five errors had a high match rate of 95 percent in both the fall and spring. Progressing to four errors yielded an 86 percent match with the publisher's ceiling in the fall, but this increased to an 89 percent match in the spring. Looking at a stop rule of three in a row, for more than 74 percent of the sample in the fall and 77 percent in the spring this rule did not change the scores. For 76 percent in the fall and 82 percent in the spring, the rule of three changed the standard score by two or fewer points. And for 85 percent in the fall and 91 percent in the spring the rule changed the standard scores by 5 or fewer points

On the Applied Problems subtest, the picture is different. Comparing six consecutive errors with the publisher's stop rule yielded a 95 percent match in the fall and a 91 percent match in the spring. A stop rule of five had an 89 percent match in the fall and an 84 percent match in the spring, and a stop rule of four had a 78 percent match in the fall and a 71 percent match in the spring. The picture is even more dramatic with a stop rule of three in a row incorrect. There was only a 64 percent match with the standard rule in the fall and 56 percent match in the spring. For 64 percent of the sample in the fall and 59 percent in the spring the score changed by 2 or fewer points. And for 81 percent in the fall and 79 percent in the spring the change was 5 or fewer points The mean scores were greater for the standard scoring in both fall and spring (see Table 4). The variance was less for the alternative scoring in the fall than for the standard scoring. In the spring, the variance was greater for the alternative scoring than for the standard scoring.

TABLE 4: WOODCOCK-JOHNSON III APPLIED PROBLEMS SCORE USING DIFFERENT STOP RULES

Time Point	Stop Rule	Mean	Standard Deviation
Fall	3	91.09	14.53
	4	92.14	14.53
	5	92.87	14.66
	6	93.19	14.72
	Standard	93.44	14.77
Spring	3	98.01	16.20
	4	98.91	16.14
	5	99.58	15.81
	6	99.86	15.67
	Standard	99.94	15.64

Without controlling for any other factors, the relationship between the fall and spring scores for the standard scoring was .68 for Letter-Word and .67 for Applied Problems (see Table 5). The stability of the score from fall to spring was similar for the adjusted scoring (r = .67 and .64 for Letter-Word Identification and Applied Problems, respectively).

The divergent validity appears stronger for the revised scoring. In the fall, the bivariate correlation between the Letter-Word Identification and Applied Problems was .47 for the standard scoring and .40 for the revised scoring. In the spring, the bivariate correlation between the Letter-Word Identification and Applied Problems was .50 for the standard scoring and .47 for the revised scoring. The math and literacy measures appear to be different with the adjusted scoring, and one might wonder if the stronger relationship with the standard scoring isn't due to a personality characteristic of children that is related to their inclination to take risks in answering. Alternatively, the increased potential for error in the

TABLE 5: BIVARIATE CORRELATIONS OF STANDARD AND REVISED SCORING OF WOODCOCK-JOHNSON III

Revised Scoring = Stop Rule 3	1	2	3	4	5	6	7	8
1. Fall Letter-Word Identification Standard Scoring	1.00							
2. Fall Letter-Word Identification Revised Scoring	.96	1.00						
3. Fall Applied Problems Standard Scoring	.47	.46	1.00					
4. Fall Applied Problems Revised Scoring	.44	.40	.96	1.00				
5. Spring Letter-Word Identification Standard Scoring	.68	.66	.43	.40	1.00			
6. Spring Letter-Word Identification Revised Scoring	.68	.67	.42	.39	.98	1.00		
7. Spring Applied Problems Standard Scoring	.44	.42	.66	.65	.50	.49	1.00	
8. Spring Applied Problems Revised Scoring	.42	.43	.65	.64	.48	.47	.96	1.00

scoring is different between the literacy and math assessments.

To examine the construct validity of the abbreviated scoring, we compared the bivariate correlations between WJ-III scores and scores on other measures of literacy and mathematics that the children took using both the standard scores and the adjusted scores on the WJ-III Letter-Word Identification scores and Applied Problems scores (see Table 6). The associations were similar for both scoring rules.

TABLE 6: BIVARIATE CORRELATIONS WITH MEASURES IN SAME DOMAINS

	Standard Scoring	Stop Rule 3
TERA with WJ Letter Word – Fall	.60	.61
TERA with WJ Letter Word – Spring	.70	.69
PPVT with WJ Letter Word – Fall	.33	.32
PPVT with WJ Letter Word – Spring	.32	.34
TOLD with Letter-Word Identification – Fall	.27	.27
TOLD with Letter-Word Identification – Spring	.33	.33
CMAA with WJ Applied Problems – Fall	.53	.52
CMAA with WJ Applied Problems – Spring	.53	.53

CMAA= Child Math Assessment-Abbreviated; PPVT = Peabody Picture Vocabulary Test; TERA = Test of Early Reading; TOLD = Test of Language Development; WJ = Woodcock-Johnson.

We examined covariate-adjusted models controlling for fall scores and demographics such as gender and race with the spring Letter-Word Identification score as the dependent variable using a model with only standard scores and a model with only adjusted scores (see Table 7). We then performed a similar analysis with the Applied Problems test (see Table 8). The results were similar, though the stop rule of 3 appears to result in less precise estimates for the Applied Problems. The model with the stop rule of 3 explains slightly less of the variance, particularly in the Applied Problems test

TABLE 7: COVARIATE-ADJUSTED MODEL: WOODCOCK-JOHNSON III SPRING LETTER-WORD IDENTIFICATION

	Letter-Word Identification Standard Score	Letter-Word Identification Standard Score Rev.-Stop 3 Rule
Male	-.054*	-.072*
African-American	.028	.031
Other Minority	.045	.041
Fall Score	.664**	.654**
R-Square	.46	.45

* p<.05; ** p<.001; all coefficients are Betas.

TABLE 8: COVARIATE-ADJUSTED MODEL: WOODCOCK-JOHNSON III SPRING APPLIED PROBLEMS

	Applied Problems Standard Score	Applied Problems Standard Score Rev.-Stop 3 Rule
Male	-.044	-.041
African-American	-.093**	-.100**
Other Minority	-.014	-.022
Fall Score	.638**	.611**
R-Square	.45	.42

* p<.05; ** p<.001; all coefficients are Betas.

We examined more carefully what the changes in these coefficients meant. On the Letter-Word test, the mean standard score difference for females was the same with both scoring rules. With standard scoring, the mean fall score for females was 102.25 and the spring was 104.70 (a difference in means of 2.45). With the stop rule of 3, the mean fall score for females was 101.11 and the spring was 103.56, a difference in means of 2.46). For males, on the other hand, the fall to spring mean score indicated a greater difference when using the stop rule of 3. With standard scoring, the mean fall score for males was 97.73 and the spring was 101.15, a difference in means of 3.42. With the stop rule of 3, the mean fall score for males was 96.01 and the spring was 99.80, a difference in means of 3.79.

On the Applied Problems test, the magnitude of the difference between fall and spring scores was different for African-American and white children. With the standard scoring rule, the difference between fall and spring scores for African-American children was 1.12 (fall=91.68; spring=90.56), while with a stop rule of 3, the difference was 0.85 (fall=88.96; spring=88.11; change=0.85). For white children, using the standard scoring rule, the difference between fall and spring scores was 0.06 (fall=98.78; spring=98.84). With a stop rule of 3, the difference between fall and spring mean scores was .29 (fall=96.15; spring=96.44).

F. Discussion

No clear conclusions can be drawn from these data at this time; further analysis is needed to come to any clear recommendations about the advisability of using a different stop rule. This study did not examine the effect of the different stop rules on the W score provided for Woodcock-Johnson measures. The W score is an interval level score that is used to examine change longitudinally. It is possible that altering the stop rule could have a stronger effect on the W score.

There are several possible explanations for why a child might miss three items in a row and then get a later item correct. One possibility is that the child can guess the correct responses to some items. This may be particularly true for the Applied Problems test, where the correct answer for many of the items is 2 or 5. The increased relationship between divergent constructs when using the standard scoring could reflect a propensity for some children toward taking risks (and guessing successfully at least some of the time) on both of these subtests, while other children are answering only when they are sure of the response.

An alternative explanation for the difference in standard scores under the different rules is that the items are not ordered well in terms of difficulty or that there are many items of similar difficulty together and a child may not know the first three but may know the correct response to one or more of the later items. This would be more likely on the Letter-Word Identification test. Because children often learn the letters in their own name before learning other letters, the ordering by difficulty may differ according to the child's names or how the letters of the alphabet were introduced to the child. More items would need to be administered to accurately estimate the child's ability. If using abbreviated scoring (stop rules less than the standard scoring) created a serious problem with the estimation of this score, one would expect that the construct validity (bivariate correlations with different measures of the same construct) or the fall to spring stability of the score would be more strongly affected by the use of the adjusted scoring rules. The differences in these relationships were small. There was, however, a greater increase for males in the standard scores from fall to spring in letter-word when using the stop rule of 3.

In terms of administration time, the use of the standard stop rule does not appreciably lengthen the Letter-Word Identification test. The child is given only three seconds per letter or word to respond on the Letter-Word Identification subtest. Although applying the standard rule does not significantly affect the timing of the test, it may affect the child's motivation to continue after experiencing so many failures. There may be different for girls and boys. Alternatively, the ordering of letters may be more associated with the frequency with which letters are found in girls' names.

On the Applied Problems test, the administration time could potentially lengthen the test by several minutes. Each problem must be read aloud to the child. A child who guesses correctly on an item could have 5 to 10 more items administered to follow the standard stop rule of6 in a row and complete the page stop rule. With large sample sizes such as those used in this study, the loss of precision in the estimates may not justify the costs involved in following the standard stop rule. However, more investigation of the differences found for African-American children is warranted to be sure that the score differences are due to guessing with the standard scoring, rather than differential item functioning for this group. If item difficulties are different for the various ethnic groups and the items are ordered for the majority group (white), then applying the stop rule of 3 could bias inferences based on the assessment results.

References

Barnett, W.S., J.T. Hustedt, L.E. Hawkinson, and K.B. Robin. *The State of Preschool: 2006 State Preschool Yearbook.* New Brunswick, NJ: National Institute for Early Education Research, 2007. [www.nieer.org]. Accessed April 14, 2007.

Mather, N., and R.W. Woodcock. *Woodcock-Johnson III Tests of Achievement: Examiner's Manual.* Itasca, IL: Riverside Publishing Company, 2001.

McGrew, K.S., and R.W. Woodcock. *Woodcock-Johnson III: Technical Manual.* Itasca, IL: Riverside Publishing Company, 2001.

Meisels, S.J., and S. Atkins-Burnett. "The Head Start National Reporting System: A Critique." *Young Children*, vol. 59, no. 1, January 2004, pp. 64-66.

Woodcock, R.W., and M.B. Johnson. *Woodcock-Johnson Psycho-Educational Battery— Revised.* Chicago: Riverside Publishing Company, 1989.

Woodcock, R.W., K.S. McGrew, and N. Mather. *Woodcock-Johnson III Tests of Achievement.* Chicago: Riverside Publishing Company, 2001.