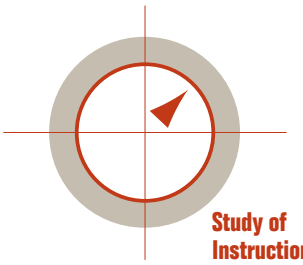

The Study of Instructional Improvement

Methodological Paper

*Administering standardized achievement tests to young children: How mode of administration affects the reliability and validity of standardized measures of student achievement in kindergarten and first grade.**

Sally Atkins-Burnett
Brian Rowan
Richard Correnti

School of Education
University of Michigan

April, 2001

* Paper prepared for the annual meeting of the American Educational Research Association, April 12, 2001.

** The research reported in this paper was partially funded by the U.S. Department of Education to the Consortium for Policy Research in Education (Grant # OERI-R308A60003), Interagency Education Research Initiative, and the Atlantic Philanthropies. Opinions expressed in this paper are those of the authors, and do not necessarily reflect the views of the U.S. Department of Education to the Consortium for Policy Research in Education, the National Science Foundation, and the Atlantic Philanthropies.

Abstract

This paper reports on an experiment conducted to examine the consequences of assessing kindergarten and first-grade students' academic achievement in group versus individualized assessment settings. In the experiment, 442 students blocked by classroom and grade level were randomly assigned to one of two assessment modes—a small group setting with 8 other students from their classroom versus an individualized setting. Students in both settings were administered the grade appropriate form of the CTB McGraw-Hill Terra Nova Tests of Achievement, Form A by trained assessors from the Institute of Social Research at the University of Michigan. Assessment results were then scored by the publisher. The results of the experiment showed that in both kindergarten and first grade, group assessment settings were more likely than individualized settings to be characterized by behavior that assessors coded as disruptive or districting for students, and that students at both grade levels who were assessed in the group setting omitted more test items and made more multiple marks on items than did students assessed in the individual setting. The study also found that kindergarten students assessed in the group setting had lower Reading, Language, and Mathematics scale scores (as estimated by the publishers' three parameter IRT model) and that these scale scores had higher standard errors of measurement for kindergarten students assessed in the group setting. However, there were no differences in measured achievement or standard errors of measurement across assessment modes among first grade students. We argue that the differences in assessment environments and item-response patterns of students in group settings call into question the validity of assessment results for young children assessed in group settings, even when such results do not result in observable differences in the measured outcomes of these children compared to students assessed individually.

Administering standardized achievement tests to young children: How mode of administration affects the reliability and validity of standardized measures of student achievement in kindergarten and first grade.

Increasingly, educational researchers and practitioners have come to recognize the critical importance of learning in the *earliest* grades for students' ultimate success in schooling. As a result, there is now heightened interest in research on the basic literacy and numeracy skills that students enter school with and how these early knowledge and skills are related to subsequent academic growth. Although research on learning in the early grades can be conducted using a variety of research methods, the focus in this paper is on the measurement of student learning in large-scale, quantitative research on schooling, especially large-scale studies of student achievement in the earliest grades. Such research has been greatly facilitated by two, recent, large scale studies of early grades achievement conducted by the U.S. Department of Education (*Prospects* and the *Early Childhood Longitudinal Study*) as well as by several other longitudinal surveys conducted by social and behavioral scientists interested in issues of human development in early childhood (see, for example, the work reported in Entwisle, Alexander, & Olsen, 1997; and Jencks and Phillips, 1998).

A prominent feature of large-scale studies of early childhood development has been the use of standardized achievement tests to assess student learning. The use of such instruments has several advantages for large-scale research, especially research seeking to understand students' academic growth trajectories during the earliest years of schooling. One advantage of standardized achievement instruments (especially commercially-developed ones) is their broad coverage of curricular content; another is the development by testing companies of scale score metrics that allow researchers to appropriately measure and analyze *growth* in student achievement. However, the use of standardized assessments to measure student learning in the earliest grades of schooling is also problematic. It is well-known, for example, that standardized assessments of student achievement have lower reliabilities and larger errors of measurement when conducted with very young children. Moreover, many observers argue that these problems are exacerbated when standardized assessments are administered in a group setting, which young children often find unfamiliar and difficult (Messick, 1983). As a result, many early childhood educators argue that individual (rather than group) administration of achievement assessments are required if researchers are to obtain valid and reliable data on student learning in the earliest grades of schooling (National Association for the Education of Young Children, 1987; Wortham, 1990). [1]

[1] It should be noted that the studies cited in the first paragraph of this paper used various standardized assessment instruments and both group and individual administration procedures to assess student achieve-

ment. As a result, our comments should in no way be considered a blanket condemnation of previous survey research in the area under discussion.

The Problem

In this paper, we investigate the effects of group versus individual modes of standardized achievement assessment with students in the very earliest grades of elementary school. In particular, we report the results of an experiment conducted with kindergarten and first grade students from six urban, elementary schools in the Midwest. In the experiment, students at these two grade levels were randomly assigned to individual or small group (n=8) administration of *Terra Nova*, CTB McGraw-Hill's commercially-available, standardized assessment series designed for use with elementary school students. In the following pages, we report on the extent to which students' behavior during testing sessions differed across group versus individual assessment modes, on the extent to which students responded differently to the *Terra Nova* assessment instrument in these different modes of administration, and on the extent to which the mode of administration therefore affected students' measured achievement and the standard errors of measurement associated with these measures.

The experiment discussed here was conducted as part of the pre-test work for a program of research that we are conducting on instructional improvement in schools pursuing three, very different, comprehensive school reform models. This larger study, which is now being fielded, is using a mixed-method approach to examine *growth* in student achievement as it occurs in the context of three instructional reform models, and it will involve the administration of the *Terra Nova* assessment series to students in grades K-5. Our review the literature on how to collect standardized assessment data with very young children (discussed below) showed that there were widely varying opinions about the use of group versus individual modes of achievement assessment in the earliest grades of schooling, as well as much variation in school system practices across the United States. However, the literature review did not uncover a clear body of empirical evidence that would help us resolve conflicting viewpoints about the appropriate mode of assessment for use with young children.

This lack of empirical evidence was frustrating. With costs as a major consideration, we would have preferred to use group modes of assessment in our study, especially since in peak years of the study we would be assessing roughly 8000 students in 130 schools on a tight timeline. In these circumstances, group assessments produce real efficiencies in testing, allowing testing to be conducted on a more rapid timeline at less cost than is possible with individual assessments. However, as we designed the study, we also were aware of the *strong* stand taken by early childhood researchers and educators (and some psychometricians) against group assessment of young children. Therefore, in the absence of a strong body of empirical evidence to guide our decision making, we decided to conduct an experiment on the effects of group versus individual modes of assessment, not only to guide our own research design process, but also to contribute to the limited empirical literature on the effects of different modes of assessment on the test-taking behavior and measured achievement of very young students.

Background

In conducting the literature review, we carefully searched the ERIC and PSYCHINFO databases in order to find empirical studies examining the effects of different methods of administration on the measurement of students' academic achievement in the earliest elementary grades. This search uncovered only two published research studies directly relevant to our problem (Wodtke, Harper, Schommer, & Brunelli, 1989; and Frisbie & Andrews, 1990). Both studies were well-conducted and addressed issues relevant to our questions, both cited roughly similar references (including two additional unpublished studies germane to these questions), and both decried the limited body of empirical work on the issue at hand. However, neither explicitly compared assessments conducted in group *versus* individual settings, and neither explicitly examined assessment outcomes across the two settings.

One important area that was addressed by both of the published studies was the extent to which teachers departed significantly from the assessment procedures specified in test manuals when conducting group assessments. The study by Wodtke and colleagues (1989) used classroom observations to examine this problem, examining teacher behavior during group administration of standardized achievement tests in ten kindergarten classrooms of varying socioeconomic composition. This study found significant teacher-to-teacher variation in the extent to which test sessions involved departures from specified administration procedures, with rates of significant departure ranging from a high of 44% of items for one teacher to a low of 14% with another teacher, with more teachers tending toward the lower rather than the higher rate of departure. The authors of this study concluded that teachers' departures from specified practice could produce real "tester" effects on student outcomes, although no data were presented to substantiate this inference. By contrast, the study by Frisbie and Andrews (1990), using similar observational procedures, came to the exact opposite conclusion. In this study of 25 kindergarten teachers conducting group administration of Iowa Test of Basic Skills in 38 classrooms across 17 schools, Frisbie and Andrews found very low rates of teacher behavior that they reasoned would significantly change students' achievement scores. Since both studies were conducted with small samples, and since neither study reports the data in ways that allow for a systematic examination of context effects on teacher behavior, we can only conclude that there is probably some variation in how teachers administer group assessments to the kindergarten students they teach, with some fraction of teachers engaging in behaviors that depart from specified procedures and that have the potential to affect student performance on standardized assessments.

The results just discussed are not particularly relevant to the experimental study reported here, since assessments in our experimental study were conducted by well-trained assessors working for the Institute of Social Research at the University of Michigan. Moreover, within the experiment, every effort was made to rotate assessors across assessment modes precisely to mitigate against the

kinds of "tester" effects hypothesized to exist by Wodtke and colleagues (1989). More relevant to the experiment discussed in this paper, however, are the analyses provided by Wodtke and colleagues (1989) and Frisbie and Andrews (1990) of *student* behavior during group assessment situations. Once again using observational methods, Wodtke and colleagues found significant classroom-to-classroom variation in the extent to which students engaged in behaviors that could invalidate their assessment results or the results of other students being assessed in the same group. Among the behaviors that Wodtke and colleagues examined were copying, calling out answers, students helping each other, students being inattentive, and students being disruptive. Across the ten classrooms observed, the percentage of item presentations by teachers which were accompanied by one or more of the child behaviors noted above ranged from 0% in one classroom to 42% in another classroom. The data presented by Frisbie and Andrews were congruent with the study by Wodtke and colleagues. That is, Frisbie and Andrews also found significant variation across classrooms in student behavior, with students in some classrooms showing more of the kinds of behaviors denoted above than students in other classrooms.

Although these well-conducted studies are helpful, especially by pointing to specific kinds of student behavior in group testing situations that might affect student performance, neither study included a comparison group of students being assessed on an individual basis. Moreover, neither study directly associated variations in student and teacher behavior in the assessment situation to student outcomes. As a result, we were forced to look elsewhere to find evidence on the consequences of group *versus* individual administration of standardized instruments on student behavior and measured outcomes. Here too, however, the number of studies involving comparisons of group versus individual administration of a standardized survey instrument to young children was limited. One study examined the scores of students on the State-Trait Anxiety Inventory for Children obtained in group versus individualized settings and indicated that the internal consistency of scale scores was higher in the individualized setting (Papay and Speilberger, 1986). Other studies found that student scores on the McCathy drawing instruments were equally valid across group and individual administrations (Reynolds, 1978). A final study found no differences in the predictive validity of the individually-administered Meeting Street School Screening Test and the group-administered Metropolitan Achievement test with a small sample of beginning first graders (Swanson, Payne, & Jackson, 1981). While each of these studies addressed issues related to the psychometric properties of group- versus individually-administered survey instruments, none was highly relevant to the questions about the assessment of student achievement of relevance to the study reported here.

Hypotheses

Lacking a set of persuasive empirical findings on which to base decisions about the use of group versus individual modes for assessing student achievement in the earliest elementary grades,

we decided to design an experiment that would examine how differing modes for administering standardized assessments might affect student behavior and measured achievement in the very earliest grades of schooling. Using previous research on the process by which respondents answer survey questions generally, we developed a simple conceptual model of how the context of instrument administration might affect the process respondents used to answer survey questions, thereby affecting their measured responses to these survey instrument (for research in this area, see Sudman and Bradburn, 1982; and Schwartz and Sudman, 1996). Building on arguments made by early childhood educators and a good many psychometricians, we developed the following hypotheses about how the use of group versus individual assessments might affect kindergarten and first grade students' responses to standardized achievement assessments and their measured outcomes:

> *H_1*: Context. In comparison to students assessed individually, students assessed in a group setting will: (a) exhibit less attention to the assessment task; (b) be less comfortable in the assessment situation; and (c) experience more disruptive behavior in the assessment situation.
>
> *H_2*: Process. As a result of exhibiting or experiencing these behaviors, students in the group assessment setting will complete the assessment differently than students experiencing individual assessment. In particular, in comparison to students assessed individually, students assessed in a group mode will: (a) be more likely to omit (i.e., fail to respond) to items on the assessment instrument; and (b) be more likely to make multiple marks (rather than a single mark) on items on the assessment.
>
> *H_3:* Response. Because the process of responding to test items differs across assessment modes, students in the group assessment mode will: (a) show lower scores on average than students in the individual assessment mode; and (b) have scores with higher errors of measurement than the scores for students assessed individually.

Sample and Procedures

In order to test these hypotheses, we conducted a study in the kindergarten and first grade classrooms of six urban, elementary schools in Michigan and Indiana.

Sample

In each school, we randomly selected a minimum of 16 students from each eligible classroom from the universe of all children in these classrooms deemed eligible for study. Students deemed *in*eligible for study included those with limited English proficiency and those in special education programs. Among the six schools studied was an early childhood center that had only kindergarten students. In another school, we therefore sampled only first-grade students. Within each participating classroom, the 16 students recruited for the study were randomly assigned to one of two modes of administration, individual administration of a standardized achievement test (n=8 students per classroom) or small group administration of the same test (n=8 students per group, one group per classroom). Each group was assessed using the grade-recommended reading and mathematics subtests from CTB-McGraw Hill's *TerraNova* Tests of Achievement, Form A. The design of the

study equalized the numbers of students per classroom participating in each assessment mode and otherwise relied on random assignment to control for differences across experimental conditions in demographic, cognitive, and motivational variables. Overall, the *TerraNova* reading and mathematics subtests were administered to a total of 442 students (221 group and 221 individual) between May 1 and June 13, 2000.

The random assignment of students to assessment mode within grades resulted in roughly equivalent samples of students experiencing each assessment mode within a given grade. For example, within each grade, students in group versus individual assessment modes did not differ in terms of percentage male and female or in average age (measured in months).[2] Overall, the students in the sample did appear to differ somewhat from the CTB McGraw-Hill standardization sample. Table 1, for example, shows the mean scale scores and standard deviations for these scores for students in the study sample and for students in the comparable CTB McGraw-Hill standardization samples. The scale scores reported here are those provided by the publisher and are based on the three parameter Item Response Model used by the publisher. As Table 1 shows, the mean scale scores in reading, language, and mathematics were lower in the study sample than in the publishers' standardization sample, whereas the standard deviations for these scales scores were larger in the study sample than in the standardization sample.

Table 1: Means and Standard Deviations in Scale Scores for the Study Sample and the CTB/McGraw Hill Standardization Sample

	Study Mean	Study SD	Standardization Sample Mean	Standardization Sample SD
Kindergarten Reading	**492**	**41.30**	**536**	**37.06**
Grade 1 Reading	**535**	**53.89**	**560**	**42.18**
Kindergarten Language	**475**	**51.94**	**524**	**43.94**
Grade 1 Language	**550**	**40.37**	**575**	**42.75**
Kindergarten Mathematics	**435**	**44.91**	**492**	**44.56**
Grade 1 Mathematics	**493**	**38.15**	**511**	**39.23**

Assessment Procedures

All student assessments were conducted by assessors employed by the University of Michigan's Institute for Survey Research, and all assessors received special training by study staff on how to administer the *Terra Nova* and observe the administration process. At each school in the study, assessors worked in teams of four. Two assessors administered individual assessments, a third assessor administered group assessments, and a fourth assessor acted as an observer of the third assessors' group administration. Assessors were rotated between individual and group assessment modes and,

[2] Tables available on request.

within the group mode, between assessor and observer roles. Rotation procedures were implemented in order to minimize "assessor" effects on student behavior and outcomes.

Student behavior during individual assessment sessions was recorded by the assessors, each of whom completed structured observation notes immediately after completing an assessment. For group assessments, the staff member performing the observer role took on-going observation notes using the same coding scheme used by assessors conducting individual assessments. In both assessment modes, these procedures allowed us to gather comparable data on the amount of time required to complete an assessment, the assessment environment, student behavior, and observer ratings of student attentiveness to and comfort with the assessment task. Upon completion of an assessment session, assessors examined each test booklet completed by students and cleaned stray marks before shipping the booklets to CTB McGraw-Hill for scoring. The assessments were scored by CTB McGraw-Hill using a three parameter Item Response Model, and the publisher provided us with a research tape that included overall scale scores for each student, the standard errors of measurement for each student on each scale, and the item-level responses of each student in the study.

Results

In the following pages, we make inferences about how students behaved during assessment sessions using the coded observation notes completed by the assessment teams. We make inferences about how students completed the assessment instrument by examining students' item-level responses to the assessment instrument. In making these inferences, we examined the extent to which students omitted responses and/or made multiple answers, as well as the specific items and locations within the assessment instrument where such responses were observed. We also examined the ratio of correct and incorrect answers to omitted and multiply marked answers. Data on outcomes come from the publisher's estimated scale scores and standard errors of measurement for each scale, as calculated for each child in the study.

Student Behavior in Different Assessment Modes

Table 2 reports data relevant to our first hypothesis—that student behavior with the potential to affect students' item response process will vary in frequency across assessment modes. In particular, we hypothesized that students in group assessments would be less attentive to and be more uncomfortable with the assessment task in group as opposed to individual assessment modes, where data on students' attention and comfort were derived from assessors' ratings. We also hypothesized that there would be more instances of disruptive behavior or behaviors that might unduly influence students' responses to test items in group versus individual modes of assessment. Data on this hypothesis were taken from assessors' field notes, where the behaviors in question included students talking to one another, calling out answers, overtly refusing to complete an assessment, whistling, breaking pencil points, crying, getting out of their seats, noisily playing with file folders or as-

sessment booklet pages, or other behaviors which might distract students from the task at hand or influence students' responses to assessment items.

We begin by reporting the ratings made by assessors of the presence or absence of disruptive and/or distracting behaviors in testing sessions. As Table 2 shows, observers were far more likely to note the presence of disruptive or distracting behaviors in the group assessment mode than in the individual mode at both of the grade levels under study. But one must take some care in interpreting these data. For example, the data presented in Table 2 show rates of disruptive and/or distracting behaviors (as rated by assessor field notes) across different assessment modes. However, the data in the table are not strictly comparable across modes. For individual sessions, the rates shown in Table 2 refer to the single student being tested and indicate whether he or she exhibited one of the disruptive or distracting behaviors listed above; for group sessions, by contrast, the rates shown in Table 2 indicate whether any one of eight students in a group testing session showed at least one of the disruptive or distracting behaviors listed above.

Table 2: Percentage of Testing Sessions Characterized by One or More Instances of Disruptive or Distracting Behaviors

	Individual Mode	Group Mode
Kindergarten	**7.3%** **(n=110 testing sessions)**	**78%** **(n=14 testing sessions)**
First Grade	**14.4%** **(n=111 testing sessions)**	**78%** **(n=14 testing sessions)**

Keeping in mind the lack of strict comparability in data in the cells in Table 2, the data in the table do suggest that rates of disruption and/or distraction per testing session were greater in group versus individual testing sessions, no matter the grade level. For example, in kindergarten, only 8 of 110 (or 7.3% of) students assessed individually exhibited behaviors that were coded as disruptive or distracted by assessors, whereas students in 11 of 14 groups (or 78% of students in the group assessment mode) were exposed to behavior coded as disruptive or distracting by observers. In first grade, only 16 of 111 (or 14.4% of students) assessed individually exhibited disruptive or distracted behavior, whereas students in 11 out of 14 groups (or 78% of all students in the group assessment mode) were exposed to disruptive or distracting behavior during a group testing session.

An alternative way to assess student behavior in test sessions is to examine data derived from assessors' ratings of students' overall attentiveness and comfort with the assessment task. Table 3 presents these data. Here, assessors rated students attention and comfort level on a standard, Likert type scale. The data from these ratings show that kindergarten students in the group assessment mode were generally rated as less attentive to the assessment task than kindergarten students in the individual mode, and that this was true for both reading and mathematics testing sessions. However,

there were no mean differences in observers' ratings across assessment modes for kindergarten students' overall comfort with the assessment task (either in reading or in mathematics testing sessions). Moreover, the data in Table 3 show that there were no statistically significant differences in observers' ratings of student attentiveness or comfort level across assessment modes in the first grade sample.

Table 3: Means and Standard Deviations for Observer Ratings of Student Attention and Comfort Level in Different Assessment Modes (by grade)

	Mean for Group Mode	SD for Group Mode	Mean for Individual Mode	SD for Individual Mode	Mean Difference Across Modes (I-G)	t
Kindergarten (n=111 students)						
Attention to Reading	**3.62**	**1.26**	**4.10**	**1.01**	**.48**	**3.01****
Attention to Mathematics	**3.55**	**1.32**	**3.96**	**1.07**	**.41**	**2.40***
Comfort level –Reading	**4.45**	**.92**	**4.26**	**.88**	**-.19**	**1.48**
Comfort level - Math	**4.26**	**.96**	**4.21**	**.92**	**-.05**	**.38**
First Grade(n=110 students)						
Attention to Reading	**4.16**	**.94**	**3.93**	**1.12**	**-.23**	**1.62**
Attention to Mathematics	**4.03**	**1.10**	**3.98**	**1.12**	**-.05**	**.31**
Comfort level –Reading	**4.07**	**.95**	**4.10**	**.97**	**.03**	**.20**
Comfort level - Math	**4.09**	**1.00**	**4.03**	**1.07**	**-.06**	**.42**

Overall, then, the data provide somewhat ambiguous support for the idea that students in different assessment modes will exhibit different types of behavior or have different levels of comfort in the assessment setting. While the data on testing sessions show that at least one student exhibited disruptive or distracting behavior in 78% of all group testing sessions, only in kindergarten did this appear to lead to lower ratings of attentiveness to the assessment task by students in the group (as opposed to the individual) assessment mode. Put differently, assessor ratings suggested that first grade students were equally attentive to the assessment task no matter what the assessment mode, whereas attentiveness among kindergartners (at least as rated by observers) was lower in group testing sessions than in individual sessions. In addition, controlling for grade level, there were no apparent differences in students' level of comfort with the assessment task (at least as rated by observers) across different assessment modes.

The Response Process

One problem with observer ratings, of course, is that they are high inference measures. In particular, it is especially difficult for observers to unambiguously determine the extent to which students are actively attending to the assessment task or to gain a window on how students are actually responding to that task. As a result, we turn now to an examination of data that, we argue, give us at

least some opportunity to observe directly how students were responding to the assessment task. The data that are discussed in this section concern the extent to which student test booklets had multiple marks and/or omissions, both of which are possible signs of student distraction and/or disruption. The hypothesis here, for example, is that the a greater frequency of disruptions per session in group versus individual modes of assessment noted above, and/or the greater frequency of distractions per session (noted above) would lead students who were otherwise equal to: (a) make more multiple marks; and (b) make more omissions.

Table 4 shows the relevant data for the reading assessments. Here, the data unambiguously support the idea that students responded differently to the assessment task depending on assessment mode. However, the pattern in the data is not as simple as our hypotheses suggest. In particular, the data in Table 4 suggest that students in both kindergarten and grade 1 who were assessed in a group mode made *more* multiple marks and omissions than students in the individual mode, but this pattern also led them to make *fewer* correct or incorrect answers than did students assessed in individual settings. Table 4 suggests that this pattern is stronger in kindergarten than in first grade (as evidenced by the t tests), but a general conclusion nevertheless seems warranted. In the individual assessment mode, students were providing a clearer "signal" about what they knew and didn't know, in large part because students in this setting made fewer multiple responses and omitted fewer items.

Table 4: Number of Items on the Reading Test Responded to in Different Ways by Students in Group Versus Individual Assessment Modes (by Grade)

	Mean for Group Assessment Mode	S.D. for Group Assessment Mode	Mean for Individual Assessment Mode	S.D. for Individual Assessment Mode	Mean Difference (Individual minus Group)	t statistic
Kindergarten						
Correct response	**19.05**	**6.76**	**22.40**	**4.97**	**3.35**	**4.19*** **
Incorrect Response	**12.33**	**4.85**	**16.58**	**4.71**	**4.25**	**6.60*** **
Multiple Marks	**.71**	**.92**	**.01**	**.09**	**-.70**	**7.92*** **
Omissions	**7.92**	**8.58**	**1.01**	**2.30**	**-6.91**	**8.16*** **
First Grade						
Correct response	**25.17**	**6.61**	**24.34**	**7.29**	**-.83**	**.89**
Incorrect Response	**14.62**	**5.37**	**17.23**	**6.42**	**2.61**	**3.29*** **
Multiple Marks	**.23**	**.72**	**.10**	**.69**	**.13**	**1.41**
Omissions	**4.98**	**5.69**	**3.36**	**5.87**	**-1.62**	**2.08***

- **p<.05, ** p<.01, *** p<.005**

A similar pattern emerges in Table 5, which shows data on response patterns for students on the mathematics portion of the Terra Nova assessment series. Here too, students assessed in the group mode made more multiple markings and omitted more items than did students assessed individually, and here too students in the group setting had fewer items that were unambiguously correct

or incorrect. Thus, once again, we see a pattern in which the "signal" about what students know is clearer in the individual mode of mathematics assessment than in the group mode, a pattern that is equally strong in both kindergarten and first grade.

Table 5: Number of Items on the Mathematics Test Responded to in Different Ways by Students in Group Versus Individual Assessment Modes (by Grade)

	Mean for Group Assessment Mode	S.D. for Group Assessment Mode	Mean for Individual Assessment Mode	S.D. for Individual Assessment Mode	Mean Difference (Individual minus Group)	t statistic
Kindergarten						
Correct response	**11.96**	**5.02**	**14.48**	**3.84**	**2.52**	**4.18*** **
Incorrect Response	**11.04**	**4.07**	**15.17**	**3.86**	**4.14**	**7.73*** **
Multiple Marks	**.53**	**.76**	**.02**	**.13**	**-.51**	**6.89*** **
Omissions	**6.47**	**6.98**	**.33**	**.78**	**-6.15**	**9.19*** **
First Grade						
Correct response	**24.73**	**8.45**	**25.86**	**7.65**	**1.13**	**1.05**
Incorrect Response	**16.92**	**6.47**	**19.41**	**7.33**	**2.50**	**2.71****
Multiple Marks	**1.41**	**1.42**	**.28**	**.59**	**-1.13**	**7.74*** **
Omissions	**3.94**	**6.34**	**1.45**	**4.95**	**-2.49**	**3.25*** **

- **p<.05, ** p<.01, *** p<.005**

In summary, the data from students' item responses strongly suggest that the process of answering questions is affected by the mode of assessment used with students. In both kindergarten and first grade, it appears that students assessed in the group mode are more likely than students assessed in the individual mode to omit items and to make multiple marks on an item. As a result, students tested in the group mode are providing assessment data with less "signal" than students tested in the individual setting. Although we cannot directly model why such a process occurs, at least two explanations seem warranted. One explanation would borrow on the observation (shown in Table 2) that levels of disruption and distraction are higher in the group versus the individual assessment mode. Here, one assumes that students who experience disruptive or distracting behavior during the assessment task will be more likely to omit items and/or make multiple responses as a result of this situation. A second explanation might lie in the differing amounts of time it took students to complete the assessment task in group versus individual settings. In our study, the average length of the reading/language arts testing session in kindergarten was 42 minutes for group assessment and 25 minutes for individual assessment; in mathematics, the average length of an assessment session was 28 minutes in the group mode and 19 minutes in the individual mode. A similar pattern prevailed in first grade testing sessions. In fact, the logs kept by assessors strongly suggest that length of the testing session was associated with increased fatigue and disruption, in both kindergarten and first

grade *group* assessment sessions. This waning of attention and increase in disruptions and distractions would explain the higher levels of omissions and multiple marks observed in group assessments.

Measured Outcomes

The final step in the analysis involves an examination of the measured achievement of students across the two assessment settings. Prior to conducting this analysis, we predicted that students assessed in an individual setting would: (a) show higher mean achievement, and (b) lower standard errors of measurement in comparison to students assessed in the group setting. Our hypothesis was based on the assumption that group settings decreased student attention to the assessment task, in part through higher levels of disruption and distraction. To test these ideas, we undertook eight, separate multiple regression analyses. In each of these analyses, we separated students into different grades, and within grades, we separately regressed students' measured achievement (in IRT-scaled scores) or the standard errors of measurement for each scale score on students chronological age (in months), sex (female=1), and assessment mode (individual=1). The results of the analyses are presented below.

We begin with Table 6, which shows the results of these analyses for students' measured achievement as estimated by the publishers' item response models. In this table, effects are reported in the scale score metric, and no variables are standardized. The data only partly confirm our expectations about the effects of assessment mode on measured student achievement. As Table 6 shows, after adjusting for age and sex, kindergarten students assessed individually have higher scale scores on all three scales than do students assessed in a group setting. Moreover, these effects are quite large, as indicated by the standard "effect size" metric (which divides the mean difference in a particular scale score across assessment modes by the standard deviation in the relevant scale score for the study sample as a whole). The standard deviations of each scale score for the study sample as a whole are provided in Table 1 (above). Using these figures, we find that the effect size of assessment mode on the reading scale is .60, on the language scale it is .39, and on the mathematics score it is .51. In contrast, after adjusting for age and sex, there were *no* statistically significant differences in scale scores for first grade students, despite the fact that our experiment provides more than sufficient statistical power to detect effect sizes of .30 or greater. In addition, Table 6 shows that the effects of sex on measured achievement are statistically insignificant in this sample, and that the effects of age, which are quite small, are statistically significant in only isolated instances.

Table 6: The Effects of Assessment Mode, Age, and Sex on Achievement Scale Scores (by Grade)

	Reading Scale Score			**Language Scale Score**			**Math Scale Score**		
	Coefficient	**SE**	**t**	**Coefficient**	**SE**	**t**	**Coefficient**	**SE**	**t**

Kindergarten									
Constant	**396.82**	**41.06**	**9.66**	**320.98**	**52.24**	**6.14**	**334.75**	**45.49**	**7.36**
Female	**1.48**	**5.50**	**.27**	**-3.80**	**7.00**	**.54**	**5.01**	**6.09**	**.82**
Individual†	**24.93*****	**5.39**	**4.62**	**20.17*****	**6.86**	**2.94**	**23.08*****	**5.95**	**3.88**
Age	**1.13***	**.55**	**2.04**	**2.00****	**.70**	**2.84**	**1.19**	**.61**	**1.94**
R^2		**.10**			**.07**			**.08**	
First Grade									
Constant	**654.89**	**57.12**	**11.45**	**591.59**	**42.89**	**13.79**	**484.70**	**41.27**	**11.75**
Female	**4.45**	**7.24**	**.61**	**10.57**	**5.44**	**1.94**	**-.17**	**5.20**	**.03**
Individual	**-6.10**	**7.21**	**.84**	**2.08**	**5.41**	**.38**	**7.95**	**5.18**	**1.54**
Age	**-1.37***	**.65**	**2.11**	**-.55**	**.49**	**1.13**	**.06**	**.47**	**.91**
R^2		**.03**			**.03**			**.01**	

† Coefficients result in ES_{sample} of .60 (Reading), .39 (Language) and .51 (Math) in SD units
*** p<.05, ** p<.01, ***p<.005**

In Table 7, we turn to the effects of assessment mode on the standard errors of measurement as calculated by the publisher. It will be recalled that these are measures of the reliability of a student's scale score, and as such, we hypothesized that students assessed individually would have lower standard errors of measurement than students assessed in the group setting. Once again, we find that assessment mode had different effects depending on grade. In kindergarten, assessment mode had a strong and statistically significant effect on standard errors of measurement across all three scales. The standard deviations of the standard errors of measurement in the study sample as a whole are 16.88 for kindergarten reading, 16.43 for kindergarten language, and 17.17 for kindergarten math scale scores. Thus, using the standard effect size just discussed, the effect size of assessment mode on reading standard errors in kindergarten is .45 for reading, it is .31 for language, and it is .44 for math. Once again, however, there are *no* statistically significant differences across assessment modes for the first grade sample.

Table 7: The Effects of Assessment Mode, Age, and Sex on Standard Errors of Measurement (by Grade)

	Reading SEM			*Language SEM*			*Math SEM*		
	Coefficient	SE	t	Coefficient	SE	t	Coefficient	SE	t
Kindergarten									
Constant	**48.83**	**17.20**	**2.84**	**50.45**	**16.72**	**3.02**	**56.31**	**17.39**	**3.24**
Female	**-.68**	**2.30**	**-.30**	**4.52***	**2.24**	**2.02**	**-3.62**	**2.33**	**1.56**
Individual†	**-7.61*****	**2.25**	**-3.37**	**-5.06****	**2.19**	**2.31**	**-7.62*****	**2.28**	**3.35**
Age	**-.27**	**.23**	**-1.15**	**-.34**	**.23**	**1.51**	**-.37**	**.23**	**1.58**
R^2		**.06**			**.06**			**.07**	
First Grade									
Constant	**6.14**	**17.67**	**.35**	**24.96**	**10.30**	**2.42**	**13.27**	**9.50**	**1.40**
Female	**-1.81**	**2.24**	**-.81**	**-1.74**	**1.31**	**1.34**	**-.04**	**1.20**	**.03**
Individual	**1.87**	**2.23**	**.84**	**-.06**	**1.30**	**.04**	**-2.19**	**1.19**	**1.83**
Age	**.20**	**.20**	**.98**	**-.03**	**.12**	**.28**	**.02**	**.11**	**.15**
R^2		**.01**			**.01**			**.02**	

† Coefficients result in ES_{sample} of .45 (Reading), .31 (Language) and .44 (Math) in SD units

- **p<.05, ** p<.01, ***p<.005**

The results in Tables 6 and 7 are interesting. Earlier (in Tables 4 and 5), we saw that student responses to the assessment instruments differed across assessment modes in both kindergarten and first grade, although the differences were greater in kindergarten than in first grade. In kindergarten especially, and to a lesser extent in first grade, students responded to proportionally fewer items when tested in a group as opposed to individual mode of assessment. In Tables 6 and 7, however, it appears that the scaling models used by CTB McGraw-Hill mitigated against finding significant differences in measured achievement outcomes among first graders assessed in different settings.[3] In contrast, the scaling models *did* produce differences in measured outcomes for kindergarten students tested in different settings, with kindergarten students tested in the group setting having both lower scores and higher standard errors of measurement than comparable students assessed individually.

Discussion

The results of the experiment presented here are less clear than we had hoped for, but then again, the analyses presented here show that the problem under consideration is far more complex than we initially anticipated. Clearly, there was evidence in this study sample that assessment setting had *strong* effects on the assessment process and on measured assessment outcomes in kindergarten, with the kindergarten students assessed in a group setting being exposed to more disruptive and distracting behavior, omitting more answers and making multiple marks on more test items, and ob-

[3] The scaling models used by the publisher provide higher standard errors of measurement when students are answering fewer questions correctly and/or when their pattern of responses indicates that they are answering more difficult questions correctly after missing easier questions. As Table 5 showed, first grade students in the two different assessment modes answered roughly the same number of items correctly. This might contribute to the lack of difference in scale scores and standard errors of measurement found in the first-grade data.

taining achievement scores that were lower and measured with more error than was the case for students assessed individually. However, the results were less dramatic for the first grade students participating in this study. Here, we found that first grade students who were assessed in a group setting were exposed to higher rates of disruptions and distractions per test session than were students assessed individually. Moreover, the first grade students assessed in a group setting also were more likely to make multiple marks in response to items and to omit more items than first grade students tested individually. However, after the assessment results for first grade students were scaled by the publisher, there were no significant differences either in obtained scale scores or standard errors of measurement for students assessed in the different modes.

The results call for more extended analyses of the problem addressed by this study. For one, it is important to note that there were only 8 students per group in the group assessment setting in this study. Lacking variation in group size, and testing only in smaller groups, we are unable to say whether larger-sized groups (as might occur in at least some American classrooms and/or in some research studies) would amplify the effects of group mode of assessment on first grade students' measured outcomes, making them statistically and substantively different from the measured outcomes of first grade students assessed individually. We would argue that experimental tests of this hypothesis are warranted by our results. In addition, we were unable in this study to assess the extent to which the mode of assessment had different effects for different types of students, for example, students from varying home backgrounds and/or prior achievement levels. This too is an issue worth exploring, but it would require a much larger sample than the one obtained here. Finally, there is a need to more carefully observe and explain differences in testing context, process, and outcomes than was done in this study. For example, while our study suggests that group and individual assessment environments differ greatly in terms of disruptions, distractions, and so on, and while the data we presented suggest that the processes students use to respond to test items differs across group and individual assessment settings, we did not attempt to associate measured features of assessment settings directly to measured student outcomes, nor did we closely observe or interview individual students to determine why and how they might have responded to test items differently as a result of different environmental conditions. These issues also are worthy of more study, although such a study would require a more intensive design than the one used here.

In the face of the need for more data, it is worth noting that our upcoming study of school reform models will conduct all assessments in grades K-2 in an *individual* setting. This is because the data on behavior in group testing environments overwhelmingly suggested the wisdom of this procedure for a number of reasons. For one thing, the notes from our assessors indicated that group assessment modes were quite trying, not only for kindergartners, but also for many first graders. Moreover, group assessments proved difficult for assessors as well. The following note by one of

our assessors captures a range of reasons why we decided in favor of individualized assessment with students in the earliest grades. The note also captures the sentiments of most assessors in this study quite well. As the assessor noted:

> A group like this and of this size places a great burden on the assessor. It is emotionally and physically draining. The clean up is horrendous as well, as you cannot control the kids who are writing in books. Some of the kids could have actually flourished in a one-to-one setting, but a group setting was detrimental to them. They left feeling like failures.

The important point is that this observation was not unusual, and it confirmed what others researchers and practitioners have reported about the nature of group assessment in the earliest grades. The group setting in our study was characterized by a variety of student behaviors that were disruptive and distracting for students. Group settings also were characterized by behaviors that had the potential to affect the validity of student responses to assessment items. As just one example, in about 63% of all kindergarten group assessments, and in 36% of all first grade group assessments, students called out answers, allowing other students to gain access to their answers. All of this raises questions about the validity of test score data derived from group assessments in the earliest grades—even if complex (and useful) scaling procedures can produce assessment scores with similar means and standard errors of measurement for students assessed in different settings. Most importantly, however, we remain concerned about the welfare of the children we assess, and believe that our observational notes and data provide firm enough evidence to suggest the wisdom of assessing students in the earliest grades on an individual basis. As we have seen, group-testing sessions were longer than individual sessions, and they were trying for a good many students. For these and other reasons, then, we have decided against the use of group-administration of standardized assessments in our own research.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Testing*. Washington, D.C.: Author.

CTB McGraw-Hill. (1997). *Terra Nova Technical Bulletin 1*. Monterey, CA. Author.

Entwisle, D. R., K. L. Alexander & L.S. Olson. (1997). *Children, Schools, and Inequality*. Boulder, CO: Westview.

Frisbie, D.A. & K Andrews. (1990). Kindergarten pupil and teacher behavior during standardized testing. *The Elementary School Journal*, *90*(4), 435-448.

Jencks, C. & M. Phillips. (1998). *The Black-White Test Score Gap*. Washington, D.C.: Brookings.

National Association for the Education of Young Children. (1987). *Standardized Testing of Young Children 3 though 8 Years of Age: A Position Statement of the National Association for the Education of Young Children*. Washington, D.C.: Author.

Papay, J. & C. Spielberger. (1986). Assessment of anxiety and achievement in kindergarten and first- and second-grade children. *Journal of Abnormal Child Psychology*, *14*(2), 279-286.

Reynolds, C. (1978). The McCarthy drawing tests as a group instrument. *Contemporary Educational Psychology*, *3*(2), 169-174.

Schwartz, N. & Sudman, S. (Eds.). (1996). *Answering Questions: Methodology for Determining Cognitive and Communicative Processes in Survey Research*. San Francisco: Jossey Bass.

Sudman, S. & Baradburn, N.M. (Eds.). (1982). *Asking Questions: A Practical Guide to Questionnaire Design*. San Francisco: Jossey Bass.

Swanson, B., D. Payne, D. & B. Jackson. (1981). A predictive validity study of the Metropolitan Readiness Test and Meeting Street School Screening Test against first-grade Metropolitan Achievement Test scores. *Educational and Psychological Measurement*, *41*(2), 575-578.

Wodtke, K.H., F. Harper, M. Schommer, & P. Brunelli. (1989). How standardized is school testing? An exploratory observational of standardized group testing in kindergarten. *Educational Evaluation and Policy Analysis*, *11*(3),223-235.

Wortham, S.C. (1990). *Tests and Measurement in Early Childhood Education*. New York: Merrill-Macmillan Publishing Co.