

**Testing What Students in the Gap Can Do**

Michael Russell

Lisa Famularo

Technology and Assessment Study Collaborative

Boston College 323 Campion Hall

Chestnut Hill, MA 02467

July 2008

Support for this investigation was funded in part by Grant CFDA #84.368 of the U.S.

Department of Education, Office of Elementary and Secondary Education, awarded to the Rhode Island Department of Education. The opinions expressed in this article do not necessarily reflect the position or policy of the Department of Education, and no official endorsement should be inferred.

## Background

Student assessment is an integral component of classroom instruction. Assessment is intended to help teachers identify what students are able to do and what content and skills students must develop further. State tests play an important role in guiding instruction. However, for some students, the tests may lead to inaccurate conclusions about student mastery of key content and skills. For other students, the tests may provide information about what they cannot do, but do not help inform teachers about what students can do. This is particularly true of mathematics items that require students to apply multiple skills to solve them. A student's failure to apply accurately any one of the required skills may lead to inaccurate responses despite the student's ability to apply all other skills accurately and appropriately.

As part of the New England Compact's Task Module Assessment System (TMAS) project (New England Compact, 2007; Bechar & Godin, 2007; Parker & Saxon, 2007a, 2007b), we sought to develop a better understanding of how test items could be altered to provide more information about what students can and cannot do. Specifically, the study aimed to develop a prototype test.

The prototype test was designed to measure students' understanding of the skills associated with the New England Common Assessment Program (NECAP) Functions and Algebra Grade-Level Expectations (GLE) and was based on a set of items from the 2005-2006 NECAP assessment. Four complex problems that required students to apply multiple skills and concepts were selected. A series of new items designed to tap skills related to the NECAP items and corresponding GLEs were developed with two purposes in mind. First, items were developed to more deeply probe students' understanding of the component skills both

individually and in combination. Second, alternate representations of items were developed to determine if modifications to the original problem would enable students to solve it.

The purpose of the study was to (a) assess the quality and usefulness of test items designed to decompose skills and knowledge required to solve complex algebraic problems; and (b) examine the extent to which students in the gap are able to succeed on decomposed items while struggling with the complex items.

### *Data Collection*

To recruit students for the study, teachers in NH, VT, and RI were contacted by email and were invited to participate in a pilot study of the prototype test. A total of 81 teachers from 64 schools in New Hampshire, Rhode Island and Vermont participated. Data were collected from 2,365 8<sup>th</sup> grade students in these states in the spring of 2006. Consistent with the demographics of the participating states, the majority of students identified themselves as White (77%), 17% as minority and 6% as Latino. Gender was divided relatively evenly with 48% of students identifying themselves as males and 50% as females (2% did not identify as either male or female).

### *Instruments*

All data collection instruments were delivered on-line. Teachers were asked to complete a questionnaire that asked them to identify whether each participating student in their class currently had an IEP, 504 Plan, or was receiving LEP services. Teachers were also asked to provide a rating of each student's proficiency in mathematics as displayed in class and a prediction of how well the student would perform on the mathematics portion of the 2005-06 NECAP/MEA test.

Students completed a background survey and the prototype test. The background survey collected information about each student's gender, race/ethnicity, a rating of their math ability, and information about their access to computers (number at home and length of time owned) and comfort with computers. The prototype test contained 43 multiple-choice items, grouped into four sets of items, which we refer to as "item families." Each item family contained one "parent item" from the 2005-06 NECAP mathematics assessment, a different representation of that item (called a "sibling item") and a series of follow-up items, referred to as "child items." The procedures for developing the test are described below.

### *Test development*

The project team worked with math specialists from two states to identify four NECAP items to use as parent items. Three criteria were used to select the items. First, the items had to be complex problems that required students to apply multiple skills and concepts. Second, the parent items had to be of moderate difficulty (the item difficulties ranged from .56 to .66). Finally, the items had to represent the four NECAP Functions and Algebra Grade Level Expectations.

The child items were developed with two purposes in mind. First, items were developed to more deeply probe students' understanding of the component skills both individually and in combination. Second, alternate representations of items were developed to determine if modifications to the original problem would enable students to solve it. For example, in some cases the context of the problem was modified or removed while for others, the presentation of information was changed (i.e. from vertical to horizontal, from a table to a graph) and in some cases, the problem was simplified (i.e. by using whole numbers rather than fractions, reducing the number of variables, or reducing the number of steps required to solve the problem). Finally,

for each item family one sibling item (a different representation or isomorph of the parent) was created.

The draft items were sent to math specialists in each state and to an experienced item developer at Measured Progress (the test developer) for review and feedback. After several rounds of revisions the final item sets were programmed for online administration. The final product was a computer-based test module containing four parent items each with one sibling and up to 13 children. In the interest of space, one of the Linear Pattern item family is described in detail below. In addition to this family, family item sets were also developed for Evaluating an Expression, Equality, and Rate of Change. Three items from one item family (Linear Pattern Family) are shown in Figure 1.

**Figure 1: Sample Items for the Linear Pattern Family**

**Parent**

Mandy is buying mats for an exercise class. The pricing chart below is missing some prices.

Number of mats	Cost for each mat
1-4	\$14.00
5-9	\$12.65
10-14	\$11.30
15-19	\$9.95
20-24	
25-29	
30+	\$5.90

Based on the pattern in the chart, how much will each mat cost if Mandy buys 26 mats?

- A. \$6.25
- B. \$7.25**
- C. \$7.93
- D. \$8.60

GLE 7.1: Identifies and extends a linear pattern represented in a table and in a problem situation.

- A. student makes a calculation mistake
- B. key
- C. balance and/or identifies the mean of the last two given prices
- D. student identifies the 20-24 price range

**Child #2**

Ellen is buying tickets for a hockey game. The pricing chart below is missing some prices.

Number of tickets	Cost for each ticket
1	\$37
2	\$31
3	\$26
4	\$22
5	
6	
7	\$16

Based on the pattern in the chart, how much will each ticket cost if Ellen buys 6 tickets?

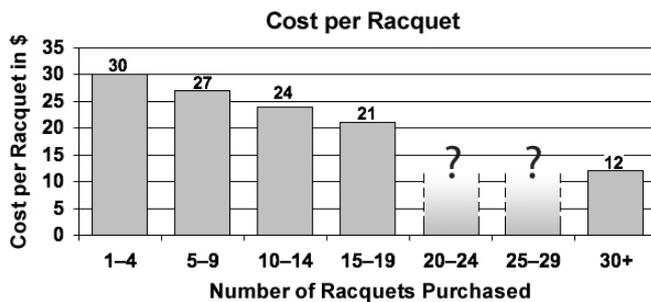
- A. \$17
- B. \$18
- C. \$19
- D. \$20

Identifies and extends a nonlinear pattern represented in a table and in a problem situation.

Pattern in the table is nonlinear. Number of tickets is not displayed as a range. Math is simplified by using whole numbers.

**Child #6**

Maria is buying racquets for a tennis tournament. The pricing chart below is missing some prices.



Based on the pattern in the chart, how much will each racquet cost if Maria buys 26 racquets?

- A. \$14
- B. \$15
- C. \$17
- D. \$18

Identifies and extends a linear pattern represented in a graph and in a problem situation.

Information is presented in a graph and the math is simplified by using whole numbers for the dollar amounts.

This item is below grade level.

### *Linear Pattern Item Set*

As shown in Figure 1, the parent item requires the student to identify and extend a linear pattern (GLE 7.1) presented in a table and in a problem situation (Mandy is buying mats for exercise class). The sibling item involves a different problem situation (Holly is training for track), but like the parent, requires the student to identify and extend a linear pattern presented in a table. There are 13 child items. Four of the child items (child 2, 5, 10 and 11) differ from the parent in that they involve non-linear patterns. Five child items (child 6, 7, 8, 9 and 10) present the information in a sequence or a graph. Other child items differ from the parent in that the pattern increases in value rather than decreases in value (child 7 and 10); simpler information is displayed in the table (child 1 and 2); the layout of the information presented in the table is presented horizontally rather than vertically (child 4 and 5); and the problem situation has been removed (i.e. child 7, 8, 9 and 10). . There are eight below grade items (child 2, 3, 5, 6, 9, 10, 11 and 12).

### *Additional Sources of Data*

Data collected during the 2005-06 NECAP assessment were also used for this study. Students' NECAP scores, teachers' judgment of proficiency in mathematics, and demographic and accommodations information from student test booklets were used to identify students in the gap. A more detailed description of how gap kids were identified is provided in the analysis section of the report.

### *Data Analyses*

Classical Test Theory (CTT) analyses were used to assess the quality and usefulness of individual test items and to examine how the items in each family relate to one another. Our

analyses began with an examination of item difficulty. Next, Exploratory Factor Analysis was used to examine whether the items in each family appear to be measuring the same skill. Then we examined the reliability of each item family and the test as a whole. A factor analysis conducted on all items revealed four factors. Each factor was comprised of items from just one family set (parent, sibling, and children). For each family, the majority of the factor loadings ranged from .4 to .7, indicating that each item set was tapping the same construct (for more complete details of the factor analysis, see Famularo & Russell, 2007). Findings from the item difficulty and reliability are presented below.

### *Item Difficulty*

Within each item family, we expected the child items to be easier than the parent item. The sibling items were intended to be alternate representations of the parent items, therefore in each item family we expected the sibling and parent items to have about the same item difficulty.

As expected, the parent item was more difficult than the child items for three item families: linear pattern, evaluating an expression, and rate of change (Table 1). The expected pattern did not hold for the equality item family. Specifically, for the equality item family there were three child items with roughly the same item difficulty as the parent and one item was much more difficult. Upon further investigation, it was evident that the more difficult item was problematic and was removed from subsequent analyses. Given the large sample size, the change in item difficulty between the parent item and the child items was statistically significant for all but one item (alpha was adjusted for multiple comparisons within each item family). In addition, effect sizes<sup>1</sup> were within .1 standard deviations for 4 child items, between .11 and .30 for 13 child items, and greater than .30 for 18 child items.

---

<sup>1</sup>Effect sizes were calculated as follows: (% Correct Child – % Correct Parent)/Standard Deviation of Parent.

The sibling items, which were intended to be the same difficulty as the parent item, did not perform as expected. As shown in Table 1, the sibling items were easier than the parent items by 5 to 29 percentage points. All of the differences between the parent and sibling items were statistically significant, with effect sizes ranging from .10 to .60. A possible explanation for this result is that the sibling items were easier for students to answer due to a practice effect. Students were always presented with the parent item first, followed by the child items then the sibling. Perhaps more students were also able to answer the sibling correctly because they answered it after they had responded to a series of items that tapped into similar skills.

Table 1  
Item Difficulty

	Linear Pattern	Evaluating an expression	Rate of Change	Equality
Parent	.58	.68	.64	.60
Sibling	.63* (.10)	.74* (.13)	.93* (.60)	.70* (.20)
Child 1	.64* (.12)	.82* (.30)	.85* (.44)	.60 (.00)
Child 2	.65* (.14)	.80* (.26)	.89* (.52)	.88* (.57)
Child 3	.78* (.41)	.84* (.34)	.82* (.37)	.58* (-.04)
Child 4	.68* (.20)	.76* (.17)	.89* (.52)	.88* (.57)
Child 5	.65* (.14)	.86* (.39)	.79* (.31)	.76* (.33)
Child 6	.79* (.43)	.83* (.32)	.84* (.42)	.22* (-.78)
Child 7	.71* (.26)			.78* (.37)
Child 8	.69* (.22)			.63* (.06)
Child 9	.89* (.63)			.58* (-.04)
Child 10	.80* (.45)			.78* (.37)
Child 11	.70* (.24)			
Child 12	.67* (.18)			
Child 13	.66* (.16)			

*\*Significant at the .05 with alpha adjusted for multiple comparisons within each item family. Effect sizes are reported in parentheses.*

### Reliability

To examine the internal consistency of the item sets, coefficient alpha was calculated for each item family as well as for the test as a whole. Item discrimination, an index of how

effectively an item differentiates between high-scoring and low-scoring students was also examined. The vast majority of item-total correlations ranged from .4 to .6 and were considered acceptable. The reliability coefficient for each item family was as follows: Linear Pattern (.84), Evaluating an expression (.84), Rate of Change (.68) and Equality (.77).

### *Identifying Gap 1 and Gap 2 Students*

A series of analyses was conducted to examine the performance of students who are in an assessment gap and those who are not. Scaled scores on the NECAP mathematics test were used to place students into subgroups. The subgroups used for the analyses are described below.

Students in gap 1 are defined as low-scoring students whose teachers rated their performance in class as proficient. In other words, there is a discrepancy between their performance on the NECAP and their teachers' rating of their proficiency. This group is comprised of students whose scaled score on the NECAP mathematics test plus one standard error of measurement (SEM) places them in the "not proficient" category (achievement level 1 or 2). The SEM was used to exclude students close to the cut point between proficient (achievement level 3) and not proficient (achievement level 2) as it is possible due to measurement error, that if tested again, these students would achieve a score that places them in the proficient category. Approximately 96 students in the sample (4.7%) were classified as students in gap 1. These students represent 26.4% of the non-proficient students for whom we have teacher ratings (n=364).

Students in gap 2 are those who scored in the lowest achievement level. This group is comprised of students whose scaled score on the NECAP mathematics test plus one SEM places them in achievement level 1. Approximately 212 students in the sample (9%) were classified as students in gap 2.

Students in the comparison group are those who scored in the proficient category (achievement level 3 or 4). This group is comprised of students whose scaled score on the NECAP mathematics test minus one SEM places them in achievement level 3 or 4. Approximately 1054 students in the sample (44.6%) were classified as students in the comparison group.

#### *Comparing performance on child items with performance on parent*

To answer the question “to what extent are students in the gap able to succeed on the child items while struggling with the more complex parent items” we compared the item difficulty of the parent item with the item difficulties of the child items. Comparisons for each item family are shown in Figures 2 through 5. The items in each chart are ordered from left to right in terms of the difference in performance between comparison and gap students with the item with the largest difference appearing on the left and the item with the smallest difference on the right.

#### *Linear Pattern Family*

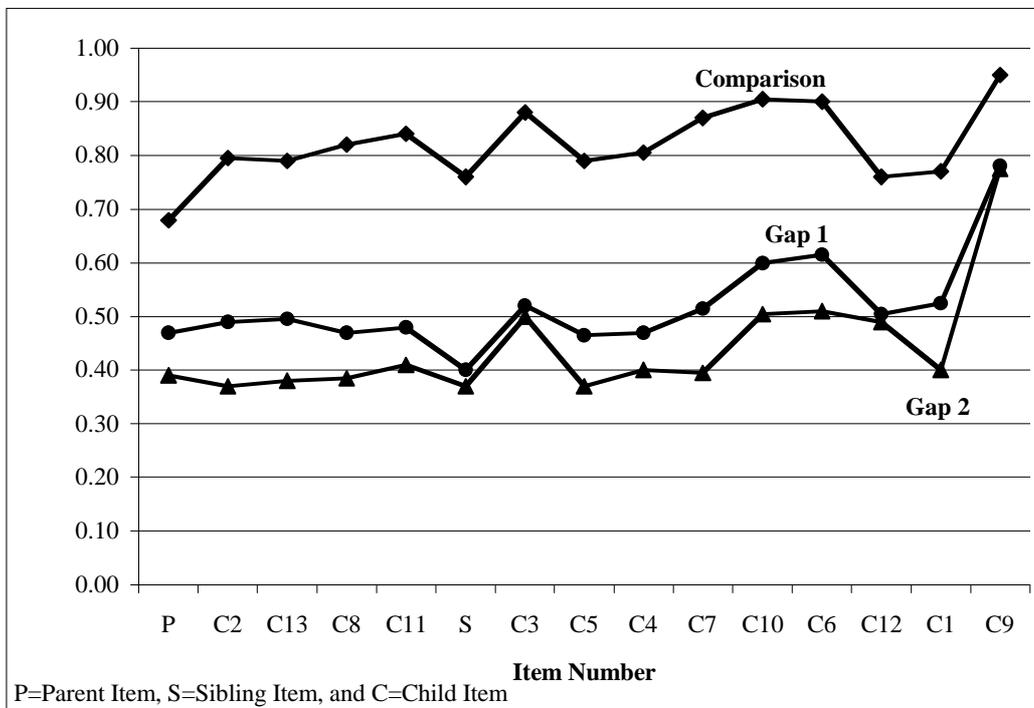
There is a fairly large difference in the performance of comparison and gap students on the parent item; there is a .23 difference between comparison and gap 1 students and a .31 difference between comparison and gap 2 students. Figure 2 reveals that almost all of the items in this family have an equally large or larger difference in performance between comparison and gap students as the parent. There is only one child item (child 9) where the difference between comparison and gap 1 students is less than that of the parent; the difference for this item is .19. There are two items (child 12 and 9) where the difference between comparison and gap 2 students is less than that of the parent; the differences for these items are .28 and .21 respectively. Child 9 and child 12 are very similar to one another and are also very similar to two

other items, child 3 and child 6. All four of these items are below grade level and all used the same pattern of numbers. Child 12 and child 3 presented the information in a chart, child 6 presented the information in a graph, and child 9 presented the numbers in a horizontal sequence without a problem situation. It is possible that the smaller difference in performance we see for child 9 and 12 resulted from modifications in the presentation of the information or perhaps it is the result of students having had two other opportunities to work with the same pattern of numbers.

Figure 2 also suggests that gap 1 students tended to do better than gap 2 students on the linear pattern items. However, there are five child items (child 2, 13, 3, 12 and 9) for which performance of gap 1 and gap 2 students is similar.

Figure 2

Child Item Difficulty by Gap Group: Linear Pattern



### *Evaluating an Expression Family*

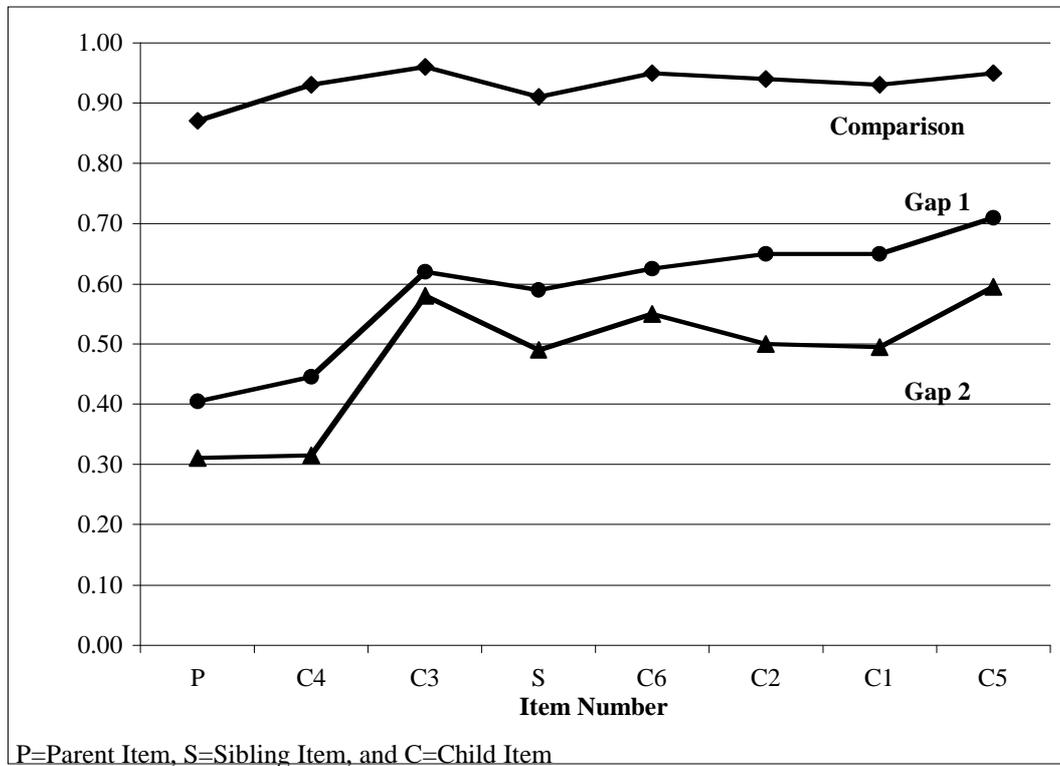
The items in the evaluating an expression family reveal a very large difference in the performance of comparison and gap students on the parent item (Figure 3). There is a .46 difference between comparison and gap 1 students and a .55 difference between comparison and gap 2 students. All but one child item (child 4) show a difference between comparison and gap students that is smaller than that of the parent. The two below grade-level items (child 5 and child 1) are the items where the gap in performance is reduced the most. The other items which appear to reduce the gap are problems where the math was simplified by using whole numbers (child 2), the problem was simplified by using variables equal to 1 (child 3), and the problem involved identifying the correct algebraic expression (child 6) as opposed to solving an equation. The item that does not appear to reduce the gap in performance is the only item in the family without a problem situation.

The parent item presented a problem situation where the student was required to solve an equation involving two variables. Performance on the child items suggests that removing the problem situation does not reduce the gap in performance between gap students and comparison students but suggests that simplifying the problem (by using whole numbers or variables equal to 1) and having the student demonstrate understanding of algebraic expressions without requiring them to solve an equation (which was the case with child 6) might.

The chart also suggests that gap 1 students tended to do better than gap 2 students on the evaluating an expression items. However, there is one child item (child 3) for which performance of gap 1 and gap 2 students is similar.

Figure 3

Child Item Difficulty by Gap Group: Evaluating an Expression



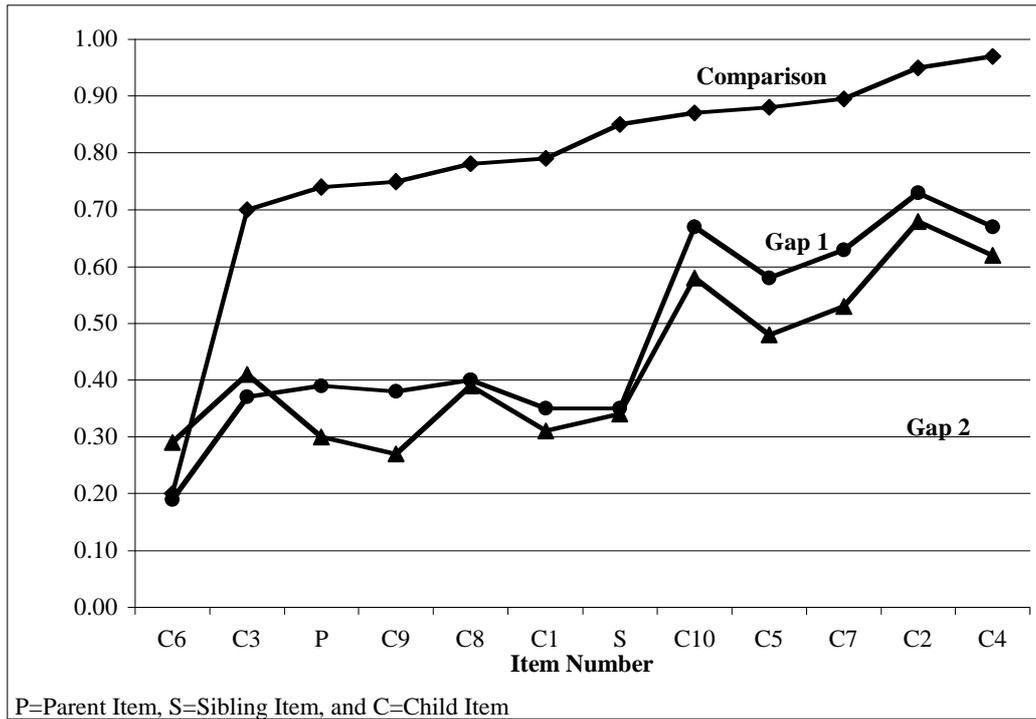
*Equality Family*

Figure 4 presents the items in the equality family and reveals a large difference in the performance of comparison and gap students on the parent item; there is a .37 difference between comparison and gap 1 students and a .45 difference between comparison and gap 2 students. Five of the child items show a difference in performance between comparison and gap students that is smaller than that of the parent – two are below grade-level items (child 2 and child 4) and three are not (child 7, 5 and 10). Child 7 differs from the parent in that it is a one variable equality problem; child 5 is a one-step simplification problem with two variables; and child 10 involves transforming information presented in an equation into a picture. Figure 4 also suggests that gap

1 students tended to do better than gap 2 students on the equality family items. However, there is one child item (child 1) for which performance of gap 1 and gap 2 students is similar.

Figure 4

Child Item Difficulty by Gap Group: Equality



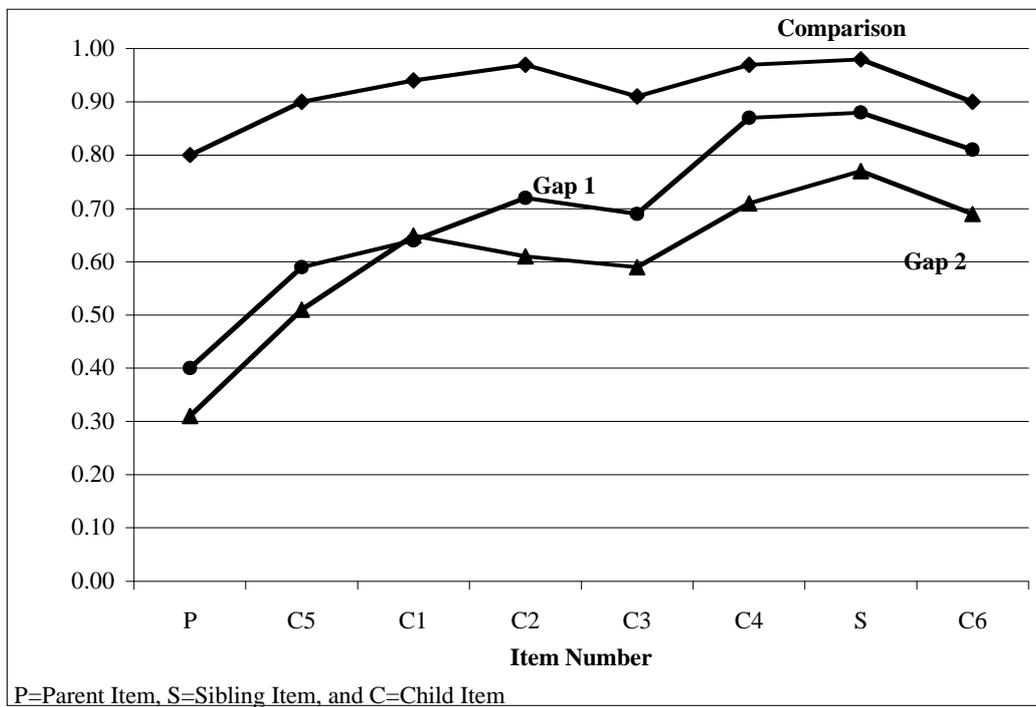
### Rate of Change Family

The items in the rate of change family (figure 5) exhibit a large difference in the performance of comparison and gap students on the parent item; there is a .40 difference between comparison and gap 1 students and a .49 difference between comparison and gap 2 students. All of the items have smaller differences in performance between comparison and gap students than the parent. The smallest differences in performance between the comparison and gap groups occur on child 6, child 4 and the sibling. These items differ from the parent in that the rate information is not presented in a table; it is presented in graphs (child 4 and sibling) or as a sequence of numbers without a problem situation (child 6).

Figure 5 also suggests that gap 1 students tended to do better than gap 2 students on the rate of change items. However there is one child item (child 3) for which performance of gap 2 students exceeded that of gap 1 students and two items (child 8 and the sibling) for which performance of gap 1 and gap 2 students is similar.

Figure 5

Child Item Difficulty by Gap Group: Rate of Change



*Examining performance on below grade level items*

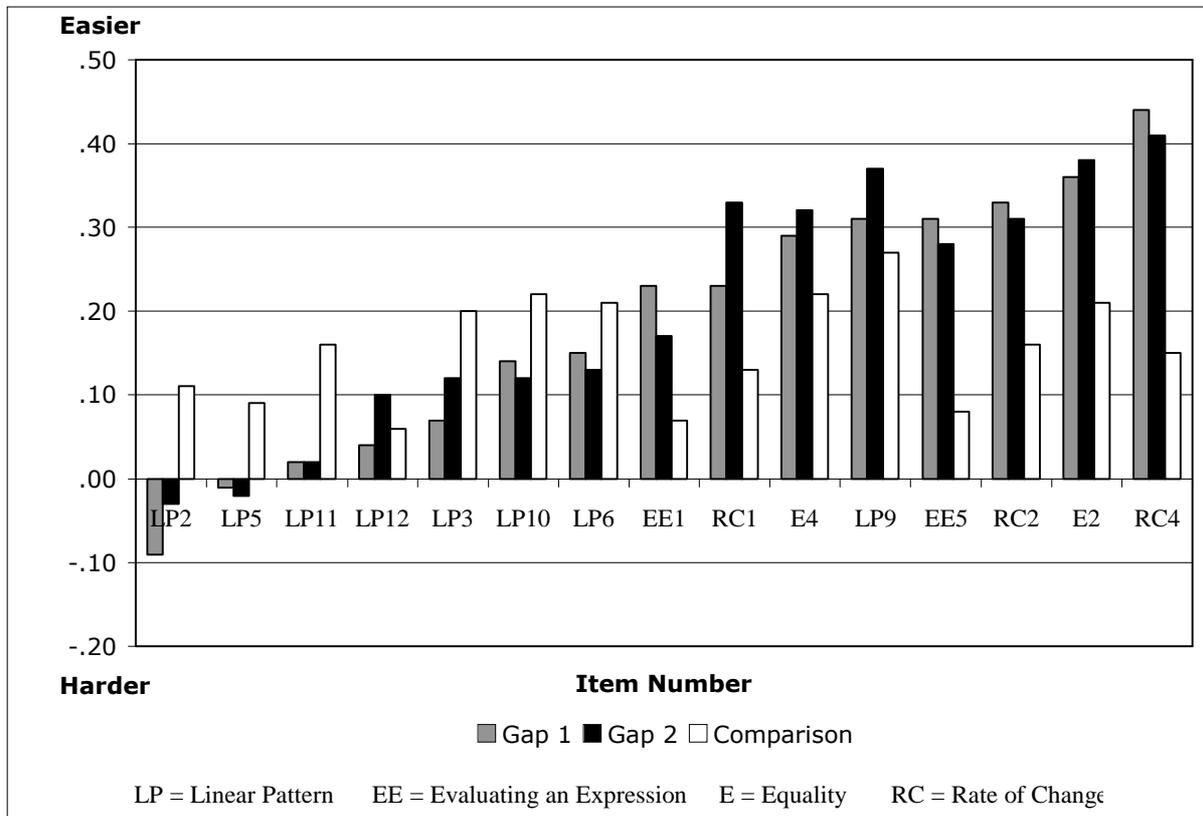
There were a total of 15 below grade level items across the four item families. For each of these items, we calculated the item difficulty separately for gap 1, gap 2 and comparison group students and subtracted it from the respective parent item difficulty. These differences are show in Figure 6 below; items with the smallest difference in item difficulty for gap students are on the left and items with the largest differences are on the right.

As shown in Figure 6, two of the below grade level items in the linear pattern family (LP) were more difficult for gap students than the parent item as evidenced by the negative difference in item difficulty. On the other hand, LP9 appears to be much easier for gap students, with about 31% more gap 1 and 37% more gap 2 students answering it correctly. The remaining below grade level items in this family are only slightly easier than the parent with differences in p values ranging from .2 to .16.

The below grade level items in the evaluating an expression (EE), equality (E) and rate of change (RC) families appear to be much easier for gap students with 25% to 45% more gap 1 and 18% to 42% more gap 2 students answering these item correctly. It is interesting to note that some items (i.e. EE1 and EE5) had a greater impact on gap 1 students while other items (i.e. RC1, E4 and LP9) had a greater impact on gap 2 students.

Figure 6

Differences in item difficulty for below grade level items



*Examining performance on items without a problem situation*

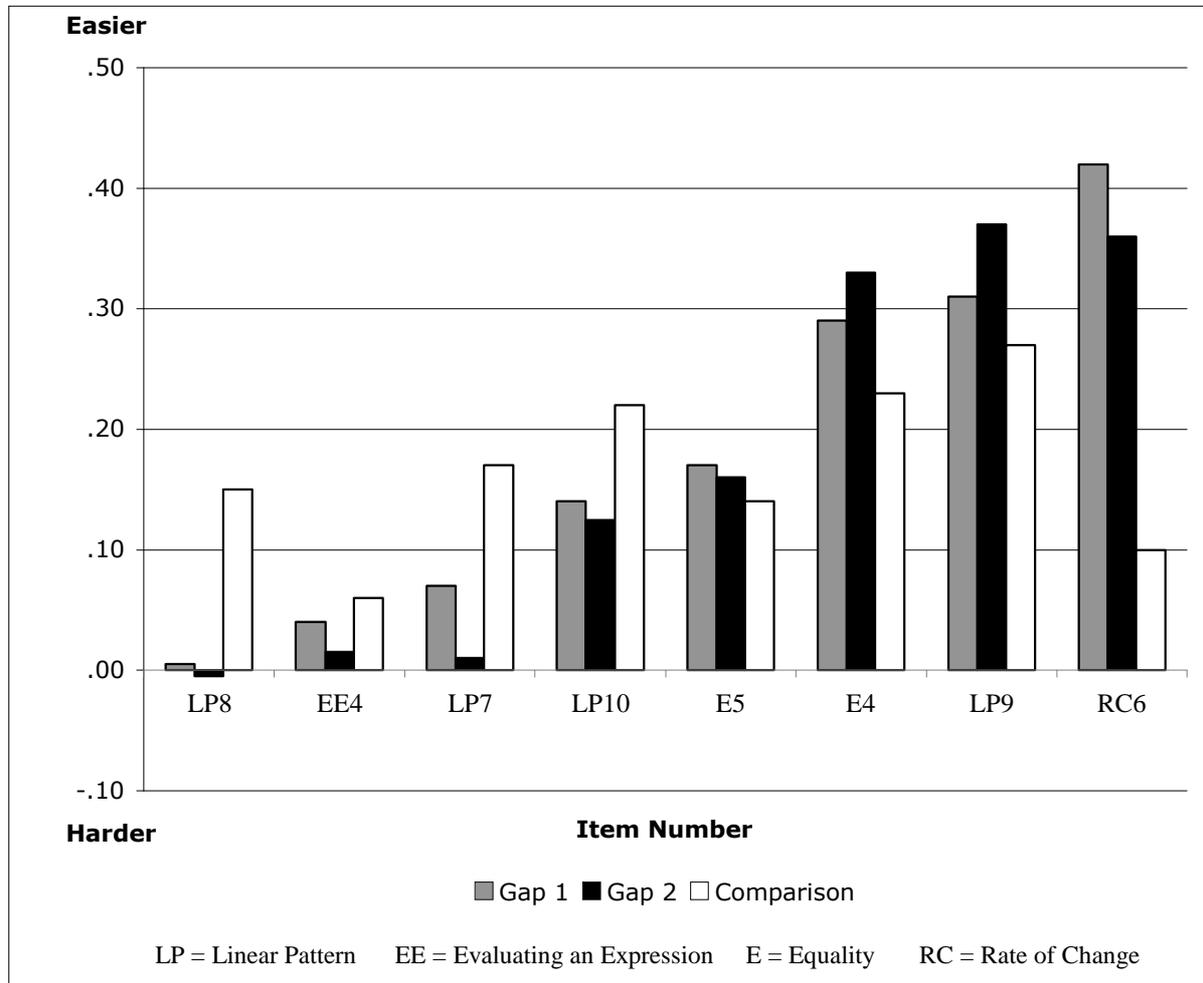
This section examines the impact of removing the problem situation (context of the problem), which in many cases reduced the linguistic complexity of the items. There were a total of eight child items that did not have a problem situation. For each of these items, we calculated the item difficulty separately for gap 1, gap 2 and comparison group students and subtracted it from the respective parent item difficulty. These differences are shown in Figure 5 below; items with the smallest difference in item difficulty for gap students are on the left and items with the largest differences are on the right.

As shown in Figure 7, three items (LP8, EE4, and LP7) were about as equally as difficult for gap 2 students as the parent and only slightly less difficult for students in gap 1. Items LP10 and E5 were slightly easier for gap students with differences in p values around .15. E4, LP9 and RC6 are much easier for gap students than were the parent items, with between 29% and 42% more gap 1 and 33% to 37% more gap 2 students answering these items correctly. It is important to note that three of these items (LP9, LP10, and E4) were also categorized as below grade level items.

Another way of addressing the question “to what extent are students in the gap able to succeed on the child items while struggling with the complex (parent) items” is to examine the correlation between the parent and child items and how often performance on the parent and child items is consistent. The phi correlation coefficients ( $\phi_p$ ) and percent consistent performance were calculated separately for students in the comparison, gap 1 and gap 2 groups. The results for each item family are shown in Tables 2 through 5.

Figure 7

Differences in item difficulty for items without a problem situation



*Linear Pattern*

For the comparison group, half of the correlations for the comparison group are below .20 and half fall in the range of .20 to .36. This indicates that for the comparison group, roughly half of the items have little if any relationship to the parent and half have a low positive correlation with the parent. The correlations for the gap groups tend to be lower than those for the comparison group. For gap 1, 9 out of the 14 correlations are below .20 and for gap 2, 13 out of

the 14 are below .20. This indicates that for the gap groups there is little if any relationship between the parent and child items. The percent consistent performance for the comparison group ranges from a low of 62.4 to a high of 74.2. For the gap groups, the percent consistent performance tends to be lower than the comparison group, ranging from 44.4 to 69.5 for gap 1 and from 43.5 to 64.0 for gap 2.

Table 2: Linear Pattern Family

Correlation with Parent and Percent Consistent Performance

	Comparison		Gap 1		Gap 2	
	$\phi_p$	%CP	$\phi_p$	%CP	$\phi_p$	%CP
Sibling	.21	67.6	.24	62.9	-.02	52.8
Child 1	.34	73.3	.33	66.0	.24	64.0
Child 2	.25	70.2	.19	60.2	.12	59.2
Child 3	.20	70.1	.05	52.2	-.05	47.5
Child 4	.36	74.2	.39	69.5	.18	60.8
Child 5	.16	66.5	.08	54.3	.08	57.3
Child 6	.23	71.2	.25	60.9	.16	57.3
Child 7	.20	69.7	.13	56.1	.06	55.6
Child 8	.25	71.2	.20	60.5	.08	57.1
Child 9	.06	68.2	-.05	44.4	.00	43.5
Child 10	.15	69.2	-.04	46.6	.14	56.7
Child 11	.13	67.3	.18	59.1	.10	56.9
Child 12	.07	62.4	-.07	46.6	-.03	49.0
Child 13	.14	65.8	.12	57.4	-.08	49.7

%CP = Percent Consistent Performance

### *Evaluating an Expression*

For the comparison group, one correlation was very low at .05 (child 6) while the others range from a low of .23 to a high of .40. In other words, for the comparison group most items have a low positive correlation with the parent. The correlations for the gap groups tend to be lower than those for the comparison group. For gap 1, 3 out of the 7 correlations are below .20 and for gap 2, 4 out of the 7 are below .20. The remaining correlations range from .24 to .44 for gap 1 and from .22 to .38 for gap 2. The percent consistent performance for the comparison

group is quite high ranging from 84.2 to 88.6. The percent consistent performance is a lot lower for the gap groups ranging from 48.3 to 72.7 for gap 1 and from 53.2 and 70.4 for gap 2.

Table 3: Evaluating an Expression Family

Correlation with Parent and Percent Consistent Performance

	Comparison		Gap 1		Gap 2	
	$\phi_p$	%CP	$\phi_p$	%CP	$\phi_p$	%CP
Sibling	.23	84.2	.14	55.0	.19	59.6
Child 1	.39	88.4	.33	62.7	.38	68.2
Child 2	.39	88.6	.30	60.7	.20	59.0
Child 3	.26	87.2	.24	58.4	.14	53.2
Child 4	.40	88.4	.44	72.7	.31	70.4
Child 5	.27	87.1	.15	51.7	.22	56.2
Child 6	.05	84.2	.03	48.3	.14	53.8

%CP = Percent Consistent Performance

*Rate of Change*

For the comparison group, all but one of the correlations (.25) were very low. In other words, for the comparison group most of the items have little if any correlation with the parent. The correlations are also low for the gap groups ranging from -.07 to .22 for gap 1 and from .12 to .23 for gap 2. The percent consistent performance is moderately high for the comparison group ranging from 77.8 to 80.8. The percent consistent performance is a lot lower for the gap groups ranging from 42.4 to 56.1 for gap 1 and from 44.7 and 58.1 for gap 2.

*Equality*

For the comparison group, 8 out of the 10 correlations are 2.0 or lower indicating that there is very little if any relationship between the parent and child items. The correlations are also low for the gap groups with 7 out of 10 correlations below 2.0 for gap 1 and 8 out of 10 below 2.0 for gap 2. The percent consistent performance for the comparison group ranges from 68.1 to 77.5. The percent consistent performance tends to be lower for the gap groups ranging from 44.4 to 76.6 for gap 1 and from 41.8 and 66.4 for gap 2.

Table 4: Rate of Change Family

Correlation with Parent and Percent Consistent Performance

	Comparison		Gap 1		Gap 2	
	$\phi_p$	%CP	$\phi_p$	%CP	$\phi_p$	%CP
Sibling	.13	80.7	.10	45.7	.12	44.7
Child 1	.11	78.9	.13	53.2	.14	50.3
Child 2	.10	79.9	-.07	42.4	.15	52.0
Child 3	.25	80.8	.22	56.1	.22	57.2
Child 4	.10	79.7	.22	51	.15	48.2
Child 5	.14	77.8	.12	54.4	.19	58.1
Child 6	.17	78.4	.16	50.0	.23	53.1

%CP = Percent Consistent Performance

Table 5: Equality Family

Correlation with Parent and Percent Consistent Performance

	Comparison		Gap 1		Gap 2	
	$\phi_p$	%CP	$\phi_p$	%CP	$\phi_p$	%CP
Sibling	.16	72.1	.12	59.1	-.09	53.1
Child 1	.30	74.7	.06	56.2	.18	65.3
Child 2	.10	74.5	.14	50.0	.23	52.0
Child 3	.20	68.1	-.07	50.0	.23	63.9
Child 4	.07	74.5	.21	55.6	-.06	41.8
Child 5	.02	70.6	.05	51.1	-.05	49.5
Child 6	*	*	*	*	*	*
Child 7	.12	73.0	.21	56.7	.05	50.2
Child 8	.38	77.5	.30	66.6	.19	62.9
Child 9	.19	69.7	.04	55.5	.15	66.4
Child 10	.09	71.1	-.03	44.4	.03	48.4

\*Item removed from analysis.

%CP = Percent Consistent Performance

Overall, the findings indicate that a majority of comparison group students perform consistently across parent and child items. While for some items a majority of gap students perform consistently, the percentage of gap students who perform consistently is typically smaller than the percentage of comparison students who perform consistently. Many items lead

to inconsistent performance for gap 1 and gap 2 students and in some cases, less than 50% of the gap students perform consistently across the two items.

## Discussion

The purpose of the study was to assess the quality and usefulness of test items designed to decompose skills and knowledge required to solve complex algebraic problems. More specifically, the study sought to answer two research questions: (1) To what extent do students who perform well on the complex items also perform well on the decomposed items; and (2) To what extent are students in an assessment gap able to succeed on decomposed items while struggling with the complex (unadjusted) items?

Our initial analyses examined whether the items in each family performed as expected. A factor analysis revealed that the items clustered together by family and reliability analysis showed moderate to high internal consistency with reliability coefficients of .84, .84, .73 and .68 for the four families, and .92 for the test as a whole. As expected, most of the child items were easier than their parent. We expected that within each family the parent and sibling items would have roughly the same item difficulty, but our analyses revealed that the siblings were easier.

Our comparison of students who did and did not perform well on the parent items showed that students who answered a parent item correctly also tended to answer the child items in that family correctly while students who answered a parent item incorrectly tended to be less consistent in their performance on the child items. Our analysis showed that students who got the parent item incorrect tended to also get the difficult child items incorrect, but were able to answer some of the easier child items correctly. Their success on some of the child items suggests that students who get the parent item incorrect may have some of the skills necessary to solve the more complex parent problem. It is interesting to note that while the below grade level

items appear to have reduced the gap in performance between the students who answered the parent correctly and those who did not, some of these items do not appear to have been more effective in doing so than the on grade level items.

Below grade level items appear to be somewhat effective at closing the gap in performance between low- and high-scoring students (gap and comparison groups) for certain content strands. In the equality and evaluating an expression families, below grade level items appear to reduce the gap in performance more than the on grade level items. On the other hand, the two below grade level items in the rate of change family do not appear to reduce the gap in performance as much as the on grade level items. In the linear pattern family, the results are mixed; some of the below grade level items appear to reduce the gap in performance while others do not.

Our analysis of gap and comparison group students' performance on the child items yielded results that provide some support for the idea that modifications to parent items may enable some students in the gap to solve them. The results vary by content strand. In the equality family, the two items that reduce the gap in performance the most differ from the parent in that they are single-step rather than multi-step problems. In the evaluating an expression family, the parent item presented a problem situation where the student was required to solve an equation involving two variables. Performance on the child items suggests that removing the problem situation does not reduce the gap in performance but suggests that simplifying the problem (by using whole numbers or variables equal to 1) and having the student demonstrate understanding of algebraic expressions without requiring them to solve an equation might. In the rate of change family, all of the child items appear to reduce the gap in performance between comparison and gap students but the smallest difference occurs on items where the rate information is not

presented in a table; it is presented in graphs or as a sequence of numbers without a problem situation. In the linear pattern family, the impact of the modification is not as clear. In this family, the two items that appear to reduce the gap are items that use the same pattern of numbers as two other child items. It is possible that the smaller difference in performance between the gap and comparison students is the result of modifications in the presentation of the information or perhaps it is the result of students having had two other opportunities to work with the same pattern of numbers (a practice effect).

Across the four families, there were eight items that differed from the parent item in that the problem situation was removed whereby reducing the linguistic complexity of the item. When these items were examined separately, we found that three of them were equally as difficult for students in the gap as the corresponding parent items. The remaining five items were easier for gap students than were the parent items but because some of these items are also below grade level we do not know which of the modifications is having a greater impact on item difficulty.

When the below grade level items from all four families were examined, we found that a majority of the items were easier than the corresponding parent items for students in the gap. However, in a few cases, students in the gap found the below grade level items to be more difficult than the parent.

In summary, this study represented a first step in exploring features of items that might be manipulated to provide valid indicators of whether students are reaching grade level expectations while also informing instruction for individual students. While no clear patterns emerged, removing the context of items, often presented as word problems, did seem to close the gap in performance for several items. Similarly, simplifying the problems presented appeared to enable

some students in the gap to solve them correctly. In many cases, however, simplifying the problem resulted in items that were considered below grade level. While below grade level items may help teachers identify the skills and knowledge that students in the gap, particularly gap 2, are able to apply in order to solve problems correctly, they do not provide information about how students are performing relative to the grade level expectations. Perhaps most importantly, these findings suggest that the dual goal of measuring student achievement relative to grade level expectations and providing teachers with information about what students can and cannot do might be accomplished through a modular test design that employed “parent” items to measure student achievement relative to grade level achievements and “child” items to measure component skills required to answer accurately complex parent items.

## References

Bechar, S. & Godin, K. (2007). Finding the real assessment gaps: A process for states to identify gaps in their assessment systems. A report prepared for the New England Compact. Accessed July 20, 2008 at: [http://www.necomact.org/Reading\\_students\\_in\\_the\\_gaps.pdf](http://www.necomact.org/Reading_students_in_the_gaps.pdf).

Famularo, L. & Russell, M. (2007). Examining the utility of a prototype assessment for assessing students in the gap. A report prepared for the New England Compact. Accessed July 20, 2008 at: [http://www.necomact.org/Reading\\_students\\_in\\_the\\_gaps.pdf](http://www.necomact.org/Reading_students_in_the_gaps.pdf).

New England Compact. (2007). Reaching students in the gap: A study of assessment gaps, students, and alternatives: Executive Summary. Accessed July 20, 2008 at: [http://www.necomact.org/full\\_research.htm](http://www.necomact.org/full_research.htm).

Parker, C. & Saxon, S. (2007a). Teacher views of students in the gaps. A report prepared for the New England Compact. Accessed July 20, 2008 at: [http://www.necomact.org/Reading\\_students\\_in\\_the\\_gaps.pdf](http://www.necomact.org/Reading_students_in_the_gaps.pdf).

Parker, C. & Saxon, S. (2007b). "They come to the test, and there is nothing to fold": Teacher views of large-scale assessments and classroom context. A report prepared for the New England Compact. Accessed July 20, 2008 at: [http://www.necomact.org/Reading\\_students\\_in\\_the\\_gaps.pdf](http://www.necomact.org/Reading_students_in_the_gaps.pdf).