north carolina

Report to the Joint Legislative Education Oversight Committee on the

Merits of Testing at the End of Second Grade for ABCs Accountability

October 2000

Public Schools of North Carolina
State Board of Education . Department of Public Instruction
Office of Instructional and Accountability Services

A Report on the Merits of Testing at the End of Second Grade for ABCs Accountability

Executive Summary

Twelve Local School Administrative Units (LSAU) were selected for participation in a legislatively required study to examine the merits of testing at the end of second grade in order to determine pretest scores for measuring growth in the ABCs of Public Education. Second grade students in participating LSAUs took a specially reformatted version of the third grade pretest in the spring of 1999 and subsequently took the third grade pretests and end-of-grade tests during 1999-2000.

There were small differences in the growth models depending on which pretest was used. However, fall-to-spring growth models tended to have the most desirable characteristics. The data from the study, recent statewide data, teacher opinion, cost and logistical considerations all contribute to the conclusion that there is no compelling reason to switch to testing at the end of second grade in order to revise the growth models for third grade reading or mathematics.

The study recommends that the State Board of Education consider updating the third grade growth models for reading and mathematics based on more recent statewide data than was available at the time of first implementation. Any changes must be coordinated with planned curriculum revisions and attendant revisions of the North Carolina testing program.

A Report on the Merits of Testing at the End of Second Grade for ABCs Accountability

Table of Contents

A Report on the Merits of Testing at the End of Second Grade for ABCs Accountability

In the Appropriations Act of 1998 the General Assembly required the State Board of Education (SBE) to study the possibility of using a test at the end of second grade to measure growth for the ABCs. Such a test would serve as a pretest for measuring growth in reading and mathematics during third grade, and could replace the pretest given in the fall of third grade for that purpose. However, existing statute [G.S. 115C-174.11(a)] prohibited the use of standardized tests in first and second grades. Therefore the General Assembly provided a special exception for the study. The requirement for the study was provided in GS 115C–174.11(c)(1) as follows:

> ... the State Board shall develop and implement a study allowing selected local school administrative units that volunteer to administer a standardized test in May, 12 months prior to the third grade end-of-grade test, in order to establish a baseline that will be used to measure academic growth at the end of third grade. Initially, the State Board shall select 12 volunteer local school administrative units that are diverse in geography and size to participate in the study. If the State Board determines that a standardized test administered in May, 12 months prior to the third grade end-of-grade test, is more reliable than a standardized test administered at the beginning of third grade for the purpose of measuring academic growth, the State Board may change the test date for additional local school units. The State Board shall report the results of the study to the Joint Legislative Education Oversight Committee by October 15, 2000.
>
> Baseline measurements administered in May, 12 months prior to the third grade end-of-grade test, are not public records as provided in Chapter 132 of the General Statutes.

Several issues had to be resolved to carry out the requirement. First, what test should be administered at the end of second grade? Second, how would the volunteers be selected? Finally, how would the study operationalize the question posed by the General Assembly (i.e., is the second grade test more reliable for measuring academic growth)?

It was decided that the most feasible solution for a test to be administered to second graders would be to reformat one form of the third grade pretest in a way that would be developmentally appropriate for students at the end of second grade. The third grade pretest was designed so that its content coverage was well aligned with the curriculum for the end of second grade and beginning of third grade. Consequently, Form A of the third grade pretest was reformatted to produce a student test booklet in which second graders could mark their answers (as opposed to a test booklet and separate answer sheet, which are generally used in grades three through eight.) Other modifications were made to type size and item placement as necessary to make the reformatted form more conducive to administration for second graders.

When local school administrative units (LSAU) were notified of the opportunity to participate in the study, the number of volunteers far exceeded the requirement specified in the legislation. Participants were selected from among the volunteers to maximize the similarity of the participants to the state in terms of geography and size, as specified in the legislation. The twelve LSAUs selected were: Anson County (040); Buncombe County (110); Caldwell County (140); Newton-Conover City (182); Cherokee County (200); Craven County (250); Franklin County (350); Halifax County (420); Jones County (520); Charlotte-Mecklenburg (600); Pitt County (740); and, Randolph County (760).

The final issue of whether a second grade test is more reliable for measuring academic growth has multiple aspects. These are described in turn below.

One question is whether the reformatted Form A and the original Form A of the third grade pretest are equivalent. The assumption is that mere reformatting would maintain the integrity of the test so that the same constructs are being measured by both versions. Unfortunately there was no way to check this assumption without administering both the reformatted Form A and the original Form A to the same students on both occasions (end of second grade and beginning of third grade.) However, this was administratively unacceptable for at least two reasons. The primary reason is that only the reformatted form was considered developmentally appropriate for second graders. The second reason is that this would have doubled the amount of testing required of those students participating in the study. This would have been an unjustifiable burden given the questionable validity of administering the original third grade pretest Form A to the second graders.

An important fact to remember about this study is that equivalence of forms cannot be demonstrated unequivocally because in this study the Form effect is confounded with the effect of maturation of the second graders over the summer months. Consequently, in this study we assume that the forms are equivalent and address the remaining questions in the context of this assumption. We will however, examine the performance of the same students on the two different forms administered in the spring at the end of second grade and the fall of third grade to see just how similar (or different) the performance actually is.

Then the major question for this study is whether one form is better than the other for measuring academic growth. Answering this question involves deriving growth models for third grade reading and mathematics in two ways. One derivation will use the second grade test as pretest; the other will use the third grade pretest. These two "new" models will be compared with the original third grade model developed for operational use in the ABCs. Of chief concern will be a comparison of the model parameters under each condition and the degree to which either the proficiency effect or the regression effect dominates in the model.

The last point is made more clear by recalling that the growth models for reading and mathematics generally involve three things: (a) an estimate of the statewide average growth, (b) an estimate of the true proficiency of the students in the school, and (c) a statistical adjustment for regression to the mean (because the same students are being measured on two successive occasions.) In the original ABCs growth models for reading and mathematics in grades four through eight, the largest component of the expected growth calculation is the statewide average growth. The other two effects (proficiency and regression to the mean) provide small adjustments to a school's expected growth to accommodate the particular students in the school and to tailor the growth expectation to the school.

In the ABCs growth models for grades four through eight it is true that in general the regression effect tends to dominate the proficiency effect in reading; the proficiency effect and the regression effect are nearly evenly balanced in the mathematics growth models. The practical implication is that in general it can be said that schools with lower pretest scores must grow more

to meet their growth expectations. This was considered to be a desirable feature of the growth models for the ABCs.

However, when the growth models were developed for third grade using the third grade fall pretest, the proficiency effect dominated the models for both reading and mathematics and the statewide average growth was substantially larger then the growth requirement for any of the other grades. The practical effect for schools is that third grade growth expectations are higher for schools that have higher pretest performance and the magnitude of the growth requirement is substantially larger than for other grades.

It was speculated that the difference in the model could be due to the fact that the third grade model was the only fall-to-spring growth model. All of the others involved spring-to-spring measurement. Although this was the major impetus for this study, an additional fringe benefit of this study is that some information will also be provided to indicate whether there may have been any change in the parameter estimates since they were originally established using the 1996-97 ABCs data.

These then are the major concerns of the study. How the study was designed, the research questions and the analyses that were planned are described next.

Design of the Study

Research Questions

The research questions that correspond to the concerns described in the introduction are as follows.

1) How do students who took both the reformatted Form A at the end of second grade and subsequently took the third grade pretest at the beginning of third grade score on each form, on average?

2) Are the third grade reading and mathematics growth models derived from the spring-to-spring data similar to the corresponding third grade growth models derived from the fall-to-spring data, with regard to:
   a) parameter estimates?
   b) impact of the proficiency and regression effects?

3) Has there been any change in the parameter estimates of the third grade reading and mathematics growth models since they were established using the 1996-97 ABCs data?

Data Collection

To carry out the legislative requirement for this study volunteer LSAUs were selected to participate in the study and special forms of the reading and mathematics tests were prepared for

administration at the end of second grade. The administration took place in the spring of 1999 in conjunction with the statewide administration of end-of-grade tests during the last three weeks of the school year. Second grade students who participated in the study took only the reformatted Form A at the end of the second grade during the spring of 1999.

These same students then participated in the fall 1999 administration of the third grade pretest to all third graders in North Carolina. Students received three forms (A, B, and C) in the fall administration, randomly distributed. Therefore, only a random third of the students who had participated in the Grade 2 spring administration received Form A of the third grade pretest during the fall administration. The others received one of the other equivalent forms of the third grade pretest.

Finally, these same students participated in the end of year testing at the end of third grade during the spring of the 1999-2000 school year. Again, all forms of the test were randomly distributed among the students.

This data plan for the study is summarized in Table 1.

Table 1
Data Collection for the Study

Timeframe	Grade	"Volunteer" Sample	Remainder of NC
Spring, 1999	2	Reformatted Pretest Form A	was not tested
Fall, 1999	3	Pretest Forms A, B, C	Pretest Forms A, B, C
Spring, 2000	3	End-of-Grade Forms	End-of-Grade Forms

This data collection design provides two alternatives for testing Research Question 2. Growth models can be derived based on data for only those students (a random third of the volunteers) who took Form A (reformatted grade two, and third grade pretest) on both occasions. Alternatively, growth models can be derived based on the data for all students who took the second grade test and who subsequently took *any* form (A, B, or C) in the fall. Because Forms A, B, and C are equivalent forms the two analyses should produce approximately the same results. This will provide a replication within the study that should strengthen the conclusions.

Analyses

To address Research Question 1, simple descriptive statistics were computed for the data from students who took both the reformatted Form A (second grade) and Form A of the third grade pretest. A t-test for dependent samples was computed to ascertain whether any observed differences are statistically significant. Additionally, the correlation between students' scores on the two forms was computed.

To address Research Question 2, growth models were first derived for reading and mathematics using the second grade tests as pretests, and then using the third grade pretests. This analysis was replicated for: (1) the group of students who took Form A on both pretest occasions, and (2) the group of students who took Form A on the first occasion (second grade) and then took any one of the three forms on the second occasion (i.e., beginning of third grade.)

To address Research Question 3, reading and mathematics growth models were derived using statewide matched data for each year since the initial year of the ABCs. The characteristics of these models were then compared with the characteristics of the operational models that are currently in effect.

Results

Research Question 1:
How do students who took both the reformatted Form A at the end of second grade and subsequently took the third grade pretest at the beginning of third grade score on each form, on average?

Table 2 shows the results of comparing the pretest means for the sample of 3,588 students who took the reformatted Form A in second grade and later took Form A during the fall pretest in third grade. Although the differences are statistically significant, their magnitude is on the order of one-half of a scale score point and is probably not educationally significant.

For the reasons mentioned earlier, it cannot be inferred that the observed differences are purely due to format differences between the two forms for reading and mathematics, because the students have experienced several months maturational experience intervening between the two administrations.

The correlation between students' performance on the two measures of reading was 0.77, while the correlation between the two measures of mathematics was 0.78. Although alternate forms reliability estimates should be higher, these correlations also reflect the influences of the summer experiences of students.

Although these differences exist, they are not so pronounced as to raise an alarm regarding the assumption that the forms are equivalent for purposes of this study.

Table 2
Comparison of the Spring and Fall Pretests

Pretest (Form A)	Mean	Standard Deviation	Difference ($3^{rd} - 2^{nd}$)	Probability
		Reading		
Second Grade	140.1	7.5		
Third Grade Pre	140.7	8.3	0.6	$p < .0001$
		Mathematics		
Second Grade	134.1	7.1		
Third Grade Pre	133.5	7.9	-0.6	$p < .0001$

Research Question 2:
Are the third grade reading and mathematics growth models derived from the spring-to-spring data similar to the corresponding third grade growth models derived from the fall-to-spring data, with regard to:

a) parameter estimates?
b) impact of the proficiency and regression effects?

Table 3 shows the results of deriving growth models for reading and mathematics for the students in the study, depending on which forms they took as pretests. The table shows the values of the parameter estimates for the growth model as defined in *Setting Annual Growth Standards: "The Formula"* (Accountability Brief, Vol. 1 No. 2, Revised June 2000.) The key values to note are the columns labeled b_0, b_1, b_2, and $2b_1 + b_2$. The first three values are associated with statewide average growth, proficiency and regression to the mean, respectively. The fourth value is an approximate indication of the relative dominance of the proficiency or the regression effect. When $2b_1 + b_2$ is positive, the proficiency effect is dominant and higher performing schools have higher growth expectations. On the other hand, when $2b_1 + b_2$ is negative the regression effect is dominant and lower performing schools have higher growth expectations.

(Note: In the original growth models for grades four through eight, the regression effect dominated the reading growth models and the two effects nearly cancelled each other for mathematics. Consequently, for the operational ABCs growth standards, generally higher growth standards are set for schools whose initial performance is low.)

As can be seen in Table 3, for reading the analyses indicate that the parameter estimates differ depending on whether the second grade test or the third grade test is used as the pretest for the

Table 3

Comparison of Growth From the End of Grade 2 and Beginning of Grade 3 to the End of Grade 3

Regression Coefficients, Standard Errors, and Other Statistics Associated with Year-to-year Change for Matched Cohorts

Reading

1999-2000 Cohort Group	Pretest Grade	b_0	s.e.	b_1	s.e.	b_2	s.e.	rmse	$2b_1+b_2$	N
Form A for Both Pretests	Second Grade	6.94	0.10	0.47	0.02	-0.87	0.03	5.94	0.08	3,588
	Third Grade	6.28	0.10	0.40	0.02	-0.82	0.03	5.83	-0.02	3,588
Both Pretests (any form)	Second Grade	6.95	0.06	0.47	0.01	-0.86	0.02	6.02	0.09	10,961
	Third Grade	6.63	0.06	0.42	0.01	-0.88	0.02	5.92	-0.04	10,961

Mathematics

1999-2000 Cohort Group	Grade	b_0	s.e.	b_1	s.e.	b_2	s.e.	rmse	$2b_1+b_2$	N
Form A for Both Pretests	Second Grade	9.74	0.12	0.43	0.02	-0.54	0.04	7.04	0.31	3,588
	Third Grade	10.26	0.12	0.39	0.02	-0.59	0.04	6.90	0.19	3,588
Both Pretests (any form)	Second Grade	9.75	0.07	0.40	0.01	-0.51	0.02	6.98	0.29	10,961
	Third Grade	11.02	0.07	0.35	0.01	-0.54	0.02	6.96	0.16	10,961

growth model. The estimate for statewide average growth in reading, b_0, is slightly larger when the second grade test is used as pretest. This is logical because of the longer time intervening between the pretest and the posttest when the second grade test is used. It is also consistent with the general belief that reading skills may continue to accumulate during the summer months between school years. The values of b_1 and b_2 also differ slightly depending on which test is the pretest.

The critical result for the growth model is the value of $2b_1 + b_2$. When the second grade test is used as the pretest, the proficiency effect dominates. When the third grade pretest is used the regression effect dominates slightly. This result is contrary to expectation, because when the current operational model was developed based on a fall-to-spring model for third grade, the proficiency effect dominated slightly for reading. In these results the fall-to-spring model shows a dominant regression effect.

The results are essentially the same regardless of whether one uses only the students who took Form A on both occasions, or the group who took Form A at the end of second grade followed by any one of the three forms at the beginning of third grade.

Similarly for mathematics, the parameter estimates differ depending on whether the spring test or the fall test is used for pretest. However, in the case of mathematics the value of b_0 (the estimate of statewide average growth) is actually larger when the fall pretest is used in spite of the fact that this represents a shorter time lapse between the pretest and the posttest. A generally accepted notion is that mathematics education is more restricted to the formal learning that takes place in school and does not improve, or may actually decline over the summer months. If so, this could explain this difference in values of b_0.

Values of b_1 and b_2 for mathematics also differ slightly depending on whether a spring or fall pretest is used. Consequently, the value of $2b_1 + b_2$ differs depending on which model is used. However, once again the result is unexpected. The proficiency effect dominates in both cases, but is less pronounced when the fall-to-spring growth model is used. Again, the results are consistent across the two groups of students analyzed.

The general message for the growth standards is clear. These data suggest that the fall-to-spring growth model is more desirable if one wants lower performing schools to have higher growth expectations. The results also suggest that the third grade growth model currently in effect may need some revision to bring it more in line with the results from more recent data. However, statewide data should be used to confirm this contention before making a recommendation. This leads to the results for Research Question 3.

Research Question 3:
Has there been any change in the parameter estimates of the third grade reading and mathematics growth models since they were established using the 1996-97 ABCs data?

Table 4 shows the results of deriving the growth models for reading and mathematics from statewide data each year since the original implementation of the ABCs. The results shown in

the table for the 1996-97 school year are the operational models currently adopted for use in third grade.

It is clear that the values of b_0 derived from more recent reading data have been somewhat larger than the value used in the operational reading growth model. This is what one would expect because the state has experienced widespread growth since the inception of the ABCs. However, the 1999-2000 results indicate a decline in b_0 from the values observed for 1997-98 and 1998-99. This is consistent with the fact that generally fewer schools made expected or exemplary growth in 1999-2000.

The values of b_1 and b_2 have been relatively stable. However it is interesting that the value of $2b_1 + b_2$ has been negative for reading every year since 1996-97, when it was positive. These more recent values suggest that the reading growth model for third grade may need to be updated, making it more consistent with the reading growth models in effect for the other grades (in terms of setting higher standards for lower performing schools.)

The mathematics analyses show that recent values of b_0 have been somewhat smaller than for the operational model. This in turn suggests that the operational model may set mathematics growth goals that are a bit too high. The values of b_1 and b_2 for the mathematics growth model have been remarkably consistent across all years. The value of $2b_1 + b_2$ has remained unchanged. However, it continues to be positive. This indicates that the proficiency effect dominates in mathematics. The effect is not as strong as suggested by the analyses for Research Question 2 however. These results taken together suggest that the current mathematics growth model is stable, but may need a slight adjustment to the value of statewide average growth.

Table 4

Comparison of Reading and Mathematics Growth Models Derived from Third Grade Data for Different Years

Regression Coefficients, Standard Errors, and Other Statistics Associated with Year-to-year Change for Matched Cohorts

Reading

Cohort Year	Grade	b_0	s.e.	b_1	s.e.	b_2	s.e.	rmse	$2b_1+b_2$	N
1996-97	3pt-3	6.23	0.02	0.46	0.00	-0.91	0.01	6.03	0.01	
1997-98	3pt-3	7.79	0.02	0.46	0.00	-0.93	0.01	6.22	-0.02	93,040
1998-99	3pt-3	8.20	0.02	0.46	0.00	-0.94	0.01	6.24	-0.01	96,376
1999-00	3pt-3	7.64	0.02	0.44	0.00	-0.92	0.01	6.09	-0.04	96,977

Mathematics

Cohort Year	Grade	b_0	s.e.	b_1	s.e.	b_2	s.e.	rmse	$2b_1+b_2$	N
1996-97	3pt-3	12.79	0.02	0.30	0.00	-0.47	0.01	7.12	0.13	
1997-98	3pt-3	12.04	0.02	0.30	0.00	-0.47	0.01	7.18	0.13	93,040
1998-99	3pt-3	11.96	0.02	0.30	0.00	-0.47	0.01	7.13	0.13	96,376
1999-00	3pt-3	11.95	0.02	0.31	0.00	-0.48	0.01	7.10	0.13	96,977

Note:

3pt-3 indicates that the growth is based on change from the third grade pretest given in the fall of each year to the end-of-grade test given to third grade students during the last three weeks of school.

Summary and Discussion

In the various analyses, we have seen evidence that student performance differs only slightly depending on whether students are tested at the end of second grade or the beginning of third grade. The growth models derived from spring-to-spring data also differ slightly from those derived from fall-to-spring data. However, switching to a spring-to-spring model would not solve the perceived problem of having a dominant proficiency effect in the third grade growth model.

Contrary to expectation, the data collected for the study as well as more recent statewide data suggest that the way to achieve a dominant regression effect (for reading at least) is to continue to adhere to a fall-to-spring growth model but to update the model parameter estimates. It appears that the mathematics growth model for third grade is characterized by a proficiency effect, but it is less pronounced with a fall-to-spring growth model. So in both cases, the data suggest a fall-to-spring growth model is preferable.

During the study, teachers were asked their opinions about the timing of the pretest. They were nearly evenly split on the issue. When teachers in the volunteer LSAUs were asked to indicate the best time to administer the pretest for third grade, 44.1% indicated the last three weeks of second grade and 45.2% indicated the first three weeks of third grade. (10.7% did not express a preference)

It is decidedly more expensive to administer the pretest at the end of second grade because a developmentally appropriate test booklet is generally longer and must be adapted for scoring. It is also more expensive and logistically complicated to score a test booklet rather than a separate single-page answer sheet.

The data from the study, recent statewide data, teacher opinion, cost and logistical considerations all contribute to the same conclusion. There is no compelling reason to switch to testing at the end of second grade in order to revise the growth models for third grade reading or mathematics.

Recommendation

The State Board of Education should consider updating the third grade growth models for reading and mathematics based on more recent statewide data than was available at the time of first implementation. Any changes must be coordinated with planned or anticipated changes that will be necessary as a result of curriculum revisions and attendant revisions of the North Carolina testing program.