2004 Colorado Student Assessment Program Alternate (CSAPA):

Interrater Reliability and Administration Fidelity

Grade 5 Math and Grades 7 & 10 Literacy

Lindy Crawford

Marilee McDonald

University of Colorado at Colorado Springs

August 23, 2005

2004 CSAPA: Interrater Reliability and Administration Fidelity
Grade 5 Math and Grades 7 & 10 Literacy

This report contains an analysis of interrater reliability and administration fidelity data related to the 2004 administration of the CSAPA for fifth grade math, and seventh- and tenth grade literacy. We viewed a collection of videotaped administrations of the assessments in order to collect data on the following: (1) teachers' fidelity, or accuracy, of administration, (2) interrater agreement related to students' performance on each of the questions comprising the three tests at three different grade levels, and (3) the clarity of directions and prompting provided teachers in the administration manual and supporting materials. We share the results of both quantitative and qualitative analyses, and where appropriate, include specific recommendations that may act to strengthen the overall reliability of the assessment.

Methodology

Dataset. Two primary data sources were used for this technical report. One set of data was provided by the Colorado Department of Education (CDE), and the second set was created by the authors. After combining pertinent information from both sources, we created a data file that included student demographics, grade level, subject area, qualitative and quantitative data related to administration fidelity, score on each indicator, and total score.

Discrepancies existed in the total number of students included in each of the two datasets at grades 5 and 7 as reported in the table below. For the reliability analysis we used only those videotapes included in both datasets. For the administration fidelity analyses we were able to use one additional tape at grade 5 that was not included in the CDE dataset.

Table 1
Number of Students in Each Dataset

	Grade 5	Grade 7	Grade 10
CDE	21	24	13
Authors	19	17	13
Shared	18	17	13

Procedures. We evaluated each videotaped administration of the CSAPA on the following dimensions: (1) student accuracy on each indicator, (2) teachers' use of accurate prompting as outlined by the scoring rubric, (3) teachers' provision of directions as outlined in the administration manual, (4) teachers' use of wait time, (5) teachers' appropriate set-up of task, and (6) teacher behaviors as recorded qualitatively. Videotapes consisted of students completing the Math (grade 5), or the Literacy assessment (grades 7 and 10). The Math assessment consisted of 74 indicators; the Literacy assessment consisted of 52 indicators at seventh grade and 50 indicators at tenth grade. On occasion teachers did not videotape portions of the test, thus we were unable to score those indicators. We also did not score indicators for which student responses were unclear.

Data Analysis

We analyzed interrater agreement by calculating a percent-agreement figure. For each indicator, we compared the score assigned by the teacher and by the researchers. More specifically, at grade 5, each indicator was scored 18 times (number of students represented in both datasets) for level of agreement using the following scale: (1) Score was an exact match between both raters, (2) Score differed by one point (on a 5-point scale), (3) Score differed by two points, (4) Score differed by three points, (5) Score differed by four points, or (6) No Score (NS) recorded on one of the two datasets. Totals were calculated and percentage of agreement for all indicators recorded. Identical procedures were followed at grades 7 and 10. Each indicator was scored 17 times at grade 7 and 13 times at grade 10.

We calculated intraclass correlations (ICCs) in order to provide a statistical measure of the degree of interrater reliability for each test. Intraclass correlations are often used as reliability coefficients among evaluations of items deemed to be in the same category or class. They compare the covariance of the ratings with the total variance of the items.

We also analyzed the accuracy of teacher set-up, prompting, use of directions, and wait time by calculating a percentage correct score for each test. We used our dataset to perform this analysis as the CDE dataset did not contain data on the accuracy of administration. And finally, we conducted a qualitative analysis of teachers' fidelity of test administration, analyzing our notes for themes related to teachers' administration practices.

Results for Fifth Grade Math Assessment

Interrater Agreement

The level of interrater agreement for *each* indicator at grade 5 is presented in Table 2. Patterns can be detected across those indicators that had the highest levels of agreement and those indicators that had the lowest levels of agreement. One way to interpret score consistency across indicators is to analyze the scores from the "Exact Match" column. Each indicator was scored 18 times. In Table 2 we have highlighted those indicators with less than 10 of 18 exact matches as a rough estimate of how well each indicator fared. Note, however, that many of the latter indicators were not scored by one of the parties, thus resulting in a lower "exact match" score.

Table 3 provides the percentage of agreement across *all* 74 indicators in the fifth grade Math assessment. These percentages are calculated by dividing the number recorded in each category by the total number of responses (n = 1160). The total number of possible responses was calculated by multiplying 18 cases by 74 indicators for a total of 1,332 responses minus the number of times that agreement could not be calculated due to missing data (n = 172). As Table 3 illustrates, a strong level of agreement was reached. Raters were in exact agreement, or disagreed by only 1 point, on 94% of their 1160 scoring opportunities.

Table 2
Fifth Grade Math Assessment: Counts of Interrater Agreement

Indicator	Exact	Off by 1	Off by 2	Off by 3	Off by 4	NS
1	16	2	0	0	0	0
2	17	0	1	0	0	0
3	14	3	1	0	0	0
4	17	1	0	0	0	0
5	17	1	0	0	0	0
6	14	3	0	1	0	0
7	13	4	0	1	0	0
8	17	1	0	0	0	0
9	14	4	0	0	0	0
10	15	2	1	0	0	0
11	13	5	0	0	0	0
12	12	5	1	0	0	0
13	9	8	1	0	0	0
14	10	4	0	1	0	3
15	13	3	0	0	0	2
16	17	0	0	0	0	1
17	13	1	1	1	1	1
18	11	4	0	0	0	3
19	15	1	0	0	0	2
20	13	3	0	0	0	2
21	15	1	0	0	0	2
22	7	6	1	0	0	4
23	11	5	0	0	0	2
24	12	4	0	0	0	2
25	12	3	0	1	0	2
26	10	5	1	0	0	2
27	7	7	1	0	0	3
28	13	2	0	1	0	2
29	10	2	1	2	1	2
30	11	4	0	0	1	2
31	11	3	1	0	0	3
32	10	6	0	0	0	2
33	10	3	0	1	0	4
34	14	1	1	0	0	2
35	9	6	1	0	0	2

Table 2 cont.

Fifth Grade Math Assessment: Counts of Interrater Agreement

Indicator	Exact	Off by 1	Off by 2	Off by 3	Off by 4	NS
36	8	5	1	1	0	3
37	6	9	0	1	0	2
38	13	2	1	0	0	2
39	16	1	0	0	0	1
40	17	0	0	0	0	1
41	13	2	1	0	0	2
42	13	2	1	0	0	2
43	13	2	1	0	0	2
44	13	3	0	0	0	2
45	12	3	0	1	0	2
46	13	1	0	0	0	4
47	11	2	1	0	0	4
48	11	4	1	0	0	2
49	11	4	1	0	0	2
50	12	2	2	0	0	2
51	11	4	0	0	0	3
52	10	4	2	0	0	2
53	10	4	1	0	1	2
54	7	6	1	1	0	3
55	8	6	0	1	1	2
56	9	6	1	0	0	2
57	9	5	0	0	0	4
58	7	6	0	0	2	3
59	9	3	1	0	2	3
60	9	2	2	0	2	3
61	10	2	2	1	0	3
62	8	5	2	0	0	3
63	6	6	1	1	0	4
64	9	4	1	0	0	4
65	10	1	2	0	0	5
66	8	3	1	0	1	5
67	6	5	1	0	0	6
68	10	3	0	1	0	4
69	9	3	0	1	0	5
70	9	4	1	0	0	4
71	6	5	1	0	0	6
72	10	3	1	0	0	4
73	10	4	0	0	0	4
74	7	2	3	1	0	5
Total	**831**	**251**	**47**	**19**	**12**	**172**

* n = 18

Table 3
Fifth Grade Math Assessment: Percentage of Interrater Agreement

	Exact	Off by 1	Off by 2	Off by 3	Off by 4	Total
Count	831	251	47	19	12	1160
Percentage	72%	22%	4%	2%	1%	101%

* Total does not equal 100% due to calculation of percentages within each category.

Interrater Reliability

The correlation coefficient for interrater reliability at grade 5 was calculated at ICC(1158) = .91. This model assumes an ordinal scale and a confidence level of 95%.

Administration Fidelity

Table 4 represents the accuracy of teacher behaviors on the fifth grade Math assessment. Teachers were least accurate in their prompting behaviors (i.e., following the sequence of prompts as denoted in administration manual) at 83% accuracy across all 74 indicators. Teachers were most accurate in gaining students' attention and providing students with wait time on each indicator.

Table 4
Fifth Grade Math: Administration Fidelity

Teacher Behavior	Total Errors	Total Trials	% Incorrect	% Correct
Set-Up	42	1273	3%	97%
Gain Attention	0	1273	0%	100%
Prompting	217	1273	17%	83%
Follow Directions	146	1273	11%	89%
Wait Time	6	1273	1%	99%

Results for Seventh Grade Literacy Assessment

Interrater Agreement

Percentage of interrater agreement for grade 7 was calculated in an identical manner as agreement at grade 5 and is reported in Tables 5 and 6. In Table 5, “exact matches” were highlighted across the 17 indicators, using the same “less than 10” criteria as used at grade 5. Again, it is important to note that some of the latter indicators were not scored by one of the parties, thus resulting in a lower “exact match” score.

In Table 6, the total number of possible responses was calculated by multiplying 17 cases by 52 indicators for a total of 884 responses minus the number of times that agreement could not be calculated due to missing data (n = 130). As was apparent in fifth grade, a strong level of agreement was reached. Raters were in exact agreement, or disagreed by only 1 point, on 93% of their 754 scoring opportunities.

Table 5
Seventh Grade Literacy Assessment: Counts of Interrater Agreement

Indicator	Exact	Off by 1	Off by 2	Off by 3	Off by 4	NS
1	13	2	0	0	0	2
2	13	3	0	0	0	1
3	15	1	0	0	0	1
4	15	1	0	0	0	1
5	12	2	0	1	0	2
6	13	1	0	0	0	3
7	10	1	2	1	0	3
8	8	5	1	0	0	3
9	11	1	1	1	0	3
10	8	3	1	2	0	3
11	11	1	1	1	0	3
12	9	5	0	0	0	3
13	9	6	0	0	0	2
14	13	2	0	0	0	2
15	12	3	0	0	0	2
16	11	3	1	0	0	2
17	12	2	0	0	0	3
18	13	1	1	0	0	2

Table 5 cont.

Seventh Grade Literacy Assessment: Counts of Interrater Agreement

Indicator	Exact	Off by 1	Off by 2	Off by 3	Off by 4	NS
19	9	3	3	0	0	2
20	8	6	1	0	0	2
21	6	7	1	1	0	2
22	13	3	0	1	0	0
23	14	3	0	0	0	0
24	8	7	1	1	0	0
25	13	1	0	3	0	0
26	12	1	0	3	0	1
27	13	4	0	0	0	0
28	11	5	0	0	0	1
29	11	4	0	1	0	1
30	13	1	1	0	0	2
31	11	2	1	0	0	3
32	4	9	0	0	0	4
33	8	6	0	0	0	3
34	9	3	1	0	0	4
35	14	2	0	0	0	1
36	10	4	1	0	0	2
37	16	0	0	0	0	1
38	13	1	1	1	0	1
39	14	2	0	0	0	1
40	9	1	0	0	1	6
41	12	4	0	0	0	1
42	10	1	1	1	1	3
43	4	2	0	0	0	11
44	9	3	1	1	1	2
45	9	4	1	1	1	1
46	8	4	1	1	0	3
47	4	3	0	0	0	10
48	8	6	0	0	0	3
49	7	4	2	0	0	4
50	8	4	0	1	0	4
51	10	3	0	0	0	4
52	8	2	1	0	0	6
Total	**544**	**158**	**26**	**22**	**4**	**130**

*n = 17

Table 6
Seventh Grade Literacy Assessment: Percentage of Interrater Agreement

	Exact	Off by 1	Off by 2	Off by 3	Off by 4	Total
Count	544	158	26	22	4	754
Percentage	72%	21%	3%	3%	1%	100%

Interrater Reliability

The correlation coefficient for interrater reliability at grade 7 was calculated at ICC(753) = .92. This model assumes an ordinal scale and a confidence level of 95%.

Administration Fidelity

Table 7 represents the accuracy of teacher behaviors on the seventh grade Literacy assessment. Teachers prompted accurately (i.e., followed the sequence of prompts as denoted in administration manual) 88% of the time across the 52 indicators. Teachers were 100% accurate in gaining attention of student before presentation of tasks and in their use of wait time.

Table 7
Seventh Grade Literacy: Administration Fidelity

Teacher Behavior	**Total Errors**	**Total Trials**	**% Incorrect**	**% Correct**
Set-Up	11	754	1%	99%
Gain Attention	1	754	0%	100%
Sequence	88	754	12%	88%
Follow Directions	42	754	6%	94%
Wait Time	0	754	0%	100%

Results for Tenth Grade Literacy Assessment

Interrater Agreement

Interrater agreement for grade 10 was calculated in the same manner as agreement at grades 5 and 7 and is reported in Tables 8 and 9 below. At grade 10 there were only 13 cases; therefore, a "high" level of "exact match" was roughly estimated at nine versus 10 and every indicator with less than nine "exact matches" is highlighted in Table 8.

In Table 9, the total number of possible responses was calculated by multiplying 13 cases by 50 indicators for a total of 650 responses minus the number of times that agreement could not be calculated due to missing data (n = 28). Raters were in exact agreement, or disagreed by only 1 point, on 96% of their 622 scoring opportunities.

Table 8
Tenth Grade Literacy Assessment: Counts of Interrater Agreement

Indicator	Exact	Off by 1	Off by 2	Off by 3	Off by 4	NS
1	8	4	0	1	0	0
2	9	3	0	1	0	0
3	12	0	1	0	0	0
4	10	3	0	0	0	0
5	10	3	0	0	0	0
6	11	1	0	1	0	0
7	9	3	0	1	0	0
8	12	1	0	0	0	0
9	12	1	0	0	0	0
10	11	1	1	0	0	0
11	12	1	0	0	0	0
12	11	0	2	0	0	0
13	10	2	0	1	0	0
14	7	3	3	0	0	0
15	10	3	0	0	0	0
16	10	3	0	0	0	0
17	11	2	0	0	0	0
18	10	3	0	0	0	0
19	13	0	0	0	0	0
20	10	2	0	0	1	0
21	11	2	0	0	0	0
22	13	0	0	0	0	0

Table 8 cont.

Tenth Grade Literacy Assessment: Counts of Interrater Agreement

Indicator	Exact	Off by 1	Off by 2	Off by 3	Off by 4	NS
23	8	4	1	0	0	0
24	12	1	0	0	0	0
25	11	2	0	0	0	0
26	12	1	0	0	0	0
27	12	1	0	0	0	0
28	7	3	1	0	0	2
29	7	4	0	0	0	2
30	13	0	0	0	0	0
31	12	1	0	0	0	0
32	11	2	0	0	0	0
33	12	1	0	0	0	0
34	9	3	0	0	0	1
35	8	3	1	0	0	1
36	9	3	0	0	0	1
37	7	5	1	0	0	0
38	10	2	0	0	0	1
39	10	2	0	0	0	1
40	12	0	1	0	0	0
41	11	1	0	1	0	0
42	11	1	0	0	0	1
43	6	6	0	1	0	0
44	8	2	0	1	0	2
45	10	3	0	0	0	0
46	5	0	0	0	0	8
47	9	1	0	0	0	3
48	9	2	0	1	0	1
49	10	0	1	0	0	2
50	7	4	0	0	0	2
Total	**500**	**99**	**13**	**9**	**1**	**28**

*n = 13

Table 9
Tenth Grade Literacy Assessment: Percentage of Interrater Agreement

	Exact	Off by 1	Off by 2	Off by 3	Off by 4	Total
Count	500	99	13	9	1	622
Percentage	80%	16%	2%	1%	0%	99%

* Total does not equal 100% due to calculation of percentages within each category.

Interrater Reliability

The correlation coefficient for interrater reliability at grade 10 was calculated at ICC(621) = .93. This model assumes an ordinal scale and a confidence level of 95%.

Administration Fidelity

Table 10 represents the accuracy of teacher behaviors on the tenth grade Literacy assessment. The accuracy of teacher prompting (i.e., following the sequence of prompts as denoted in administration manual) was the lowest at 86% accuracy across all 50 indicators. Teachers were much more accurate on other behaviors including gaining students' attention and use of wait time.

Table 10
Tenth Grade Literacy: Administration Fidelity

Teacher Behavior	Total Errors	Total Trials	% Incorrect	% Correct
Set-Up	37	622	6%	94%
Gain Attention	17	622	3%	97%
Sequence	90	622	14%	86%
Follow Directions	60	622	10%	90%
Wait Time	9	622	1%	99%

Qualitative Information on Test Protocols and Fidelity of Administration

Test Protocols

On occasion, critical directions are placed at the bottom of indicators in both the grade 7 and grade 10 administration manuals. Teachers often miss these directions causing confusion for both student and teacher as well as affecting student accuracy on indicator or subsequent indicators. These notes, such as directing teachers to do things like show students the "Dog Facts" in the seventh grade manual (Indicator 35) are easily missed and if the teacher has not adequately reviewed the administration manual before beginning the test, they do not provide the directions to the students. This was not perceived as a problem in fifth grade administration manual.

Fidelity of Administration

In general, one conclusion made after reviewing videotaped administrations of the CSAPA across all three grade levels is that teacher behaviors deeply impact student behaviors and thus, greatly affect a student's final score on the assessment. Although all teachers are provided with extensive training, standardized materials, and verbatim directions, much variation exists in the manner which teachers prepare for the test, administer the test, and score student responses. Some teachers were observed to give up far too easily, while others persevered and provided their students with every possible opportunity to perform tasks successfully. Some teachers repeated the Level 5 prompt up to seven times while others gave the prompt once and then immediately scored the student's answer as a "1" without giving him or her the opportunity to successfully complete the task with more scaffolding or encouragement. Most teachers were painstakingly patient and should be commended for their care of the students in their custody; whereas, a minority of teachers seemed overburdened, rushed, and irritated with the test. In the end, even with extensive directions, leveled prompts, and explicit scaffolds, teacher behaviors varied considerably.

Teachers also seemed unclear as to how many times the stimulus should be presented. On occasion, teachers would ask the same question numerous times in an attempt to get the student to answer the question. Sometimes, students would need this much repetition to answer correctly but many times students' had already provided their best answer and teachers continued the questioning. At other times, the student's answer appeared to be correct but the teacher would continue prompting or would count answer as incorrect and proceed. Multiple prompts were most apparent at levels 2 and 5.

Even though test directions were more explicit at all grade levels in the 2004 administration manuals than in the 2003 manuals, teachers still failed to consistently provide students with Level 3 prompts. On at least one occasion during most test sessions , teachers presented the indicator (Level 5 prompt), repeated it (Level 4 prompt) and then skipped to a forced choice (Level 2 prompt), or quit and moved to the next question.

Finally, teachers administering the test in as many as four sessions did not always take breaks at logical points. Sometimes when teachers resumed testing, students were required to remember the main idea of a particular cluster of questions; they rarely remembered. At times, teachers remarked that the test was taking too long and their schedules did not allow for one large block of time in which to administer the test in its entirety.

In the next three sections, observations unique to each grade level are shared and, when appropriate, suggestions are made for changes to indicators.

Grade 5

The CSAPA Math assessment at grade 5 consists of 74 questions. Many teachers were recorded as saying that the test was far too long and in our viewings we noted that many students became tired and performed less well toward the latter half of the test. Obviously, teachers could spread the testing session out over numerous short sessions and some teachers did. Other teachers, however, moved the student straight through beginning with Indicator 1 and ending with Indicator 74. The 50 and 52 questions comprising the Literacy assessments at grades 7 and 10 seemed to be a more appropriate length.

Other observations related to the fifth grade Math protocol are as follows:

Indicator 6: Students are supposed to estimate quantity but many get the answer “4” by using 1:1 counting. Teachers scored the student’s answer as correct even if they counted each individual domino as opposed to using estimation skills.

Indicator 16: This question is worded as a forced choice from the beginning – “Which has more dominoes, the game board or the box?”

Indicator 23: Indicator is unclear on Level 3 prompt. Teacher writes the number 7, student traces it. Then student is asked to write the number 9 but does the teacher write the number 9 for student to trace because it is a Level 3 prompt? Some teachers provide students with a scaffold by writing the number first and asking students to trace it, while other teachers direct students to write without the scaffold.

Indicator 27: Teachers may benefit from having more explicit directions and scaffolding for this question. Teachers often respond, “But we just did that,” because of their work on Indicator 26.

Indicator 29: According to test directions, students should say, “Thirteen plus three equals,” but students rarely say “equals” and teachers do not count it as wrong.

Indicator 35: Test question is measuring whether or not students understand the value of "sets of 10," but most teachers give students a "5" if they count each domino. On the other hand, a few teachers would not allow students to use 1:1 correspondence in order to arrive at an answer.

Indicator 42: Directions are a bit unclear. Some teachers expect students to show them all of the circles in the entire set while other teachers expect students to show them one circle. Question difficulty is based on each teacher’s understanding of directions.

Indicator 53: There is no possible correct answer. The question asks students to identify the shape represented "the most" on the graph, *but* there exists an equal number of squares and triangles.

Indicator 56: Directions are somewhat unclear. Sometimes the task is simple because teachers leave all puzzle pieces together on the table and students only have to pick them up and place them on the puzzle again. In other cases, the teacher asks students to tell them what shapes they need for the puzzle and then the teacher provides the shapes (puzzle pieces).

Indicator 57: Directions are unclear. Teachers do not understand what they are supposed to ask students to do. Often, they point to entire page of pictures as opposed to television (as suggested in parentheses). Sometimes, teachers read the directions and complete the tasks themselves instead of having students complete the tasks.

Indicator 58: Level 2 prompt does not make sense; there are no triangles on the “enviro” graph.

Indicator 60: Teachers can greatly affect student’s success on this indicator when they choose to give students blank graph paper as opposed to pre-made graph paper.

Indicator 65: Students never really understand what the question asks them to do. One suggestion is to make student directions longer and more explicit.

Indicator 66: Directions for students need to be more explicit. "Now it is your turn," does not communicate the desired outcome that student repeats pattern. Often, students create whatever pattern they want.

Indicator 67: See #65 and #66.

Indicator 69: The directions for Prompt 2 in the administration manual are incorrect because there are no green circles to use in the materials pack.

Indicator 73: Directions for Prompt 3 are incorrectly written, should be "two, four, six" not "two, three, six.”

Indicator 74: The implicit definition of a “pattern” is unclear throughout this indicator. Shouldn’t the shapes be repeated at least once in order to create a pattern? Level 2 prompt is especially unclear.

Grade 7

In the seventh grade Literacy assessment, teachers are presented with two versions of the story: *The Mistake.* Teachers are to choose one of the versions to use with students. The versions, however, are not the same length. Version A is considerably longer. Most of the teachers in this sample chose the shorter version but on occasion a teacher would choose the longer version, providing the student with more information than the amount of information provided to students exposed to Version B. Comparability of forms is thus compromised.

As mentioned previously, at grade 7 some indicators provided teachers with directions that went unnoticed because of their placement on the page and their font size and format. At grade 7 these three indicators were particularly affected:

Indicator 2: Teacher does not review all the words in the glossary as directed.

Indicator 27: Teachers rarely follow directions at bottom of page prompting them to reread the entire article to the student.

Indicator 35: Teachers sometimes miss the directions telling them to read the ”Dog Facts” from web page. One result of this is that students answer the next indicator incorrectly.

The following indicators were at times problematic for other reasons:

Indicator 9: Teachers would benefit from explicit instructions on how desk top should look. Teachers were not always ready with the picture cards; they often times arranged them on desk in the correct order: beginning, middle, and end.

Indicator 11: Teachers do not always scaffold this task correctly when they remove the incorrect cards after the Level 5 prompt leaving students with one choice or no choices and automatically reducing students' possible score to a 1 or 2.

Indicator 15: Students' responses to this indicator seem to be highly dependent on the prop teachers' use (for example, use of a burning candle for “fire,” etc.)

Indicator 29: The indicator asks, “who” is in the article? The student may choose the “dog,” or the “woman” picture cards. The test considers only the “woman” picture cards as correct for the answer to “who?” Some teachers consider the “dog” a correct response. Other teachers include the picture of the dog and the woman together and consider this a correct response. More explicit directions may help standardized scoring across different teachers.

Indicator 32: The wording of this indicator, "What does Sadie mean?" is sometimes confusing to students. Students try to describe the meaning of the word Sadie as opposed to describe Sadie's intent. Also, “Sadie” is not a common proper noun and its unfamiliarity seems to throw students off at times – in particular, its pronunciation.

Indicator 42: A common error is that teachers do not ask students to tell them about what they learned about dogs and then write about what they learned. Instead, teachers ask students to *only* write or *only* tell. Or, teachers ask them to do both and students tell one thing but write another.

Grade 10

As with grade 7, directions to teachers on some indicators went unnoticed, specifically Indicators 2, and 15-18. Other indicators posed challenges to teachers for reasons other than ambiguous directions as described below.

Indicator 1: Question One reads: "What will be the best way for you to read and understand this story?" But, teachers do not provide students with list or actual AT devices so students do not understand what is meant by the "best way to understand the story." Even at the Level 3 prompt, when teachers say, "I want you to pick or look at something that will assist you in reading this story," they still do not always provide students with choices.

Indicator 2: Often teachers do not follow directions at bottom of indicator telling them to review all vocabulary words before going on. This may impact student success on successive indicators.

Indicator 11: Teachers do not always assume correctly when deciding whether or not students would benefit from use of picture cards. Directions may need to state that teachers should use the cards until they know for sure that the student does not need them. Or, always present the cards whether student needs them or not. A different issue is that very few teachers follow directions in describing the picture cards as they are placed in front of students.

Indicator 17: Often teachers do not use the Notes Organizer, negatively affecting student responses on questions that follow. A note to teachers on Indictors 15-18 prompts them to use Notes Organizer but it is in regular font at bottom of indicator. The Notes Organizer is mentioned (used) again on Indicator 30 in bold print. Therefore, some teachers don't seem to be aware of the Notes Organizer until Indicator 30.

Indicator 32: Even though teachers are directed to place cards randomly, many do not and place the choices in correct order so students get the question right; unclear whether or not they truly understand the concept of sequence.

Conclusions

In summary, both percent-agreement scores and interrater reliability statistics strongly support the interrater reliability of the CSAPA in fifth grade Math and seventh- and tenth grade Literacy.

In terms of administration fidelity, teachers appeared to provide more consistent prompting during the 2004 than the 2003 administration. The majority of teachers, however, tended to skip scaffolding levels on at least one indicator per administration, thus resulting in prompting accuracy percentages ranging from 83-88%. As previously noted, a large degree of variation exists across the prompts teachers administer, the support devices they offer students, the fluidity of their test administration, and how quickly they “give up” on student responses to each individual indicator. These differences in fidelity of administration appear to impact the accuracy of students’ answers thus introducing construct irrelevant variance to their total test scores; variance that may be reduced with more explicit directions and increased training.