



Florida Alternate Assessment Validity Studies

2008–2009

Florida Alternate Assessment Validity Studies

2008–2009

Three studies were conducted during the 2008–09 testing year. The purpose of the studies was to contribute evidence about the validity of the Florida Alternate Assessment. Each study focused on gathering different aspects of validity evidence for the assessment scores. The studies consisted of (1) the Teacher Rating Survey, in which teachers were asked to classify students into one of the 9 performance levels (Level 1 through Level 9) and their ratings were compared to the performance levels awarded to the students based on their assessment results; (2) the Video Scoring and Administration Rating Study, in which test administrations were video-recorded and then re-scored by a second rater; and (3) the Test-Retest Reliability Study, in which students were presented all three levels (participatory, supported, and independent) of each item, regardless of how they performed on the lower levels, to see whether the administration procedures negatively biased students' scores.

Presented below are more complete descriptions of each study, as well as a summary of the results. Limitations of the studies are also described, and recommendations for future study are outlined.

I. Teacher Rating Survey

Objective

Teachers are a key resource for student evaluation as they have significant daily interaction with students and best understand each child's strengths and areas in need of academic improvement. Given the uniqueness of each teacher's evaluation criteria, the possibility exists for teachers across the state of Florida to rate student performance differently.

The Teacher Rating Survey was an online survey designed to compare a teacher's rating of daily student performance with the performance level achieved by the student on the 2009 Florida Alternate Assessment. Teachers were asked to first review three sets of descriptors and choose the one that best fit the student based on the student's daily instruction. The descriptors were grade and content specific (e.g., grade 3 teachers were asked to rate mathematics and reading, while grade 8 teachers were asked to rate mathematics, reading, science, and writing). Essentially, the descriptors represent the three levels of complexity (participatory, supported, and independent) but were not labeled as such to avoid bias. Teachers were then asked to use indicators of student performance (such as student work products, performance during classroom activities, IEP progress reports, teacher observation, data charts, and classroom assessments) to rate the student's performance on content-specific skill sets analogous to three performance levels (basic, proficient, and advanced, which were the labels previously used for reporting) within the level of complexity descriptor chosen for the student. The teacher would rate the student as being able to demonstrate less than 50%, between 50% and 75%, or more than 75% of the skills presented. Performance levels were labeled as percentages to avoid bias. Input collected from teachers was then compared with student performance in each content area (reading, mathematics, writing, and science) tested within a particular grade. Approximately 27,000 students across grades 3–11 took the 2009 Florida Alternate Assessment. Teachers were encouraged to complete the survey for all students.

Design

The online survey was a cost-effective and easily accessible way for teachers throughout the state of Florida to provide information about their students. The survey was designed to be quick and easy, given that teachers were requested to provide input for more than one student. Teachers provided

feedback for each student based on a unique identifier. The site was set up so that student data could be accessed by a series of drop-down lists, from which teachers selected the school district, school name, and grade, and then entered the student's first and last names. Fields for the student's ID and date of birth were automatically populated so that the teacher could confirm he/she had selected the appropriate student. The survey rating process (outlined previously) progressed from that point on. Participation rates for the study, by grade and content area, are shown in Table I-1.

Table I-1. Teacher Rating Study Participation Rates by Grade and Content Area

	Mathematics		Reading		Science		Writing	
	Study N	Percent of tested	Study N	Percent of tested	Study N	Percent of tested	Study N	Percent of tested
Grade 3	253	10.8	253	10.7				
Grade 4	279	12.5	278	12.4			277	12.4
Grade 5	271	11.7	274	11.8	270	11.9		
Grade 6	306	13.2	307	13.3				
Grade 7	259	11.4	260	11.4				
Grade 8	322	12.3	321	12.3	320	12.3	320	12.4
Grade 9	323	12.8	320	12.7				
Grade 10	420	15.2	419	15.2			416	15.4
Grade 11					415	14.7		
Total	2,433	12.5	2,432	12.5	1,005	13.1	1,013	13.5

The survey contained explicit directions and evaluation criteria for teachers to use in rating student performance. Directions and evaluation criteria were drawn up by Measured Progress with input and final approval from the Florida Department of Education (FLDOE). Teachers were instructed to look at a variety of student-related information, such as daily instruction, student progress reports, and grades, to support the student performance rating.

Measured Progress prepared and sent an informational letter approved by the FLDOE to school administrators in fall 2008 outlining the goals, expectations, timing, and resources required of teachers to participate in the Teacher Rating Survey. A similar letter was sent to special education district coordinators and teachers, outlining the type of information and required resources for participation in the survey. This survey was conducted in December 2008. Teachers were given two weeks to complete the survey.

Analysis

Ratings provided by the teachers were compared with actual student performance on the 2009 assessment. Specific comparisons included correlation analysis and computation of the percentages of exact and adjacent agreement. Given that both classifications have degrees of uncertainty, the proportion of agreement may be inflated by chance agreement. To adjust for agreement by chance, kappa statistics were also calculated. Kappa can be thought of as the chance-corrected proportional agreement, and possible values range from +1 (perfect agreement) via 0 (no agreement above that expected by chance) to -1 (complete disagreement).

Results

Table I-2 below compares the percentages of students classified into each performance level based on the assessment results and the teacher ratings. For mathematics, for example, 11.7% of students were categorized as Level 1 according to their score on the assessment, while 29.5% were categorized as Level 1 according to their teacher's rating. Conversely, while 4.6% of students fell into Level 9 according to the mathematics assessment, only 2.4% received a rating of Level 9 by their teachers. This pattern is

consistent across the four content areas: generally speaking, teachers tended to rate students lower than indicated by the scores on the assessment. Finally, a closer look at Table I-2 shows that teachers were noticeably less likely to rate students at Level 3, 6, or 9 than at any of the other levels. This pattern is likely an artifact of the scoring system that was in use at the time that teachers completed the survey. Specifically, teachers were asked to rate student performance using descriptors of content-specific skill sets analogous to three performance levels (basic, proficient, and advanced) within the level of complexity at which instruction was occurring for the student. The results of the analyses indicate that, overall, teachers used a high standard for categorizing students as advanced regardless of the level of complexity.

**Table I-2. Comparison of Assessment Results and Teachers' Ratings
Percent of Students Classified into Each Performance Level**

	Math (N = 2433)		Reading (N = 2432)		Science (N = 1005)		Writing (N = 1013)	
	Actual	Teacher	Actual	Teacher	Actual	Teacher	Actual	Teacher
Level 1	11.7	29.5	11.8	31.5	10.5	33.9	10.0	33.1
Level 2	10.3	11.7	9.7	13.2	8.4	12.6	8.7	10.0
Level 3	12.7	3.8	12.1	4.6	14.8	3.9	20.6	3.5
Level 4	11.3	18.7	7.6	14.3	9.6	12.9	5.5	17.5
Level 5	17.4	14.6	10.4	14.3	13.0	14.2	12.1	11.4
Level 6	10.2	2.1	9.3	1.8	16.3	1.5	10.3	1.8
Level 7	9.5	8.6	11.1	7.2	8.2	7.3	10.9	12.3
Level 8	12.4	8.6	13.7	10.0	8.7	8.5	10.9	7.8
Level 9	4.6	2.4	14.3	3.3	10.6	5.2	11.1	2.8

Tables I-3 through I-6 show summary statistics (mean and standard deviation) of the assessment results vs. teachers' ratings as well as the correlation coefficient between the two sets of ratings. These tables also clearly show that teachers' ratings are lower than the assessment results: overall, across all grades and content areas, teachers tended to award ratings 1 to 2 performance levels lower than those obtained on the assessment. Correlations between the two sets of ratings are moderate, ranging from a low of 0.45 for grade 11 science to a high of 0.70 for grade 5 mathematics.

**Table I-3. Summary Statistics of Assessment Results and Teachers' Ratings
Mathematics**

	N	Actual		Teacher Classification		Correlation
		Mean	SD	Mean	SD	
Grade 3	253	4.2	2.4	3.0	2.2	0.64
Grade 4	279	5.2	2.4	3.8	2.3	0.65
Grade 5	271	4.3	2.3	3.7	2.5	0.70
Grade 6	306	4.7	2.5	3.7	2.4	0.64
Grade 7	259	4.8	2.3	3.8	2.3	0.56
Grade 8	322	4.6	2.3	3.9	2.5	0.63
Grade 9	323	4.9	2.3	3.9	2.6	0.55
Grade 10	420	4.7	2.3	3.9	2.7	0.62
Overall	2433	4.7	2.4	3.8	2.5	0.62

Table I-4. Summary Statistics of Assessment Results and Teachers' Ratings
Reading

	N	Actual		Teacher Classification		Correlation
		Mean	SD	Mean	SD	
Grade 3	253	4.7	2.9	3.0	2.3	0.67
Grade 4	278	5.3	2.6	3.9	2.5	0.60
Grade 5	274	4.8	2.7	3.6	2.5	0.65
Grade 6	307	4.9	2.7	3.9	2.6	0.63
Grade 7	260	5.5	2.6	3.8	2.5	0.55
Grade 8	321	5.2	2.7	4.0	2.6	0.59
Grade 9	320	5.7	2.7	3.6	2.7	0.53
Grade 10	419	5.5	2.7	3.7	2.7	0.53
Overall	2432	5.2	2.7	3.7	2.6	0.58

Table I-5. Summary Statistics of Assessment Results and Teachers' Ratings
Science

	N	Actual		Teacher Classification		Correlation
		Mean	SD	Mean	SD	
Grade 5	270	5.3	2.8	3.5	2.5	0.66
Grade 8	320	4.6	2.3	4.1	2.7	0.61
Grade 11	415	5.0	2.4	3.5	2.7	0.45
Overall	1005	4.9	2.5	3.7	2.6	0.54

Table I-6. Summary Statistics of Assessment Results and Teachers' Ratings
Writing

	N	Actual		Teacher Classification		Correlation
		Mean	SD	Mean	SD	
Grade 4	277	5.2	2.5	3.6	2.4	0.62
Grade 8	320	5.1	2.7	3.8	2.5	0.59
Grade 10	416	4.7	2.5	3.7	2.7	0.58
Overall	1013	5.0	2.6	3.7	2.6	0.59

Table I-7 shows kappa coefficients and percentages of exact and exact or adjacent agreement. As mentioned previously, kappa coefficients are a measure of proportional agreement corrected for the amount of agreement that can be expected based on chance. The kappas are low, ranging from 0.06 for grade 4 writing to 0.18 for grades 5 and 8 mathematics.

Table I-7. Kappa Coefficients and Percentages of Exact and Exact or Adjacent Agreement

Grade	Content Area	Kappa	<i>N</i>	Percent Exact	Percent Exact or Adjacent
03	Mathematics	0.14	253	26.1	52.2
04		0.11	279	21.1	48.0
05		0.18	271	28.4	63.5
06		0.17	306	27.1	52.9
07		0.13	259	23.2	51.0
08		0.18	322	28.3	56.5
09		0.14	323	23.8	47.7
10		0.11	420	21.4	53.1
03	Reading	0.12	253	22.9	49.4
04		0.10	278	18.7	45.3
05		0.16	274	25.9	52.9
06		0.16	307	26.1	51.1
07		0.10	260	20.8	42.7
08		0.13	321	22.4	49.8
09		0.14	320	22.8	44.4
10		0.09	419	18.9	40.1
05	Science	0.12	270	21.1	45.2
08		0.12	320	21.6	56.6
11		0.08	415	16.6	41.0
04	Writing	0.06	277	14.8	41.2
08		0.12	320	21.3	49.4
10		0.10	416	18.0	45.2

Figures I-1 through I-4 provide a visual representation of the relationship between teacher ratings and assessment results. The figures show the distribution of teacher ratings within each test performance level. In Figure I-1, for example, the sizes of the boxes in the column for Level 1 indicate that the vast majority of students who received a Level 1 according to their test score were also rated as a Level 1 by their teacher. Similarly, very few of the students who received a Level 1 test score were assigned a level other than Level 1 by their teacher. Although students who received a Level 9 according to their test score tended to be assigned higher performance levels by their teachers, the teachers tended to show less agreement with the assessment score for these students than they did for the lower performing students.

The figures show the same patterns described below:

- Teacher ratings tended to be lower than the assessment-assigned levels. If one envisions a line running diagonally from the bottom left square to the top right square (i.e., through the boxes that represent students who received the same score from both the assessment and the teacher rating), the boxes below the line, in aggregate, consistently represent a greater proportion of the students than the boxes above the line.
- Teachers were noticeably less likely to rate students at levels 3, 6, or 9 (looking across the figure at the rows corresponding to those three levels) than at the remaining levels.
- The teachers' ratings agreed much more closely with the assessment scores for students who scored in the lowest performance levels on the assessment; for students who scored in the middle and higher levels according to the assessment, the teachers' ratings were much more variable.

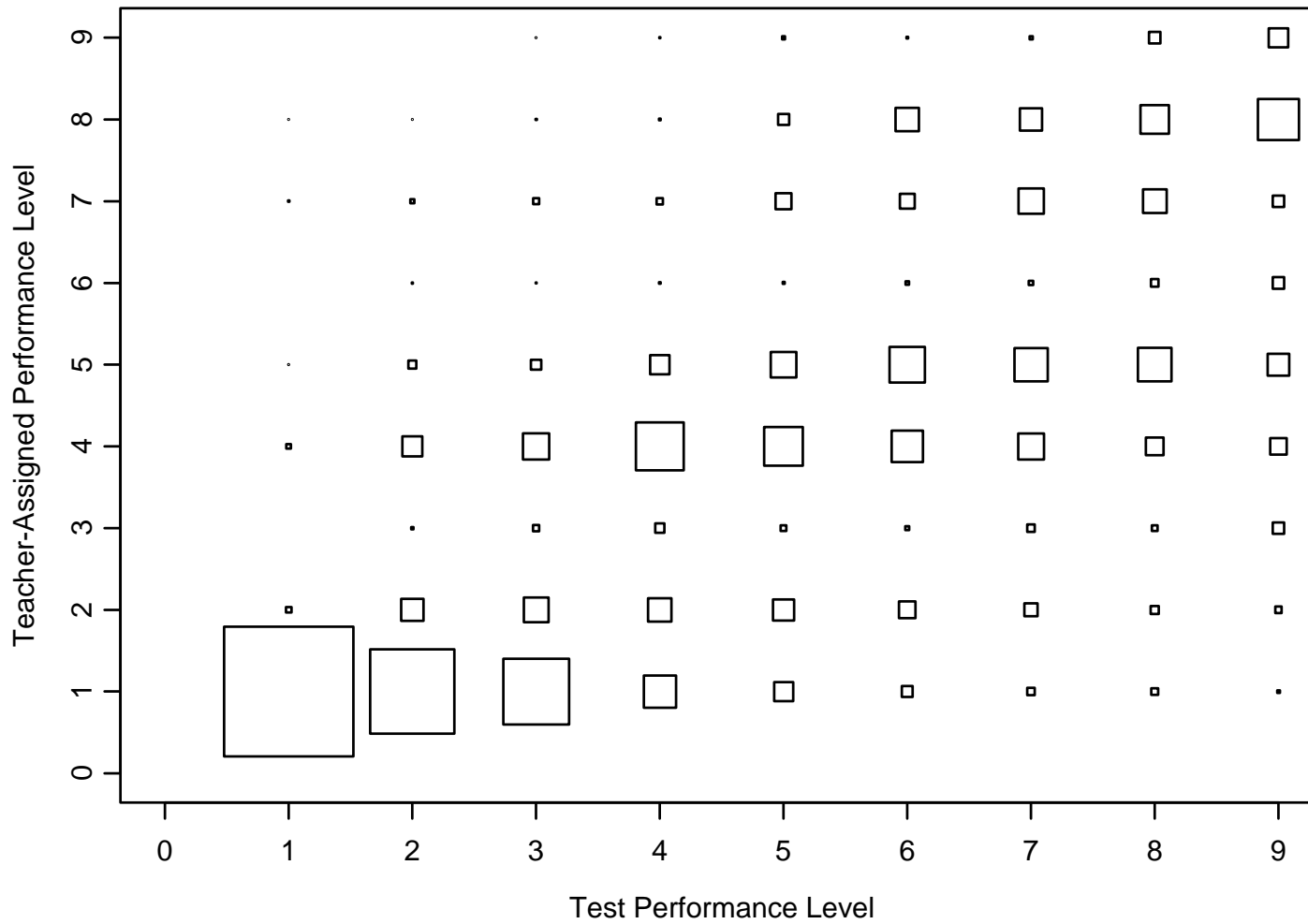


Figure I-1. Teacher Classification vs. Assessment Results for Mathematics

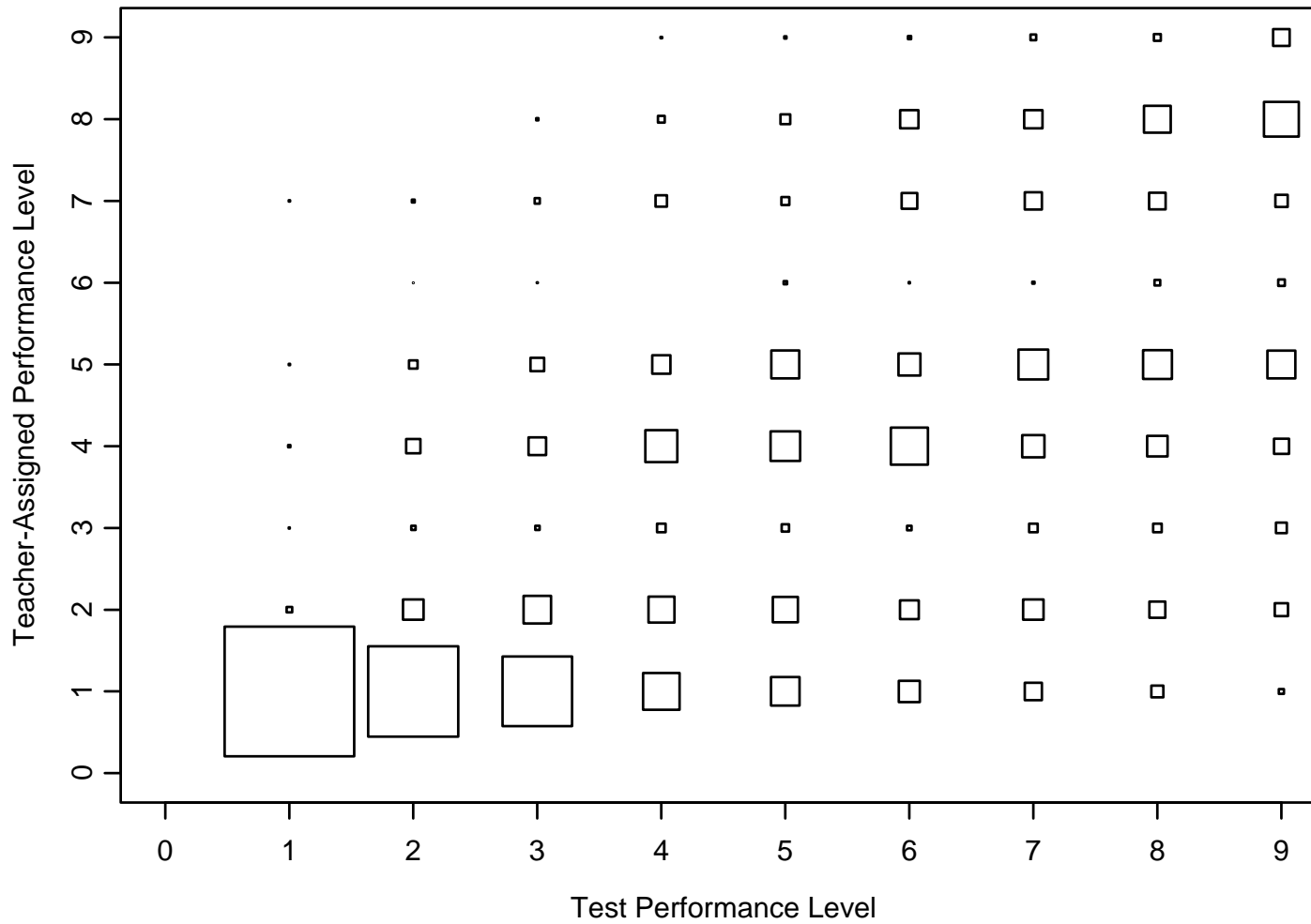


Figure I-2. Teacher Classification vs. Assessment Results for Reading

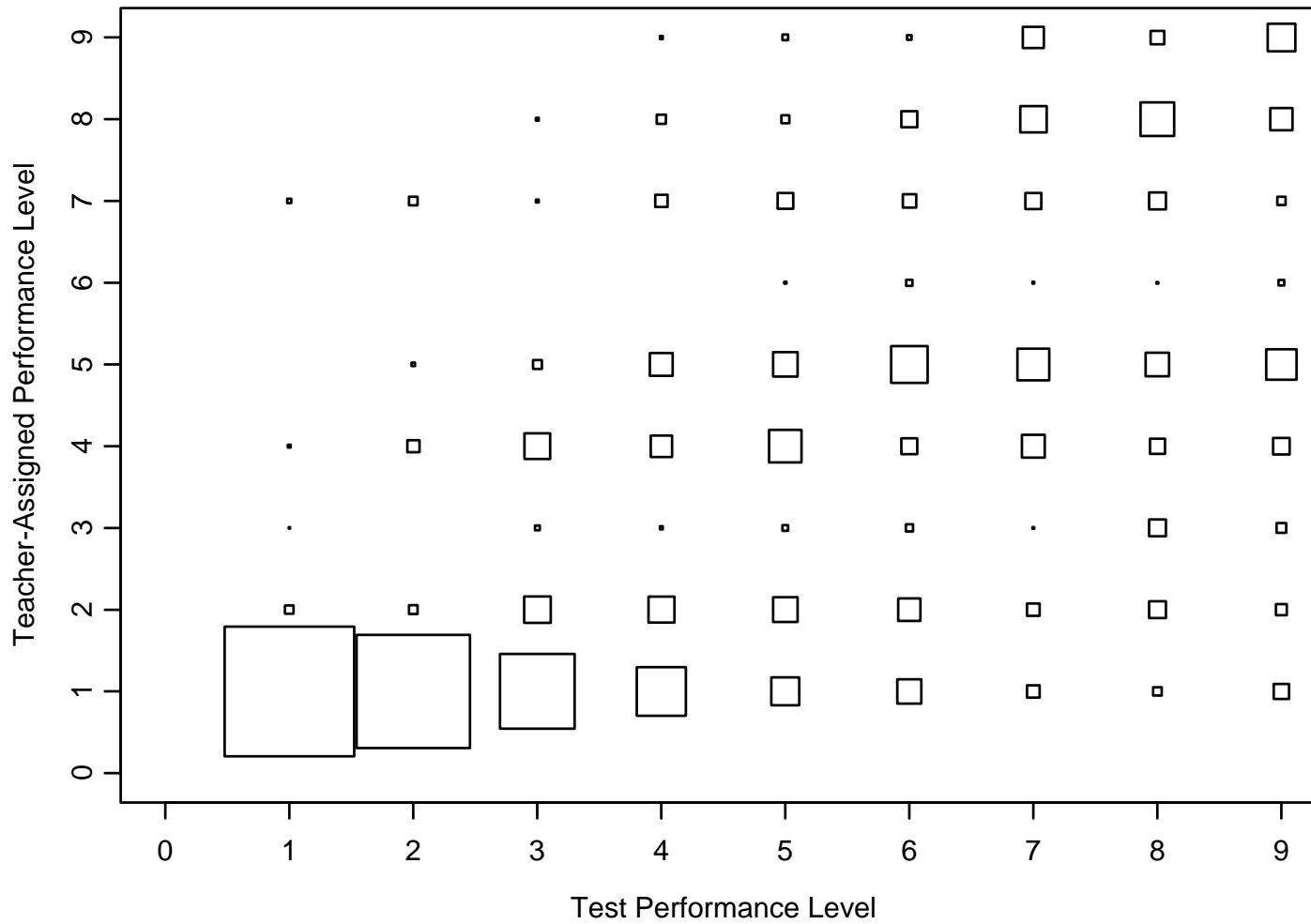


Figure I-3. Teacher Classification vs. Assessment Results for Science

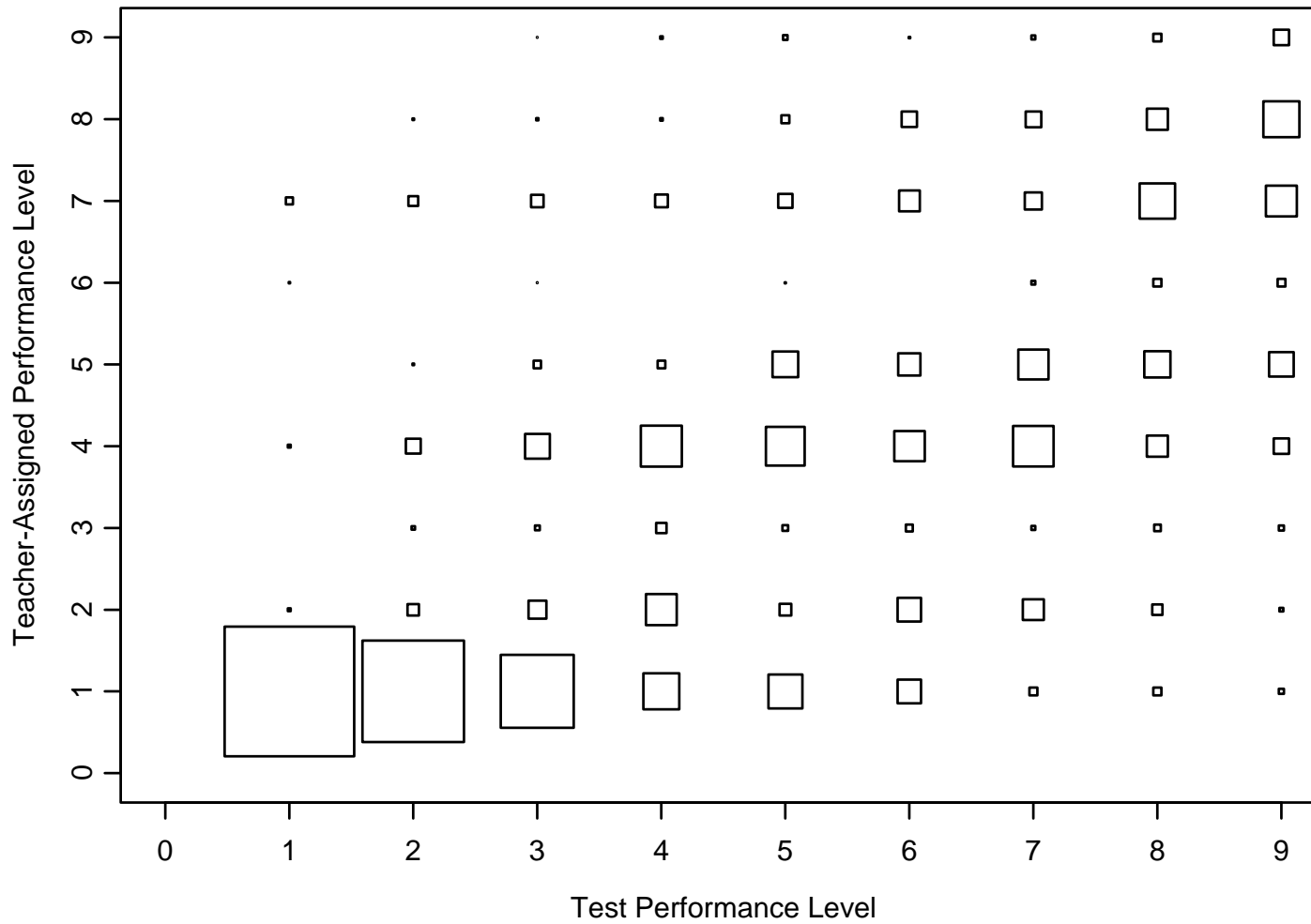


Figure I-4. Teacher Classification vs. Assessment Results for Writing

Summary

Although the results of the survey indicated fairly modest agreement with the assessment scores, there are several factors that should be kept in mind:

- The performance level definitions are based on the access point skills and are relatively new to the field. Teachers have not had much time to familiarize themselves with the access point skills, use them during daily instruction with students, or use them to assess student performance. As teachers become more familiar with the access point skills and increase their understanding of how to base instruction on them, it is likely that their ratings will be more consistent with the assessment results.
- The task that teachers were asked to complete—rating students’ level of performance relative to classroom performance—is fundamentally different from assessing student performance in the context of administering a formal assessment. Although teachers were asked to make judgments on the same performance levels, the assessment scores are obtained using a narrower range of activities. Therefore, complete agreement between teachers’ ratings and assessment results is not to be expected.

II. Video Scoring and Administration Rating Study

Objective

The design of the Florida Alternate Assessment is such that human judgment is an important factor that can affect the perceived performance of the examinees. The manifestation of human judgment is foremost in the test administration and scoring of performance assessment tasks. To ensure that results of the assessment are reliable and valid, different measures are observed to maintain procedural validity.

The objective of the study was to observe test administration and determine whether (1) the assessment is being administered consistently with test administration protocols created for this assessment program and (2) the scores being assigned by the teachers administering the assessment are consistent with scoring protocols for this assessment program.

Design

This study was implemented for the grade 5 mathematics and grade 10 writing assessments. Students were selected for each of the content area and grade combinations using stratified random sampling. A total of twenty-six grade 5 students and twenty-four grade 10 students participated in the study. The following stratification variables were used:

- School type (center school or not)
- Type of disability, including modes of response (eye-gazing, students with physical mobility limitations, etc.)
- Gender
- Ethnicity
- Urbanicity
- Score on the assessment (from prior year; grade 5 only)

The test administration for each student selected for this study was recorded on video. Grade 5 mathematics was included in the study because (1) the students are old enough to focus on the test rather than on the camera; and (2) for the mathematics content area at this level, the response booklet tool is primarily used and there are minimal cards/strips. The administration of the grade 10 writing content area

was also ideal for inclusion in this study given that this content area does not use a response booklet but instead employs the combination of strips and cards as student response tools (in conjunction with open responses). The unique tools used within each content area provided panelists the unique opportunity to observe distinct stylistic differences in teacher administration.

The recordings were viewed by two different panels:

- Teacher scorers. A panel of 18 teachers recruited from the group of teachers who administered the test in 2009 was selected for this study; teachers with the most experience administering the Florida Alternate Assessment were chosen. A list of the Video Scoring Study panelists is provided in Appendix A.
- Checklist reviewers. A panel of 15 reviewers participated in this portion of the study. The panel included both teachers and administrators. A list of the Administration Rating Study panelists is provided in Appendix B.

Each panel meeting took place over the course of a day. Each meeting began with the FLDOE providing an overview of the study. Measured Progress then trained the panelists in either scoring of the video recordings or use of the checklist with the video recordings. All facilitation and setup were performed by Measured Progress, while the FLDOE provided guidance for the desired locations and schedule.

Each teacher scorer watched the videos and scored each student's performance. Each video was scored by two separate teacher scorers and no teacher scorers scored videos submitted from their district. Scores provided by the teacher scorers were compared with each other and with the scores originally received.

The checklist reviewers watched the videos to ensure that proper test administration protocols had been followed. Using two separate test administration checklists (Administrator Checklist and Coordinator Checklist), prepared by the FLDOE and Measured Progress, checklist reviewers rated the test administration. Each item on these two checklists addressed the fidelity between what was in the test administration manual and how the assessment was actually implemented.

Focus questions were also prepared and the panelists participated in a facilitated discussion on the guidance provided for the two checklists as far as the use and ease of the checklists for administrators and coordinators.

A complete description of the logistics of the Video Scoring Study and the Administration Rating Study is provided in Appendix C.

Analysis

Scores provided by the teacher scorers were compared with the scores given by the original test administrator, as well as to each other. Specific comparisons included correlation analysis and computation of the percentages of agreement—both exact agreement and exact or adjacent agreement.

For each item on the checklist, the percentage of the ratings for each grade and content area combination was calculated. The qualitative information collected during the focused discussion was also compiled.

Results of Video Scoring

Tables II-1 and II-2 below compare the teacher-assigned (original) scores and the scores awarded by the video rescorsers, by item. Because each student was rescored by two video scorers and approximately 24 students participated for each grade, the data are based on approximately 48 observations. In some cases, video scorers were unable to rescore an item due to the limited perspective offered by the video recording; therefore, most of the *Ns* in the two tables are slightly less than 48. Included in the tables are correlation coefficients between the teacher-assigned (original) and video rescorer-awarded scores, as well as the percentages of exact and exact or adjacent agreement between the two sets of scores.

In general, the correlations and percentages exact and exact or adjacent agreement indicate a high to very high level of agreement between the two sets of scores for most of the items. For grade 5 mathematics, the correlations range from 0.79 for item 1 to 0.97 for items 8, 17, and 20. For grade 10 writing, the correlations are very similar, ranging from 0.77 for item 17 to 0.97 for five of the items. Similarly, percentages exact agreement range from a low of 67% (for writing item 7) to a high of 96% (for mathematics item 17). Percentages exact or adjacent agreement range from 89% for mathematics item 2 to 100% for a number of the items.

**Table II-1. Teacher-assigned (original) Scores vs. Video Rescores by Item
Mathematics Grade 5**

Item	N	Correlation	Percent Agreement	
			Exact	Exact or Adjacent
1	45	0.79	84	93
2	46	0.89	72	89
3	47	0.90	81	98
4	45	0.95	87	100
5*	47	0.93	87	96
6	46	0.82	89	91
7	48	0.87	83	96
8	42	0.97	93	98
9	40	0.91	83	98
10*	42	0.81	88	93
11	47	0.87	91	91
12	39	0.93	74	97
13	45	0.88	82	96
14	47	0.88	87	94
15*	46	0.91	91	93
16	48	0.80	83	96
17	46	0.97	96	98
18	44	0.92	91	95
19	45	0.92	80	98
20*	46	0.97	87	100

*Denotes field test items.

**Table II-2. Teacher-assigned (original) Scores vs. Video Rescores by Item
Writing Grade 10**

Item	N	Correlation	Percent Agreement	
			Exact	Exact or Adjacent
1	46	0.97	89	100
2	44	0.92	82	98
3	42	0.97	86	98
4	42	0.95	88	100
5*	45	0.94	84	98
6	40	0.92	78	95
7	43	0.93	67	100
8	41	0.97	85	100
9	43	0.91	81	100
10*	42	0.97	86	100
11	43	0.92	77	100
12	43	0.87	88	95
13	42	0.83	83	93
14	43	0.92	74	98
15*	46	0.97	87	100
16	40	0.92	88	95
17	43	0.77	81	93
18	43	0.97	86	100
19	40	0.91	88	95
20*	42	0.93	83	95

*Denotes field test items.

Figures II-1 and II-2 show a visual representation of the agreement between the teacher-assigned and video scores for two sample writing items, one with a high correlation (Figure II-1) and one with a lower correlation (Figure II-2). Note that because of the scoring rules for the assessment items, scores of 3 and 6 and scores of 6 and 9 are considered adjacent. Note that, in Figure II-1, the majority of the observations appear along the diagonal line while the boxes off the diagonal line represent video scores that are adjacent to the corresponding administration scores. Figure II-2, on the other hand, shows a noticeably less orderly relationship between the two sets of scores. (A complete set of item-level graphs are presented in Appendix D.)

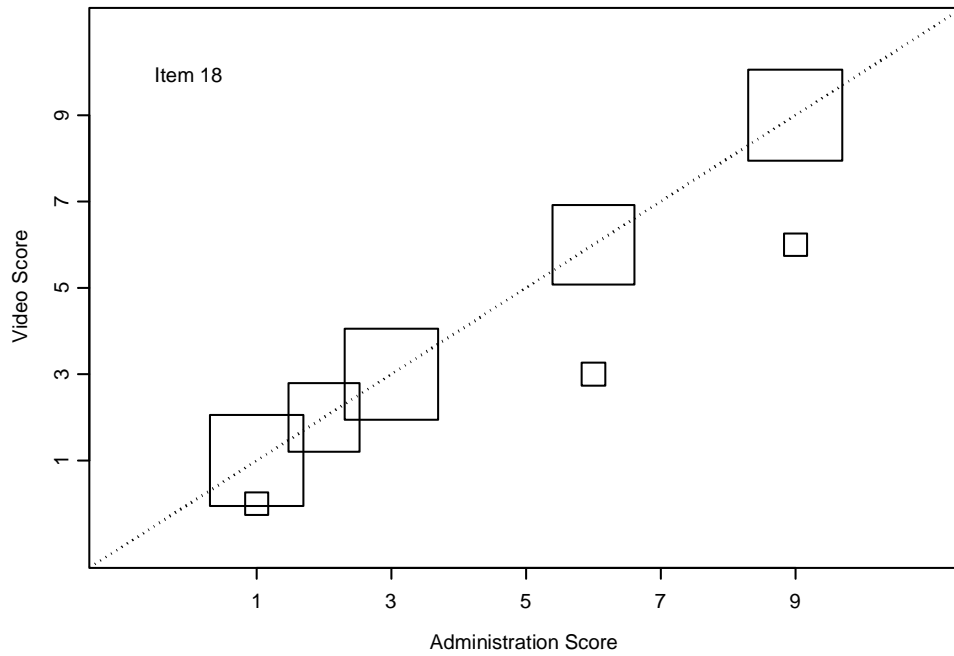


Figure II-1. Sample Writing Item with High Degree of Agreement between Original Administration and Video Scoring

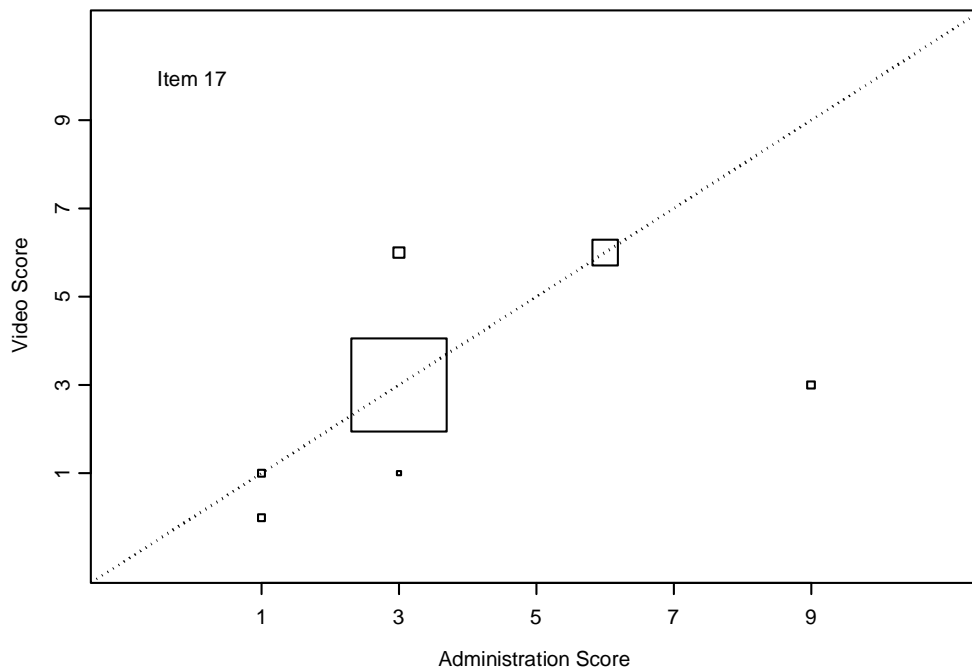


Figure II-2. Sample Writing Item with Low Degree of Agreement between Original Administration and Video Scoring

Tables II-3 and II-4 again compare the teacher-assigned (original) scores and the video-based scores, for each video scorer. In this case, the unit of observation is the assignment of an item-level score by a video scorer. So, for example, if a video scorer observed four recordings, the *N* for that rater would be expected to be approximately 80 (since 20 items are administered to each student). Therefore, the *N*s vary due to different numbers of videos scored by the different raters as well as the restricted perspective issue described above. As with the previous tables, correlation coefficients and percentages of exact and exact or adjacent agreement are included. These tables provide information about how much agreement with the initial scores varied for different raters.

For the most part, the correlations were high: approximately two-thirds 0.90 or higher, and all but four above 0.80. Similarly, the percentage of exact and the percentage of exact or adjacent agreement were high overall, with some variability across raters. Examination of Tables II-3 and II-4 reveals the following:

- It is possible to obtain very high levels of interscorer agreement but there is a fair degree of variability across raters.
- In some cases, raters had difficulty in one content area/grade level only (e.g., #011 in math; #049 in writing), while, in other cases, the lack of agreement applied to both content areas/grades (#027).
- In one case (#034 in writing), the rater's correlation was quite high (0.94) but the percentage exact agreement was low (48%). This indicates that the rater's judgments were consistently either more or less stringent than the original scorer's judgments.

**Table II-3. Teacher-assigned (original) Scores vs. Video Rescores by Rater
Mathematics Grade 5**

Rater	N	Correlation	Percent Agreement	
			Exact	Exact or Adjacent
003	51	1.00	100	100
011	60	0.56	75	83
016	56	0.85	89	98
017	77	0.94	70	96
021	40	0.87	80	95
025	17	0.98	94	100
027	20	0.68	10	50
030	20	0.91	80	100
031	69	0.96	93	99
033	57	0.88	93	98
034	93	0.98	95	100
036	60	0.98	95	100
043	40	0.94	85	95
049	53	0.88	79	94
051	57	0.85	68	88
052	57	0.99	98	100
054	56	0.96	96	98
058	18	1.00	100	100

**Table II-4. Teacher-assigned Scores vs. Video Scores by Rater
Writing Grade 10**

Rater	N	Correlation	Percent Agreement	
			Exact	Exact or Adjacent
003	41	0.94	95%	98%
011	80	0.97	90%	100%
016	27	0.93	89%	100%
017	31	0.80	84%	90%
021	53	0.83	72%	92%
025	40	0.85	75%	100%
027	38	0.73	61%	87%
030	56	0.96	89%	100%
031	20	1.00	100%	100%
033	36	0.99	97%	100%
034	60	0.94	48%	93%
036	60	1.00	98%	100%
043	40	0.93	85%	98%
049	30	0.76	77%	97%
051	51	0.96	92%	100%
052	73	0.97	68%	100%
054	70	0.96	93%	99%
058	47	0.99	96%	100%

Tables II-5 and II-6 show correlations for all possible score comparisons by item. In this case, the unit of observation is an item-level score for a given student, restricted to those for which all three scores (teacher administration, video rescore #1, and video rescore #2) are available. The maximum obtainable N for an item, therefore, is 24; the actual Ns range from 17 to 24.

Here again we see that the correlations overall are high or very high. Virtually all the correlations are 0.80 or higher, and a substantial number of them are 0.90 or higher. There do not appear to be any consistent trends by comparison, i.e., the correlations in any given column are not consistently higher or lower than the correlations in the other columns.

**Table II-5. Correlations for All Score Comparisons by Item
Mathematics Grade 5**

Item	N	Correlation		
		Administration & Video 1	Administration & Video 2	Video 1 & Video 2
1	21	0.83	0.77	0.90
2	22	0.87	0.90	0.88
3	23	0.95	0.88	0.86
4	22	0.95	0.95	0.97
5*	23	0.96	0.91	0.95
6	22	0.82	0.79	0.92
7	24	0.87	0.86	0.97
8	19	1.00	0.99	0.99
9	17	0.96	0.86	0.95
10*	20	0.88	0.71	0.82
11	23	0.93	0.80	0.84
12	17	0.95	0.91	0.93
13	21	0.88	0.89	0.84
14	23	0.92	0.90	0.84
15*	22	0.92	0.88	0.92
16	24	0.79	0.80	0.98
17	22	1.00	0.96	0.96
18	20	1.00	0.91	0.91
19	21	0.97	0.95	0.95
20*	22	0.97	0.97	0.98

*Denotes field test items.

**Table II-6. Correlations for All Score Comparisons by Item
Writing Grade 10**

Item	N	Correlation		
		Administration & Video 1	Administration & Video 2	Video 1 & Video 2
1	22	0.99	0.97	0.98
2	21	0.97	0.89	0.92
3	19	0.99	0.94	0.95
4	20	0.98	0.95	0.98
5*	21	0.91	0.96	0.92
6	17	0.92	0.92	1.00
7	19	0.95	0.93	0.95
8	17	0.96	0.98	0.99
9	19	0.89	0.92	0.97
10*	19	0.97	0.99	0.98
11	21	0.90	0.96	0.93
12	19	0.98	0.86	0.89
13	19	0.89	0.74	0.83
14	21	0.96	0.88	0.94
15*	23	0.95	0.99	0.97
16	17	0.92	0.92	1.00
17	20	0.82	0.75	0.90
18	21	0.99	0.96	0.98
19	19	0.92	0.92	1.00
20*	21	0.93	0.92	0.99

*Denotes field test items.

Results of Administration Rating Study

Tables II-7 and II-8 present a summary of the results of the observation checklists. For both checklists, answers of “yes” indicate that the teacher was following the protocols in administering the assessment.

As shown in Table II-7, which presents the results of the administrator checklist, the “yes” percentages were quite high for grade 5 mathematics, ranging from 83% to 100%. Seventeen percent of raters disagreed that “the teacher made sure the student was focused on the item before beginning that item.” For the remaining questions, the raters judged that the administrators followed administration protocols almost perfectly. For grade 10 writing, the “yes” percentages ranged from 88% to 100%. Twelve percent of raters disagreed that “the test was administered in an area where the student could focus” and 8% disagreed that “the teacher made sure the student was focused on the item.”

Table II-7. Summary of the Results of the Administrator Checklist by Content Area and Grade

Checklist Item	Mathematics Grade 5		Writing Grade 10	
	Yes	No	Yes	No
The assessment was administered one on one.	100%	0%	100%	0%
The test was administered in an area where the student could focus (quiet area and away from distractions).	96%	4%	88%	12%
The teacher made sure the student was focused on the item before beginning that item.	83%	17%	92%	8%
The teacher had all of the appropriate booklets, and/or cut outs within the student’s reach.	100%	0%	98%	2%
If mathematics was being administered, the teacher had a calculator, number line, and/or counters on the work surface.	96%	4%	NA	NA
The teacher recorded the student’s response to the item during the test administration.	96%	4%	94%	6%

The percentages in Table II-8, which presents the results of the district coordinator checklist, are also quite high for the most part, but with a few notable exceptions. For grade 5 mathematics, only 33% of raters answered yes to the question “Did the teacher follow the scripting verbatim?” In addition, 14% said no to two of the questions, “Did the teacher follow the process outlined in the Scoring Rubric Flowchart?” and “Did the teacher repeat the item to the student up to two times, for a total of three times as needed?” For grade 10 writing, 50% answered no to the question “If an item had cut outs, did the teacher place the cards/strips in the order specified in the test booklet?” In addition, 46% answered no to the question “Did the teacher follow the scripting verbatim?”

**Table II-8. Summary of the Results of the
District Coordinator Checklist by Content Area and Grade**

Checklist Item	Mathematics Grade 5		Writing Grade 10	
	Yes	No	Yes	No
Did the teacher place any booklets and cut outs required within the student's reach?	100%	0%	96%	4%
If an item had cut outs, did the teacher place the cards/strips in the order specified in the test booklet?	100%	0%	50%	50%
Did the teacher follow the process outlined in the Scoring Rubric Flowchart?	86%	14%	96%	4%
Did the teacher use scaffolding, when necessary, at the participatory level of complexity, but never for supported or independent levels?	96%	4%	96%	4%
Did the teacher focus the student on each item before beginning the item?	100%	0%	92%	8%
Did the teacher follow the scripting verbatim?	33%	67%	54%	46%
Did the teacher repeat the item to the student up to two times, for a total of three times as needed?	86%	14%	92%	8%
When an item required the student to give more than one response did the teacher cue the student for another response?	92%	8%	96%	4%
Did the teacher mark the student responses in the test booklet or directly on the scan sheet as the teacher administered the assessment?	100%	0%	100%	0%

The results of the checklists indicate that, while test administrators did a good job overall of following the administration protocols, there is some need for improvement in the training on some aspects of test administration.

The most common panelist feedback received from the Administrator Checklist Study (in the form of written comments/notes from grade 5 mathematics and grade 10 writing checklists) related to the teacher ensuring that the student was focused on the item before beginning the item. Panelists made notes indicating that the student was already engaged at the beginning of an item; hence there was no reason for the teacher to re-check for student focus. In other instances, a teacher may have focused the student on the

item frequently at the beginning of the item, but did not focus the student in the event that the student disengaged during the item administration process or after the item was administered the first time (directly prior to the teacher repeating the item).

Most comments made by panelists on the District Coordinator Checklist (grade 5 mathematics) indicated that teachers did not adhere to the scripting in the *Teacher Will* section of the test booklet. Panelists also indicated that teachers on the videos often paraphrased the instructions from the test booklet, did not read number or word/picture cards to a student, or used the phrase *Show me/tell me* together rather than *Show me* or *tell me* based on the student's mode of communication. One panelist noted that the teacher misidentified a shape although she [teacher] was not supposed to identify the shapes as part of the stimulus. This particular teacher also identified pictures and numbers when not prompted to do so and manipulated the counters instead of requiring the student to do so per directions. Panelists also reported varying styles and use of teacher-gathered materials outlined in the test booklet.

Comments from the District Coordinator Checklist (grade 10 writing) primarily related to teachers not placing the writing cut out strips in the correct order as outlined in the test booklet. One panelist noted that the teacher on the video used a random placement of cut outs. A panelist evaluating a different video noted that a student was rearranging cards/strips as the teacher separated cards from the writing assessment stack. On occasion, panelists indicated that cut outs were not placed within the student's reach. One panelist believed that the teacher of a student who uses eye-gazing as a means of communication did not place cut outs far enough apart on the board to discern student response.

In addition to the collection of comments gathered from panelists who rated videos using the checklists, a roundtable discussion was held after the use of each of the Administrator Checklist and District Coordinator Checklist, to elicit feedback from panelists about the setup and content of the checklists themselves. Comments ranged from suggested edits for clarifying the criteria under evaluation to group consensus related to three items being the "perfect" number to observe.

Summary

Information gathered from this study can be used to improve upon aspects of the assessment that might threaten validity. For example, the item information presented in Tables II-1, II-2, II-5, and II-6 can help identify items with lower interrater consistency. The scoring rubrics for these items can then be evaluated to see whether they need to be refined. Alternatively, more training on scoring these items may be warranted, or refinements to the administration protocols may be needed.

The rater-level information presented in Tables II-3 and II-4 indicates that a high degree of interrater agreement can be obtained, but that there is some variability among raters. These results point to the need for careful training of raters. In addition, ideally, checks should be put into place to monitor whether raters are following the scoring and administration protocols accurately.

Scores on the checklist indicate that, overall, test administration protocols appear to be followed fairly well. However, several aspects of the assessment program can be improved, either by implementing improvements in the teacher training or by tweaking parts of the protocol that have been subject to misinterpretation during test administration.

The timing of the onsite video review studies in April 2009 permitted clarification of instructions and insertion of additional guidance related to administration practices in key sections of the *Florida Alternate Assessment Administration Manual 2009–2010*. In addition, feedback from the studies was integrated into the train-the-trainer meetings held in July 2009 and will be incorporated into subsequent teacher trainings held throughout the state of Florida prior to the 2009–2010 assessment administration window.

III. Test-Retest Reliability Study

Objective

The Florida Alternate Assessment is based on a tiered level of difficulty. Sunshine State Standards access points approved by the Florida State Board of Education create the frameworks upon which alternate assessment items are constructed. A single item consists of three questions, one at the participatory level of complexity (least challenging), one at the supported level of complexity, and one at the independent level of complexity (most challenging).

Each student starts at the participatory level of complexity question of an item. A student completing the participatory level of complexity question accurately and without assistance moves to the supported level of complexity question. A student completing the supported level of complexity question accurately moves on to the independent level of complexity question. In this way, the student moves up through the access points as long as he or she is able to respond accurately and independently.

The student's final score for the item is based on the highest level at which it was answered correctly. If the student is unable to complete the question at the participatory level of complexity, he or she receives scaffolding and will be awarded a score of 1 or 2, depending on the amount of assistance given. If the student answers the question without assistance at the participatory level, but is unable to complete the question at the supported level of complexity, he or she retains the **3-point score** from the participatory level of complexity. If the student is able to complete the question at the supported level of complexity, the teacher will next administer the independent level of complexity question. If the student is unable to complete the independent level of complexity question accurately, a score of **6 points** is awarded. If the student completes the independent level of complexity question accurately, the teacher will record a score of **9 points**. If the student will not engage or actively refuses at any point within the participatory level of complexity question, the student will be scored at **0 points**.

This method of test construction theoretically permits an increasing level of complexity for the questions within an item. In order to confirm that the questions within each developed item are in an order of hierarchical difficulty, it becomes necessary to compare the scores of the administration method described above to an administration method that provides the opportunity for a student to respond to all questions within an item (irrespective of achieving a correct score at any level of complexity).

This study examined the hypothesis that student scores will not improve when the assessment is readministered in entirety using the new administration guidelines.

Design

This study involved students who participated in the 2009 Florida Alternate Assessment. The relevant students of interest were those individuals who consistently scored at the participatory, but not supported, level of complexity. Samples of 50 students each were selected to re-take reading in grade 8, and mathematics in grade 5. One grade 5 student dropped out of the study due to unexpected illness. The readministration window for the test-retest study was approximately April 27, 2009, through Friday, May 29, 2009.

The FLDOE reviewed data from the spring 2008 assessment to gain a sense of which students theoretically qualified as candidates to participate in the study from prospective schools across the state. The FLDOE then set expectations regarding which students at a particular school within a district were suited for participation. A preliminary student roster was drawn up by with FLDOE; a backup list of students who consistently scored at the participatory level was given to Measured Progress in the event a student was not able to participate. The FLDOE also contacted alternate assessment coordinators to confirm which districts and schools were selected for the study. In turn, Measured Progress provided

alternate assessment coordinators with district-specific student rosters. Alternate assessment coordinators contacted teachers at each student's school to provide study materials. Students not able to participate in the study were replaced with a student from the backup list supplied by Measured Progress.

A FLDOE-approved informational letter was sent out by Measured Progress to school administrators and teachers of selected districts outlining the goals, expectations, timing, and resources required of teachers to participate in the Test-Retest Reliability Study. In addition, Measured Progress prepared a new administration flowchart, scoring instructions, and student scannable. These materials instructed teachers on how to administer the retest and fill out the student scannable appropriately.

Piedra Data Services provided alternate assessment coordinators with assessment materials. Coordinators were asked to pull out the content-specific materials for each grade and distribute the assessment materials along with the study-related materials sent to coordinators by Measured Progress. Alternate assessment coordinators were also the point of contact for return of all study-related materials. After the student retest administration was completed, coordinators returned both the assessment materials and study-related materials to Measured Progress.

Each box or envelope returned to Measured Progress had district identification so that materials could be sequestered by district. Materials were logged in by district to ensure all assessment materials (secure) and study-related information (including student information and completed scannables) were accounted for.

Analysis

The design of the Florida Alternate Assessment is adaptive in nature. That is, the sub-item that the student responds to depends on his/her performance on the previous sub-item. For the students selected in this study, their adaptive scores and non-adaptive scores on the assessment were compared. Statistical tests were performed to explore whether a non-adaptive administration of the assessment improved student performance. A *t*-test was performed to test the null hypothesis of no improvement on overall test performance (i.e., total raw score).

Results

Tables III-1 and III-2 compare the original and retest item scores for grade 5 mathematics and grade 8 reading, respectively. The unit of observation is an item score; therefore, if 50 students participated in the retest study and all students completed all 20 items, the *N*s would be expected to sum to 1000. For grade 5 mathematics, of the 110 student responses originally scored as 0, 77% of those were rescored as 0 at the participatory level in the retest, while 100% were scored at 0 at both the supported and independent levels (i.e., students responded incorrectly to the supported and independent levels of the items).

Interestingly, students with an original score of 6 were not overly consistent when presented with the supported item on the retest: for mathematics, 32% of students answered the supported item correctly on the retest, while for reading, 38% answered correctly. The same is true for students who originally received a 9: only 29% of them answered the independent item correctly on the retest for mathematics, and only 21% for reading. These results should be treated very cautiously given the small numbers of student responses on which they are based. However, results presented in the tables strongly suggest that a fair amount of variability in scores can be expected for this population of students.

While the focus of this study is the effect of the administration mode on students' total scores, information about item-level effects may also help identify individual items that are particularly problematic. Item-level versions of Tables III-1 and III-2 are, therefore, provided in Appendix E.

**Table III-1. Comparison of Original and Retest Item Scores
Mathematics Grade 5**

Original Score	N	Percentage at Each Score on the Retest							
		Participatory				Supported		Independent	
		0	1	2	3	0	6	0	9
0	110	77	17	5	1	100	0	100	0
1	372	12	61	16	11	95	5	94	6
2	169	4	36	30	31	89	11	93	7
3	140	4	33	21	42	92	8	94	6
6	22	0	23	27	50	68	32	95	5
9	7	0	0	43	57	57	43	71	29

**Table III-2. Comparison of Original and Retest Item Scores
Reading Grade 8**

Original Score	N	Percentage at Each Score on the Retest							
		Participatory				Supported		Independent	
		0	1	2	3	0	6	0	9
0	70	36	49	6	10	93	7	97	3
1	493	12	64	13	11	96	4	98	2
2	146	1	46	30	23	92	8	94	6
3	151	1	22	34	42	87	13	91	9
6	26	0	15	38	46	62	38	88	12
9	14	0	14	21	64	79	21	79	21

Tables III-3 and III-4 show basically the same information as Tables III-1 and III-2, but focus on the categories that more directly address the question of interest. Specifically, comparisons are made for items on which students originally received a score of 0, 1, or 2 (i.e., cases in which the student would not have been presented the supported level of the item) and for items on which students originally received a score of 0, 1, 2, or 3 (cases in which the student would not have been presented the independent level of the item).

For mathematics, for items on which the students would not have been presented either the supported or independent level of the item, approximately 6% could be expected to get the supported level correct if it had been presented to them, and approximately 5% could be expected to get the independent level correct. For reading, the corresponding percentages are 5% and 3%. For items on which the students *did* see the supported level of the item (but did not answer correctly), and did not see the independent level of the item, the percentages for mathematics are 6% and 5%, while those for reading are 7% or 4%. These results indicate that, overall, the odds that students' scores are artificially depressed by the mode of administration are quite low. In fact, comparing these results to the variability in scores described above for students who originally received a 6 or a 9 on an item suggests that this variability may outweigh any potential disadvantage of not being exposed to all of the test items.

**Table III-3. Comparison of Item Scores by Items Originally Presented
Mathematics Grade 5**

Original Score	N	Number (and Percent) at Each Score on the Retest			
		Supported		Independent	
		0	6	0	9
0,1,2	651	614 (94)	37 (6)	617 (95)	34 (5)
0,1,2,3	791	743 (94)	48 (6)	748 (95)	43 (5)

**Table III-4. Comparison of Item Scores by Items Originally Presented
Reading Grade 8**

Original Score	N	Number (and Percent) at Each Score on the Retest			
		Supported		Independent	
		0	6	0	9
0,1,2	709	672 (95)	37 (5)	688 (97)	21 (3)
0,1,2,3	860	803 (93)	57 (7)	826 (96)	34 (4)

To supplement the results shown in the tables above, a unidirectional paired *t*-test was also conducted, comparing the overall raw scores for the original administration and the retest. The results of the *t*-tests were found to be nonsignificant at the 0.05 significance level for both grade 5 mathematics and grade 8 reading. These results indicate that being presented with the supported and independent levels of the items did not result in a significantly higher total score for students.

Summary

The results of this study suggest that not being presented with the supported and independent levels of the items does not significantly impact the scores that students who are performing primarily at the participatory level would be expected to receive on the assessment. Although scores for students did increase somewhat overall as a result of being given the opportunity to answer the higher-level items, that increase in scores is small compared to overall variability in item scores, and not greater than would be expected due to chance.

**Appendix A: Video Scoring and Administration
Rating Study List of Video Scoring Study Panelists**

Video Scoring Study Panelists (April 23, 2009)					
First Name	Last Name	District	District Size	Selected For	Position
Terri	Messer	Brevard	Large	Video Scoring only	ESE Teacher
Kelly	Stevenson	Collier	Large	Video Scoring only	VE Teacher
Kim	Garman	Escambia	Large	Video Scoring only	ESE Teacher
Marilyn	Halsey	Jefferson	Small	Video Scoring only	Teacher
Michelle	Smith	Lee	Large	Video Scoring only	ESE Life Skills Teacher
Freida	Strickland	Levy	Small	Video Scoring only	Teacher
Celeste	Middleton	Pasco	Large	Video Scoring only	Teacher (ASD)
Deborah	Cotney	Polk	Large	Video Scoring only	TMH Teacher
Kelly	Tacy	Sarasota	Large	Video Scoring only	ESE Teacher
Andria	Tichy	St. Johns	Medium	Video Scoring only	ESE Teacher
Jean	Collins	Clay	Medium	Checklist & Video Scoring	Intellectual Disabilities Teacher
Pamela	Stolsworth	Flagler	Medium/Small	Checklist & Video Scoring	ESE Teacher (PI)
Marie	Schwartz	Hardee	Small	Checklist & Video Scoring	ESE Teacher
Sue	Berg	Hernando	Medium	Checklist & Video Scoring	ESE Teacher / ESE Dept. Chair
Maria	Rivas	Hillsborough	Very Large	Checklist & Video Scoring	SPMH Teacher
Linda	Pillows	Lake	Large	Checklist & Video Scoring	Teacher, ESE
Patricia	Elkin	Lee	Large	Checklist & Video Scoring	ESE Teacher & School Counselor
Dianne	Febles	Nassau	Medium/Small	Checklist & Video Scoring	Participatory Non-Ambulatory Teacher

**Appendix B: Video Scoring and Administration
Rating Study List of Administration Rating Study Panelists**

Admin Rating (Checklist) Study Panelists (April 24, 2009)					
First Name	Last Name	District	District Size	Selected For	Position
Catherine	Anderson	Bay	Medium	Checklist only	ESE Resource Teacher
Karl	Amundson	Citrus	Medium/Small	Checklist only	Alt Coordinator
Jeris	Bookhard	Duval	Very Large	Checklist only	Alt Coordinator
Margie	Haugh	Lee	Large	Checklist only	ESE Program Specialist, Alternate Assessment
Jill	Brookner	Miami-Dade	Very Large	Checklist only	Alt Coordinator
David	Hill	St. Johns	Medium	Checklist only	ESE Program Specialist
Susan	Reaves	Volusia	Large	Checklist only	Alt Coordinator
Jean	Collins	Clay	Medium	Checklist & Video Scoring	Intellectual Disabilities Teacher
Pamela	Stolsworth	Flagler	Medium/Small	Checklist & Video Scoring	ESE Teacher (PI)
Marie	Schwartz	Hardee	Small	Checklist & Video Scoring	ESE Teacher
Sue	Berg	Hernando	Medium	Checklist & Video Scoring	ESE Teacher / ESE Dept. Chair
Maria	Rivas	Hillsborough	Very Large	Checklist & Video Scoring	SPMH Teacher
Linda	Pillows	Lake	Large	Checklist & Video Scoring	Teacher, ESE
Patricia	Elkin	Lee	Large	Checklist & Video Scoring	ESE Teacher & School Counselor
Dianne	Febles	Nassau	Medium/Small	Checklist & Video Scoring	Participatory Non-Ambulatory Teacher

Appendix C: Video Scoring and Administration Rating Study Logistical Details

Solicitation of Districts to Participate in the Study

The Florida Department of Education (FLDOE) contacted alternate assessment coordinators throughout the state to request volunteers to participate in the Administration Rating and Video Scoring Study. Each district was asked to present two teacher candidates, one teacher each for grade 5 and grade 10. Efforts were made to recruit teachers with different amounts of tenure; experience teaching students working at participatory, supported, or independent level access points; experience with various accommodations given to students with significant cognitive disabilities (e.g., students with hearing, visual, and/or physical impairments); experience with teaching English Language Learners; exposure to different types of training associated with the Florida Alternate Assessment; and experience administering the Florida Alternate Assessment to students.

In total, five teachers from very large districts, thirteen teachers from large districts, five teachers from medium districts, four teachers from medium/small districts, and six teachers from small districts participated in the study.

While each teacher participating in the study could select the student of interest to be video recorded for the study, teachers were asked to select students with a wide range of significant cognitive disabilities. Teachers were provided a form for noting each student's mode of communication so that video rescorsers would know whether the student communicated by sign language, eye gazing, assistive technology, Braille, pointing to objects, and/or verbal speech. Space was provided on the form so that the teacher could include additional notes, such as comments about breaks taken or any particular challenges experienced while the video was being recorded.

Detailed instructions for the teacher and videographer were provided for guidance related to the study and video creation process. In addition, a parent/guardian consent form was provided as a resource that teachers could use in conjunction with existing district-specific paperwork.

Video Receipt at Measured Progress

Each video recording received at Measured Progress was labeled with a unique student ID number, date of birth, first name, last name, grade, and district number. Audio and video quality were checked by a qualified technician; any issues were noted at the time of occurrence in the recording. In one case, student last name was edited out of the video prior to receipt into Measured Progress to protect student identity.

All recordings submitted on videotape were converted to DVD for the purpose of standardizing the media used during onsite review. File format was checked to ensure compatibility with common video software programs such as Windows Media Player.

Student mode of communication forms were turned in with videos. Teachers were contacted in the event this information was missing.

Of the 58 candidates confirmed to participate in the study, 7 participants dropped out of the study due to one of the following reasons: video recording production issues, withdrawal of parent/guardian consent for student participation, or student illness. One video recording received was of grade 5 science rather than mathematics. Due to the design of the study, this particular video was not used in conjunction with the scoring and checklist studies. In total, 26 videos of grade 5 mathematics and 24 videos of grade 10 writing were used in the study.

Video Scoring Study—Onsite Review (April 23, 2009)

Panelists selected to act as blind scorers were teachers who had been video recorded while administering the assessment to a student and had submitted a viable video to the study.

Representatives from Measured Progress and the FLDOE were present throughout the video scoring process. As a group, panelists were given an overview of the Florida Alternate Assessment and training for video scoring. Each panelist then signed out a grade 5 and grade 10 test booklet containing content relevant questions administered on the spring 2009 assessment. Each test booklet could be used as a reference guide to help panelists follow along with each video recording viewed.

DVDs were separated by grade and alphabetized by student first name so that the student's mode of communication form could be matched up with the video. Each panelist selected the first available alphabetized video on the table; a cross-check was completed to ensure that a panelist did not receive his/her own video submitted to the study, nor did the panelist receive a video from his/her district. The panelist ID number and video was recorded on a separate check-out list to ensure no panelist would review the same video twice. Panelists were also provided with a scoring rubric and scannable answer sheet pre-populated with student first name, student ID, date of birth, district number, school, and grade. A comment form was supplied so that panelists could individually make notes for each video throughout the rating process.

Notebook computers and headphones were placed on tables around the room to permit independent review of each video. Directly prior to scoring each video, panelists were asked to cross-check that the information on the DVD label, mode of communication form, and pre-populated scan sheet all matched. In addition, each panelist recorded his/her unique ID number on the lower right-hand corner of the scan sheet, took out the relevant test booklet for the grade, and reviewed the student mode of communication form prior to scoring.

Instructions were provided to help panelists score selected-response and open-response items. Items were not scored if a panelist was not able to accurately discern the student response shown on the video. Common reasons why an item could not be scored related to the position of the camera and audio recording equipment; either the camera did not capture the student response accurately enough to enable the panelist to understand the answer, or the camera (focus locked on a fixed point rather than zooming in/out or the view panning across a surface) was not able to capture the materials as they moved around the surface of the table or desk. The capture of auxiliary materials was particularly challenging for some of the writing items that involved student manipulation of multiple cards and strips. In addition, an item may not have received a score due to discrepancies between the teacher's administration of the item and the item administration outlined in the teacher administration manual. Examples include teachers who used their hands (rather than pieces of paper) during the scaffolding process, teachers who read the entire *Teacher will* section of the test booklet (including teacher directions that are not supposed to be read to the student), and teachers who did not place cards/strips in the correct order as outlined in the test booklet. In addition to scoring, panelists were asked to provide written comments related to the scoring of an item or to the video itself.

Upon completion of scoring each video, panelists returned the DVD, student scan sheet, mode of communication form, and comment form. The process then started again once another DVD was provided. Throughout the day panelists reviewed a mix of grade 5 mathematics and grade 10 writing videos. On average, each panelist viewed and scored approximately six videos.

Checklist Administration Study—Onsite Review (April 24, 2009)

Videos used for the Video Scoring Study were also used for the Checklist Administration Study. A series of three checklists (Administrator Checklist, District Coordinator Checklist, and Teacher Self-Evaluation Checklist) were created by Measured Progress and reviewed by the FLDOE prior to the onsite meeting. For the purposes of onsite review, panelists were primarily focused on using the videos in conjunction with the Administrator Checklist and District Coordinator Checklist to evaluate whether the assessment was being administered consistent with test administration protocols.

Panelists consisted of a mix of teachers who had participated in the video scoring held on Thursday, April 24, teachers from around the state of Florida with expertise in special education and familiarity with the Florida Alternate Assessment, and district alternate assessment coordinators.

DVDs were randomly assigned a set of three consecutive items a panelist would review in conjunction with a particular checklist. Panelists were asked to observe the administration of at least three items to gain a tangible sense of administration style prior to completing each rating checklist. The starting time for the first item in the series was noted on the DVD cover so that panelists could fast-forward to the start of the item administration.

Representatives from Measured Progress and the FLDOE were present throughout the checklist process. As a group, panelists were given an overview of the Florida Alternate Assessment and training for how to rate videos using the Administrator Checklist and District Coordinator Checklist. Instructions were provided to help panelists understand the different criteria present on each checklist, including specific examples of suitable administration techniques that could be used for this population of students with disabilities. For example, a teacher may orient the student to the assessment materials by pointing at the materials, taping the desk, or speaking the student's name prior to administering an item. The Administrator Checklist had a total of six criteria while the District Coordinator Observation Checklist had a total of ten criteria that panelists would use to rate videos.

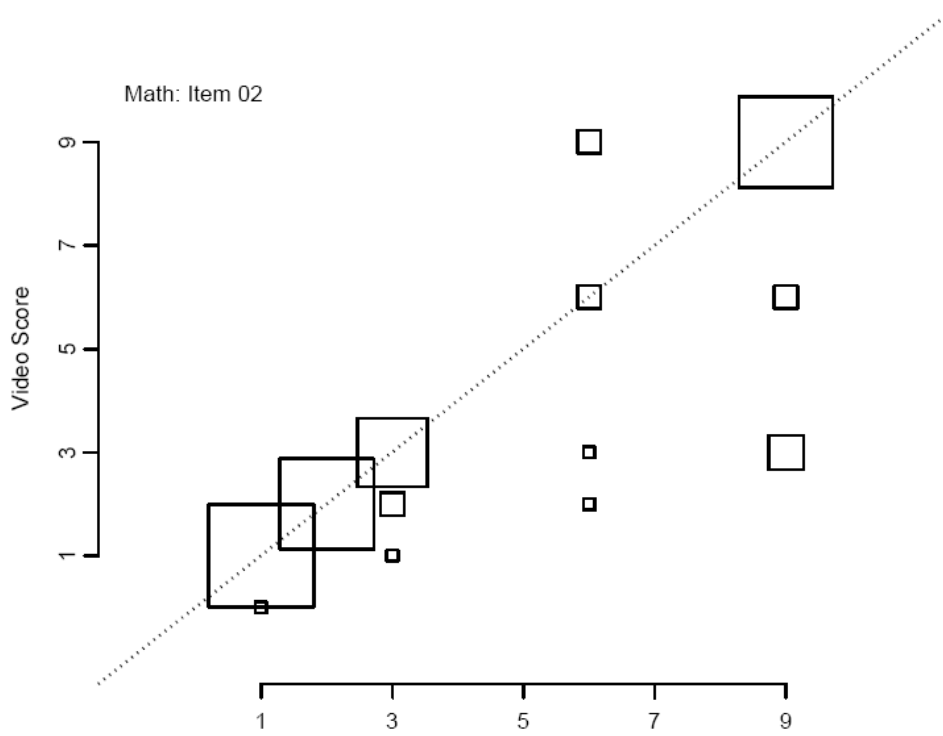
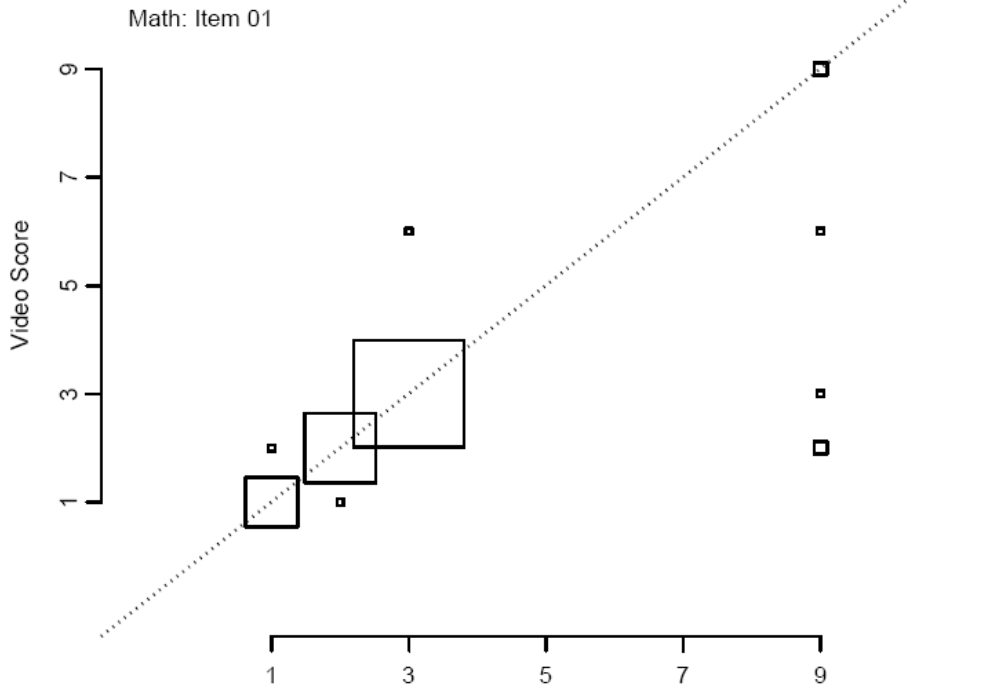
Each panelist signed out a grade 5 and grade 10 test booklet containing content relevant questions administered on the spring 2009 assessment. A test booklet could be used as reference guide to help panelists follow along with each video recording viewed.

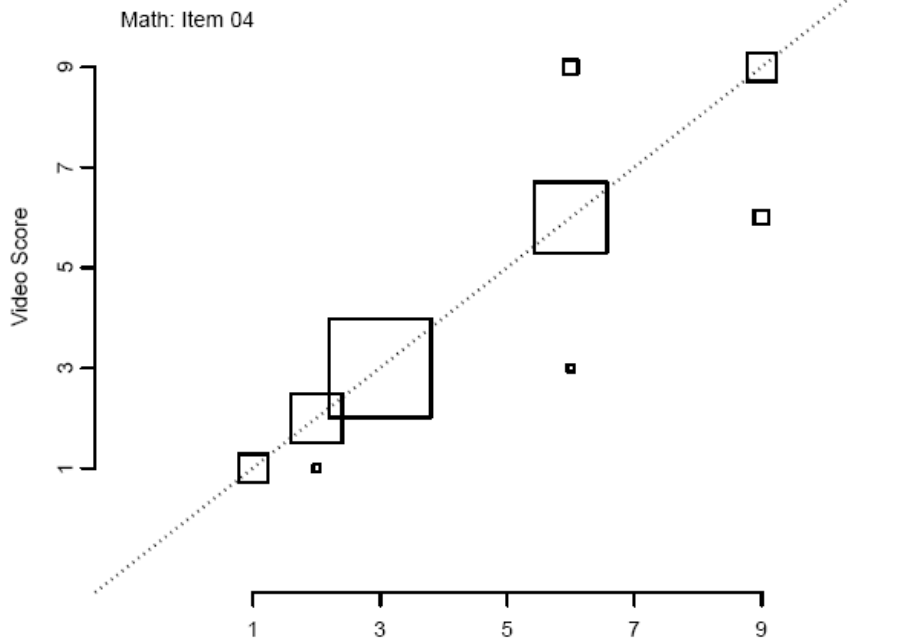
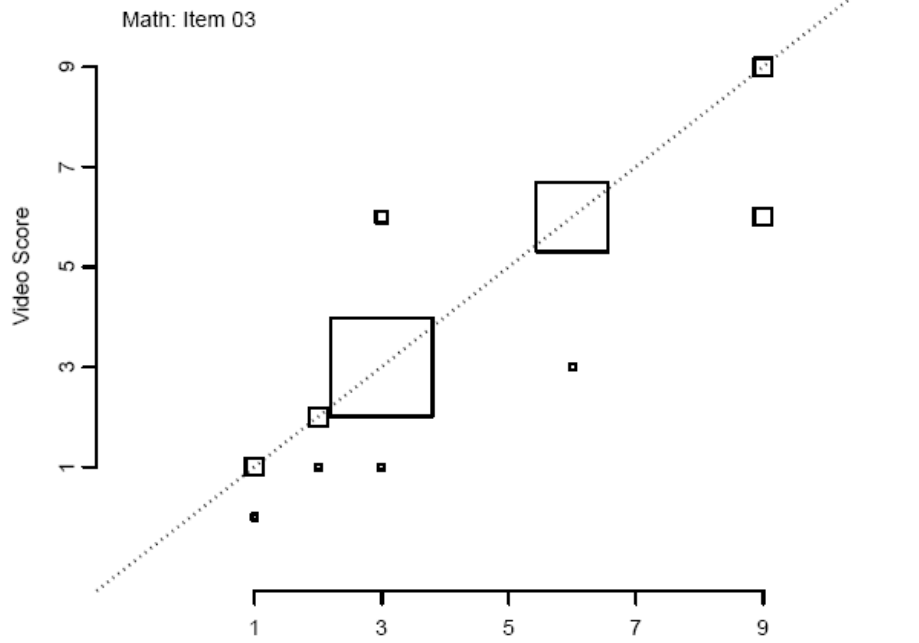
DVDs were separated by grade and alphabetized by student first name so that the student's mode of communication form could be matched up with the video. Each panelist selected the first available alphabetized video on the table; a cross-check was completed to ensure that a panelist did not receive his/her own video submitted to the study, nor did the panelist receive a video from his/her district. The panelist ID number and video was recorded on a separate check-out list to ensure no panelist would review the same video twice on a particular checklist. Panelists were also provided with a scoring rubric, checklist, and labels pre-populated with student first name, student ID, date of birth, district number, school, and grade (to be affixed to each checklist used). A comment form was supplied so that panelists could individually make notes for each video throughout the checklist rating process.

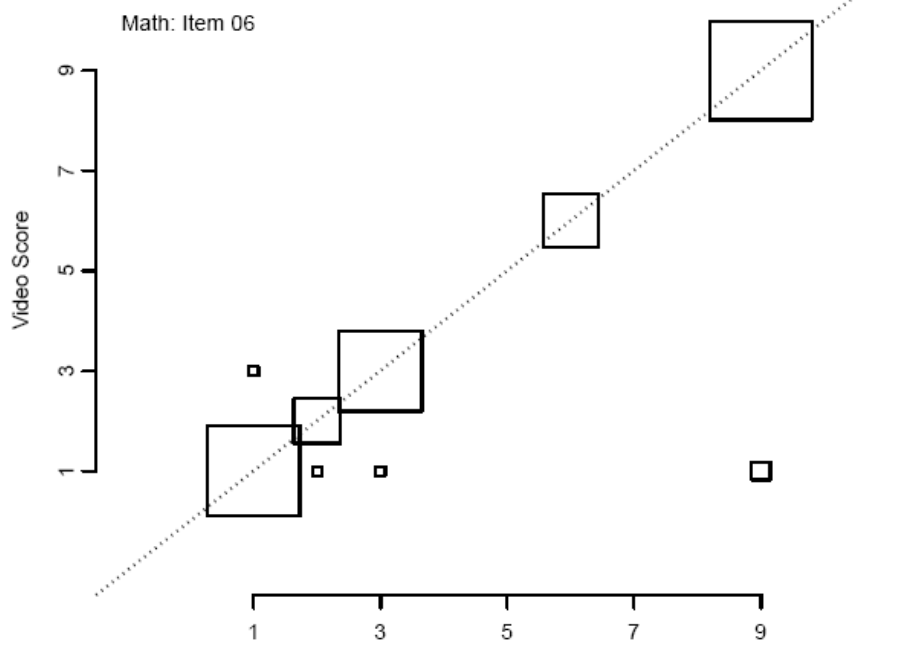
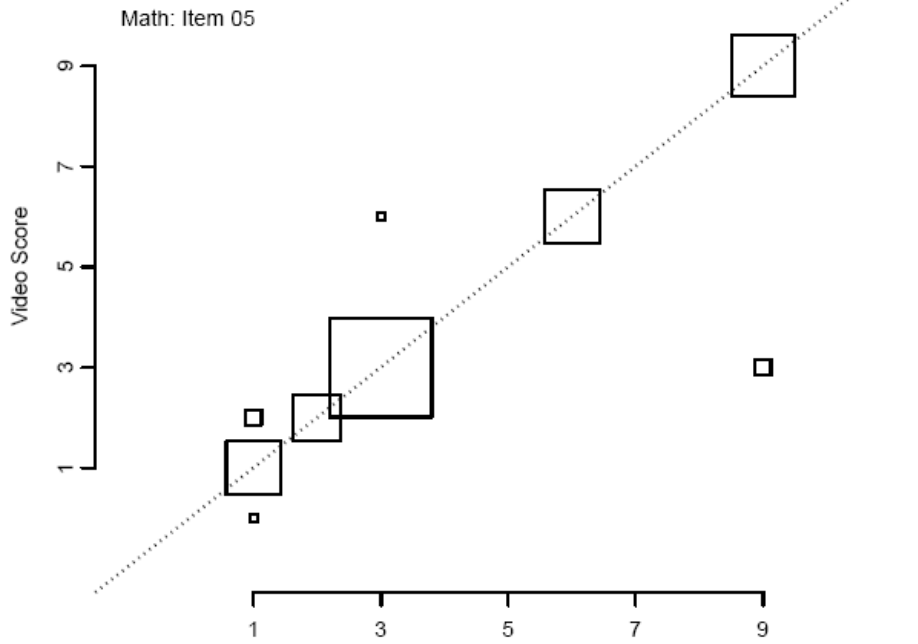
Notebook computers and headphones were placed on tables around the room to permit independent review of each video. Directly prior to scoring each video, panelists were asked to cross-check that the information on the DVD label, mode of communication form, and checklist matched. In addition, each panelist recorded his/her unique ID number on the lower right-hand corner of the checklist, took out the relevant test booklet for the grade, and reviewed the student mode of communication form prior to scoring.

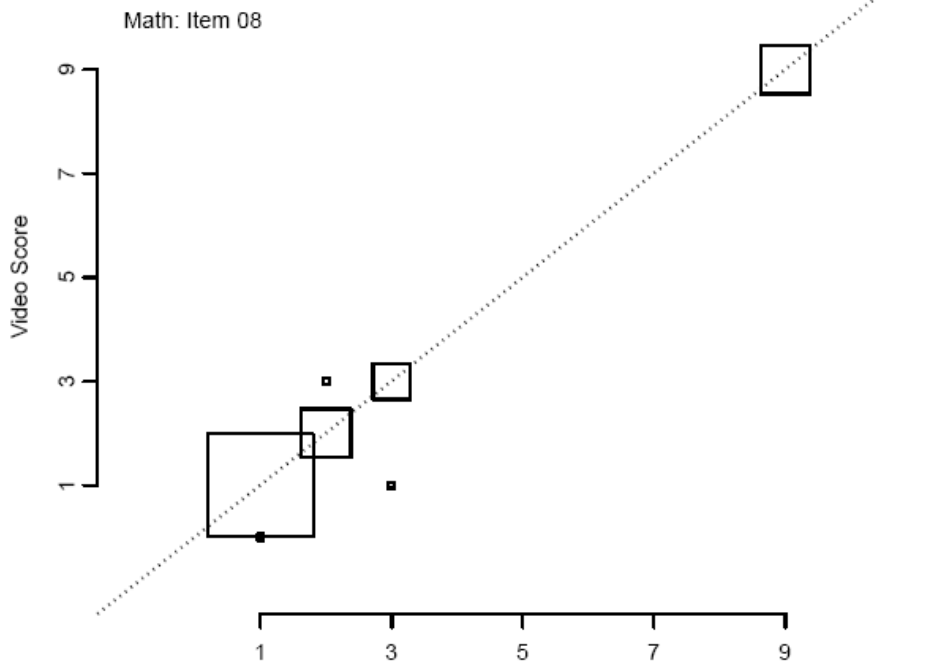
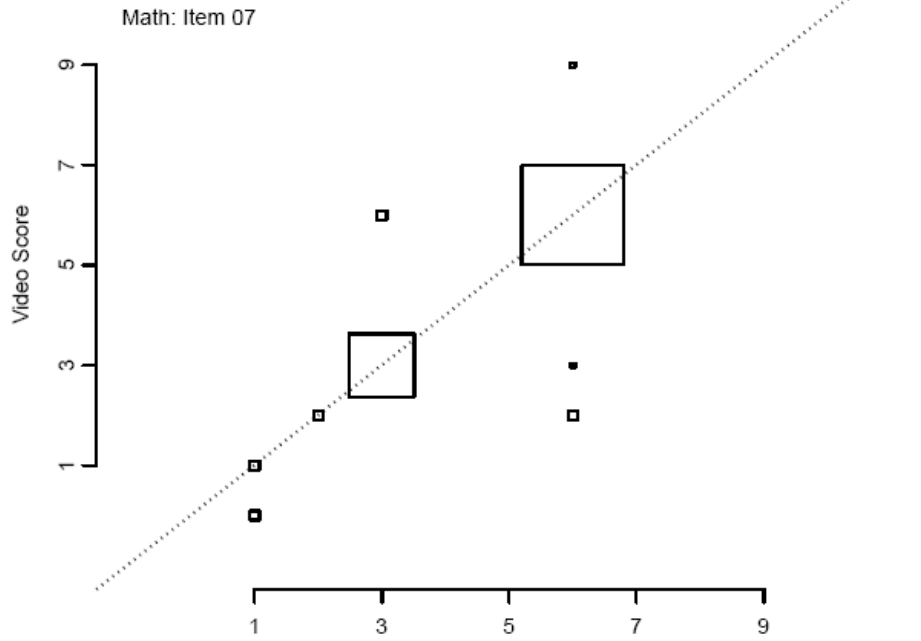
Upon completion of scoring each video, panelists returned the DVD, student scan sheet, mode of communication form, and comment form. The process then started again once another DVD was provided. Throughout the day panelists reviewed a mix of grade 5 mathematics and grade 10 writing videos.

**Appendix D: Video Scoring and Administration Rating Study
Teacher-assigned vs. Video Score Graphs by Item**

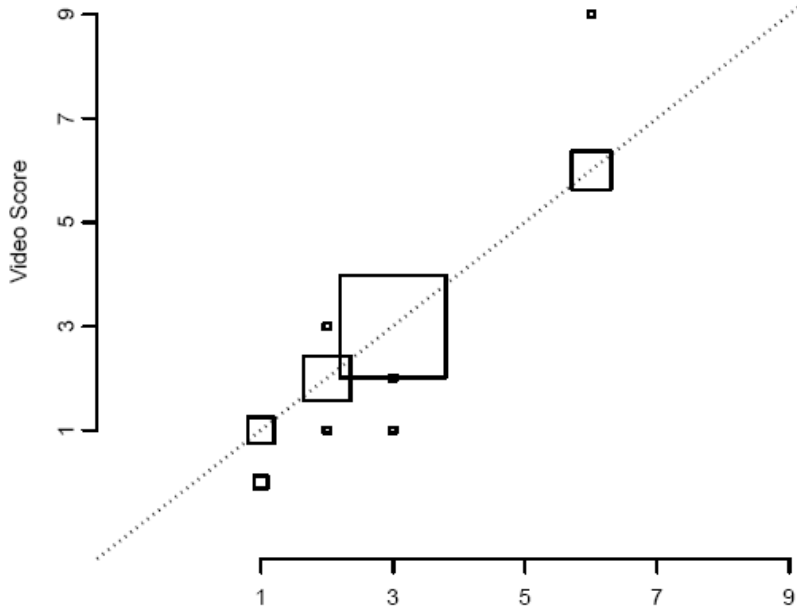




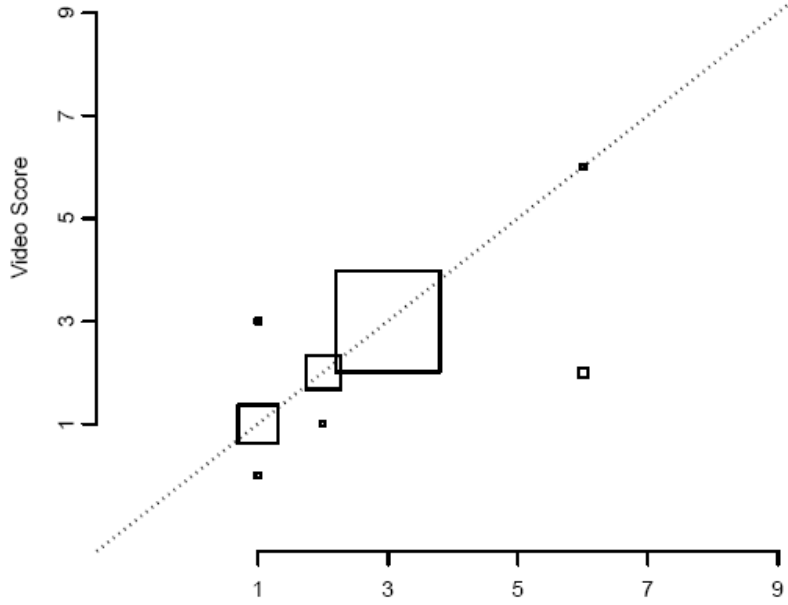


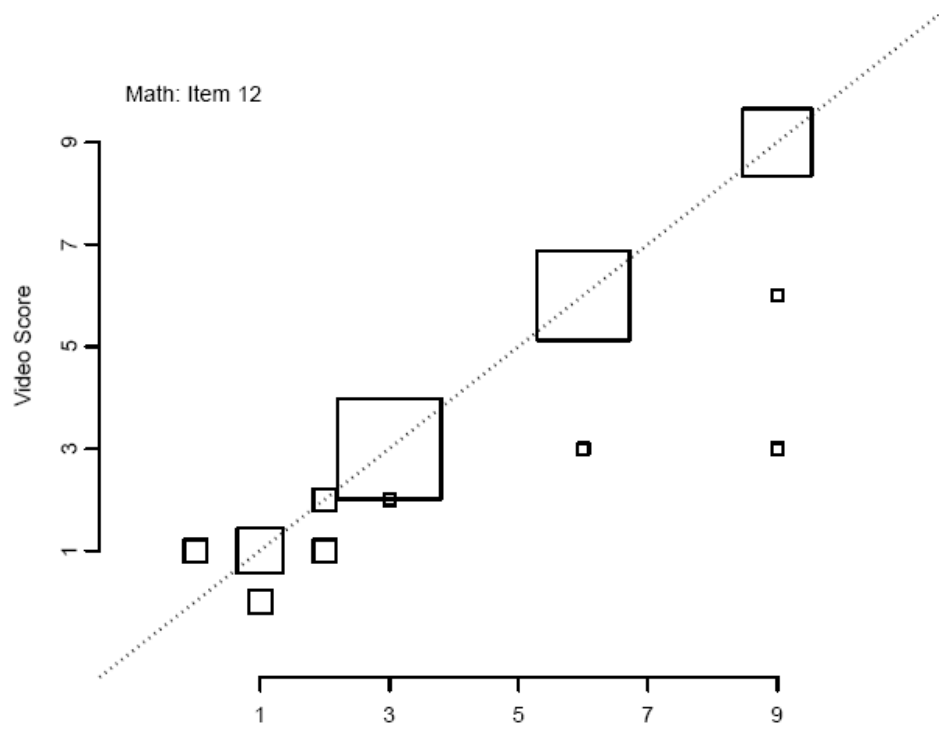
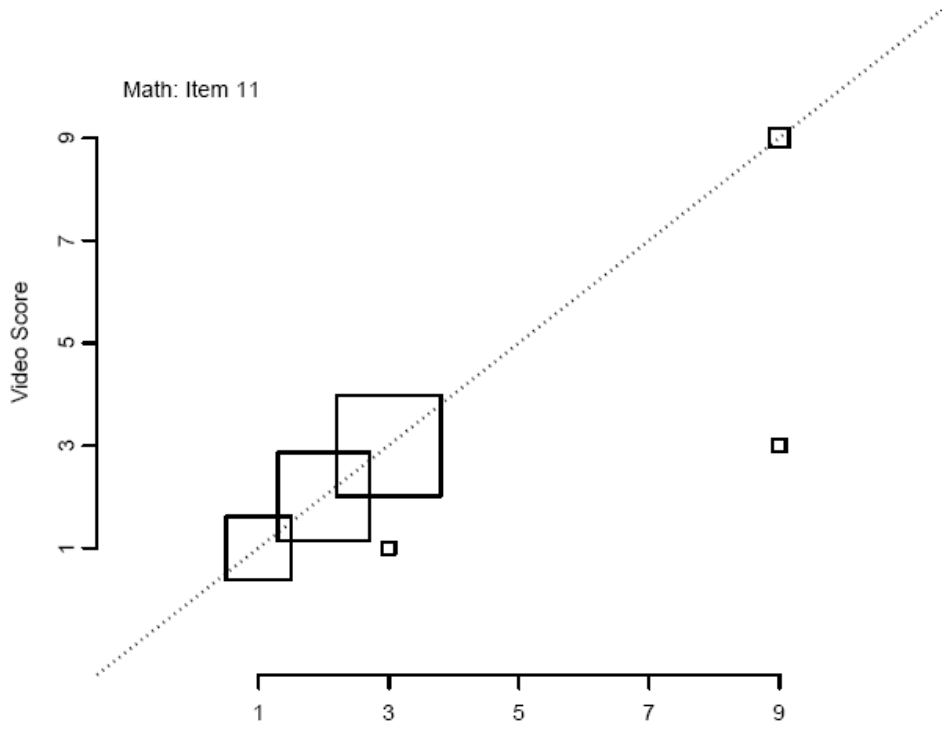


Math: Item 09

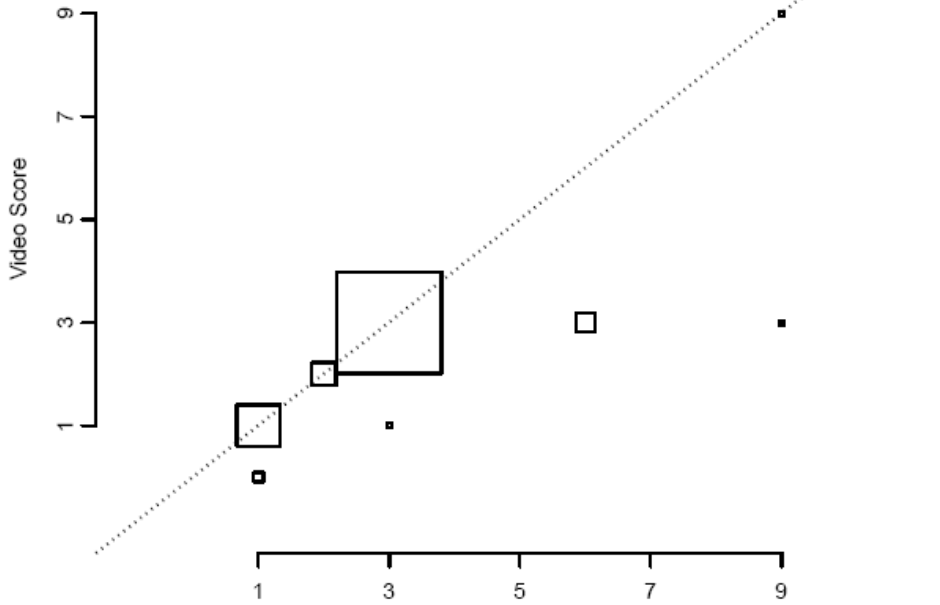


Math: Item 10

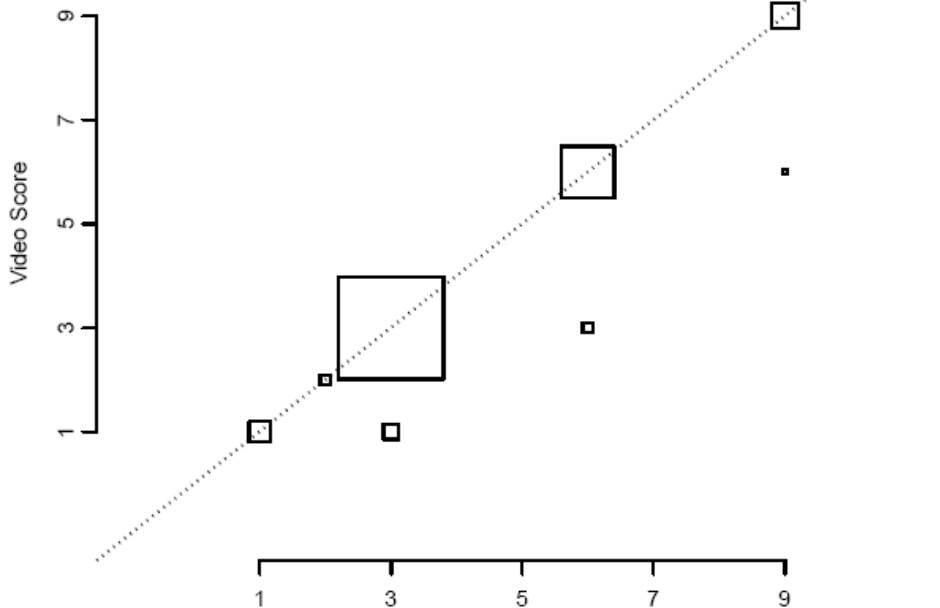


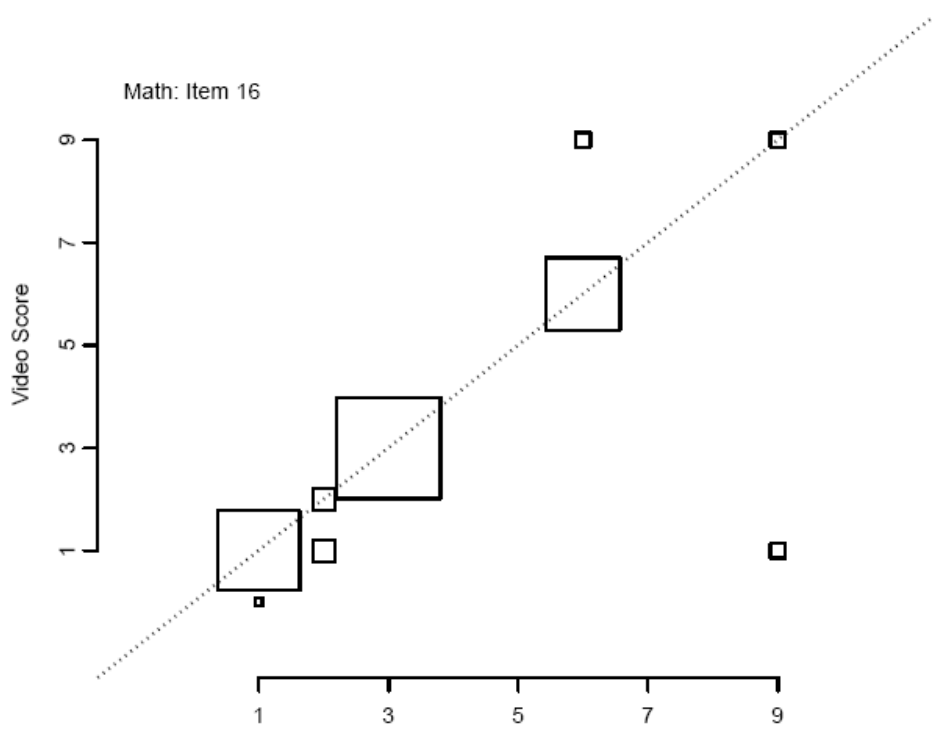
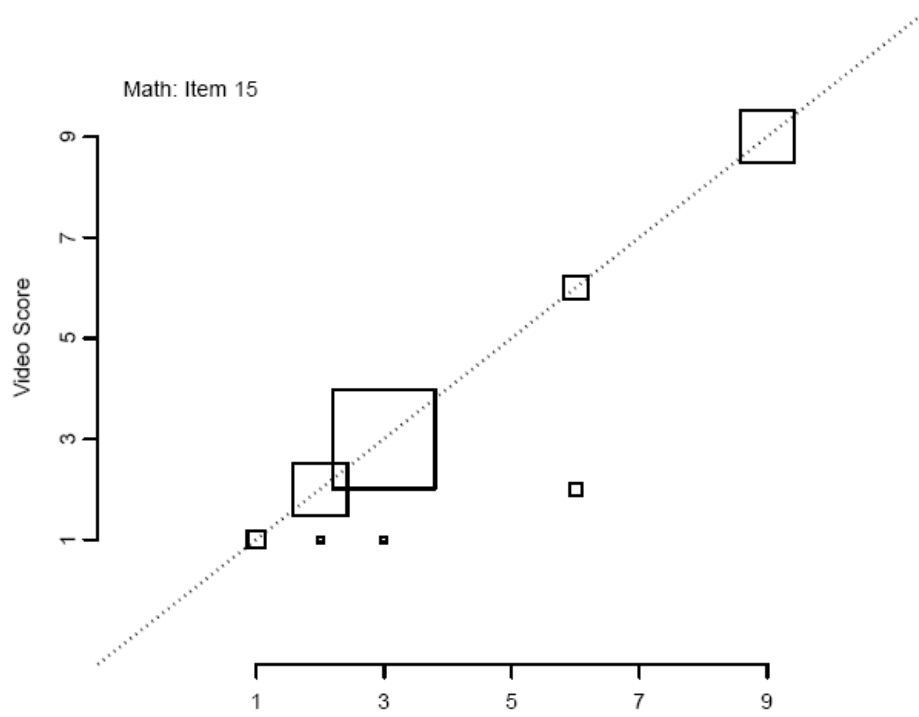


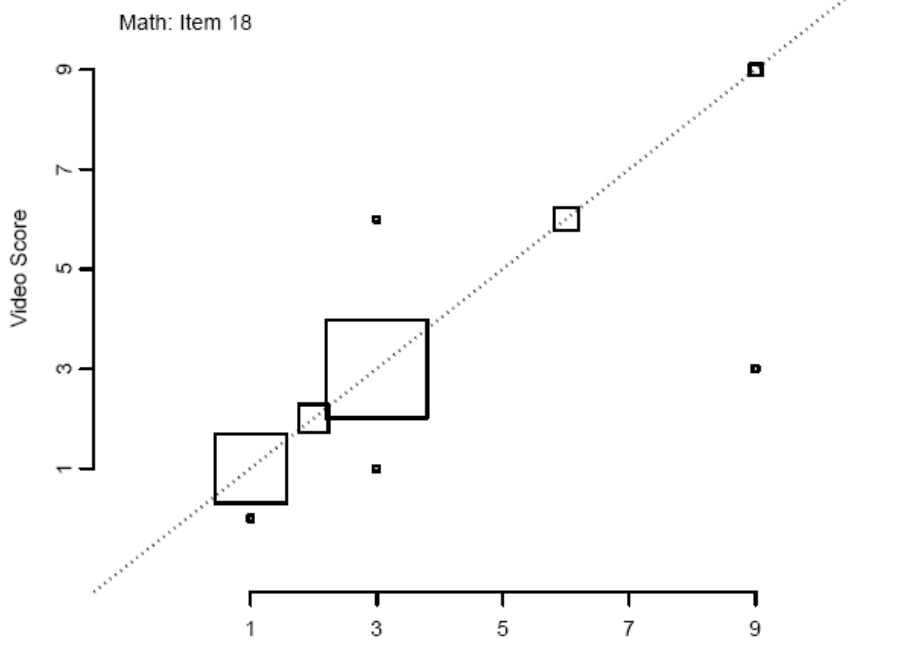
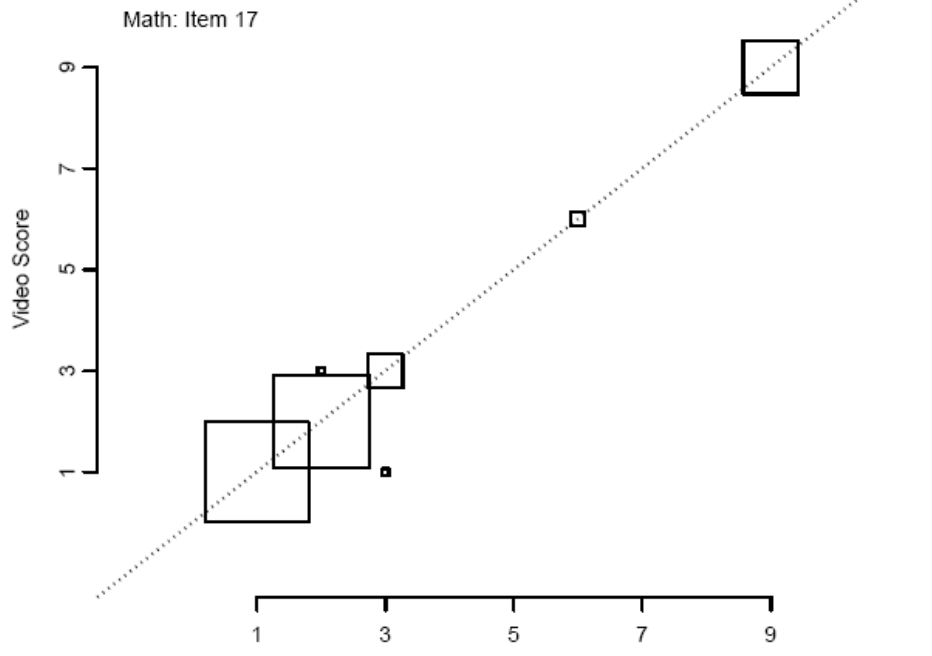
Math: Item 13

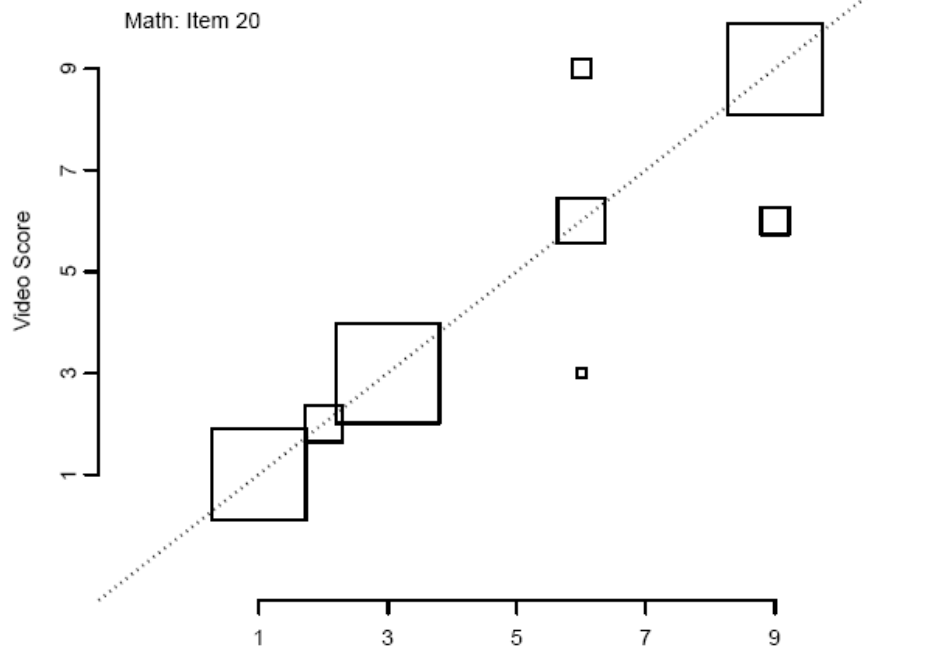
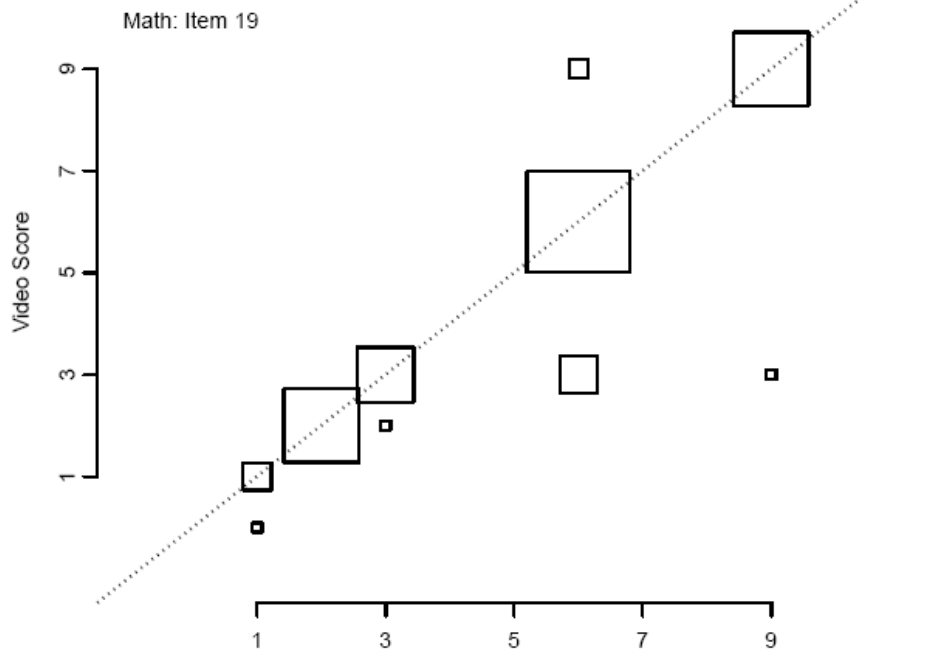


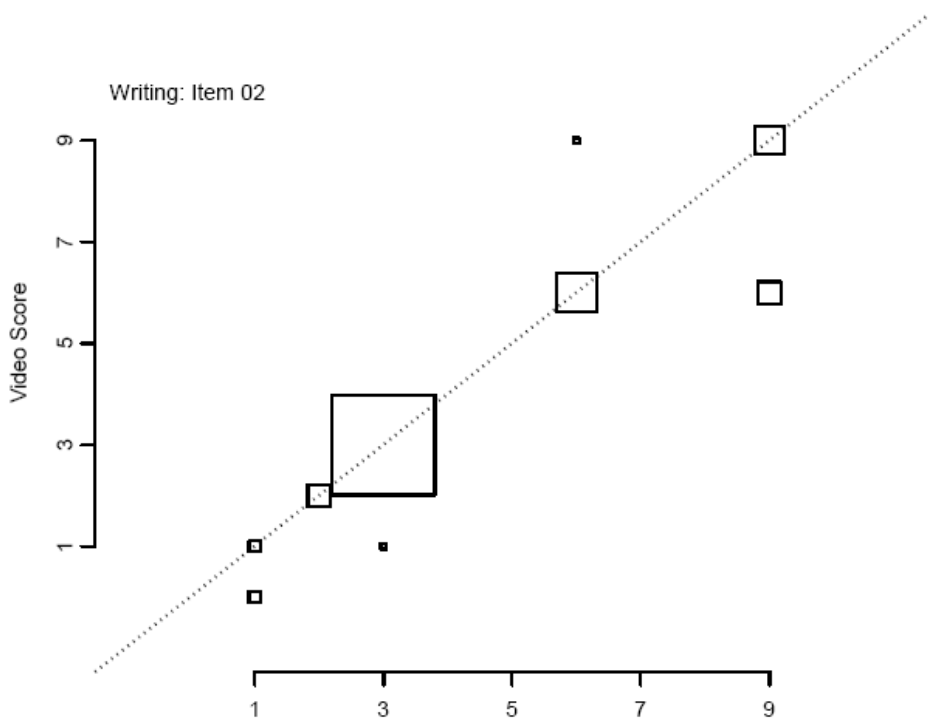
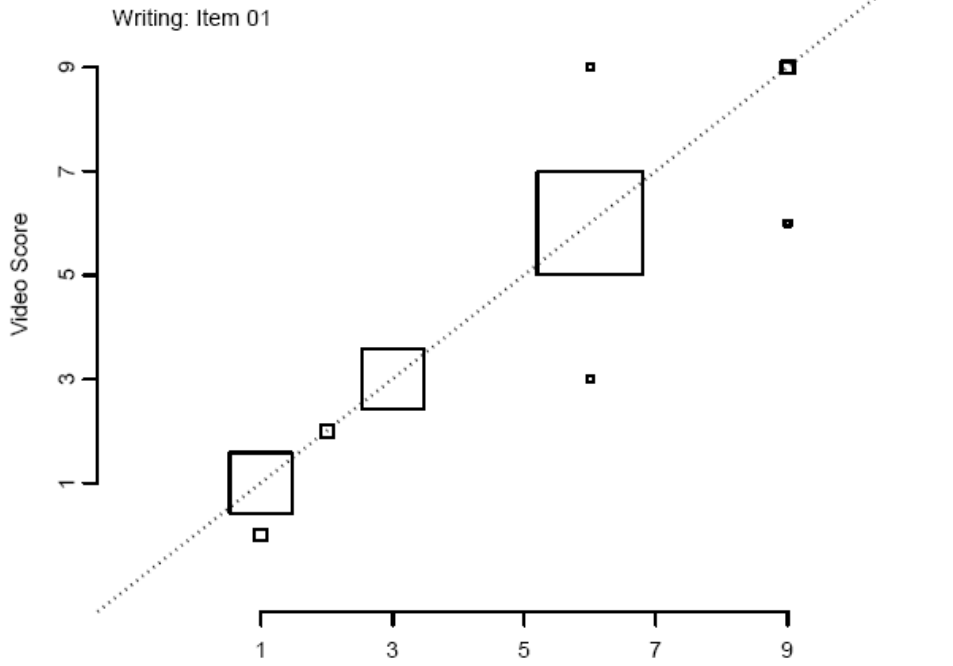
Math: Item 14

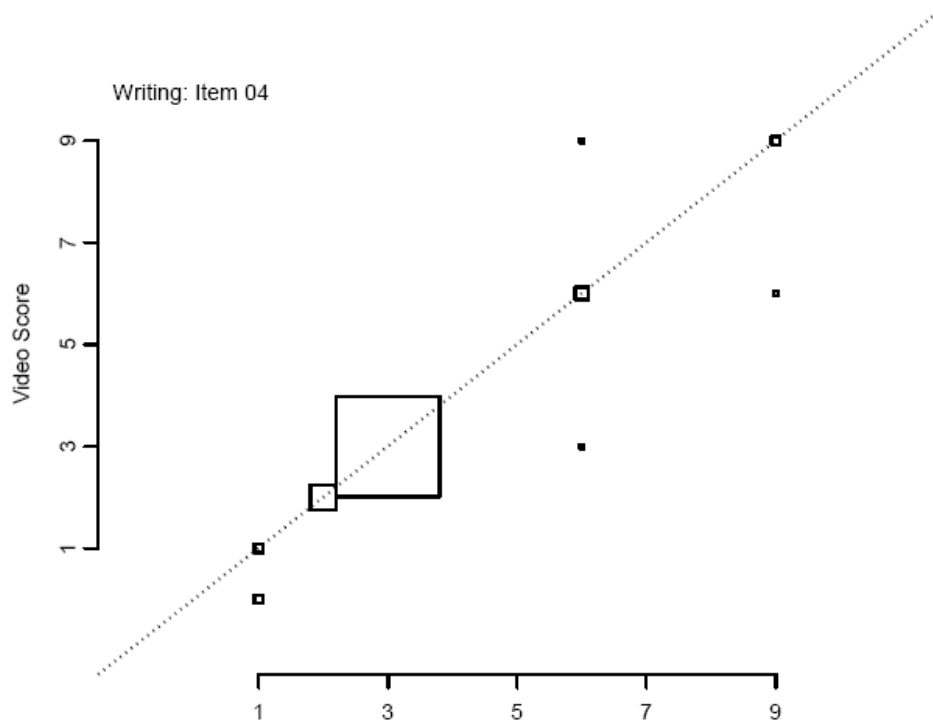
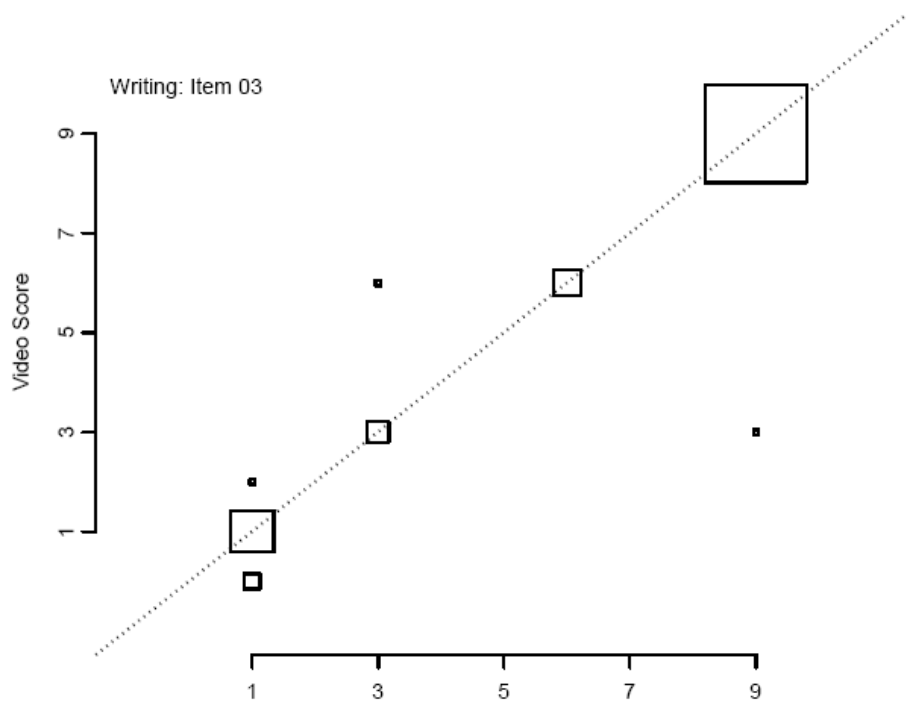


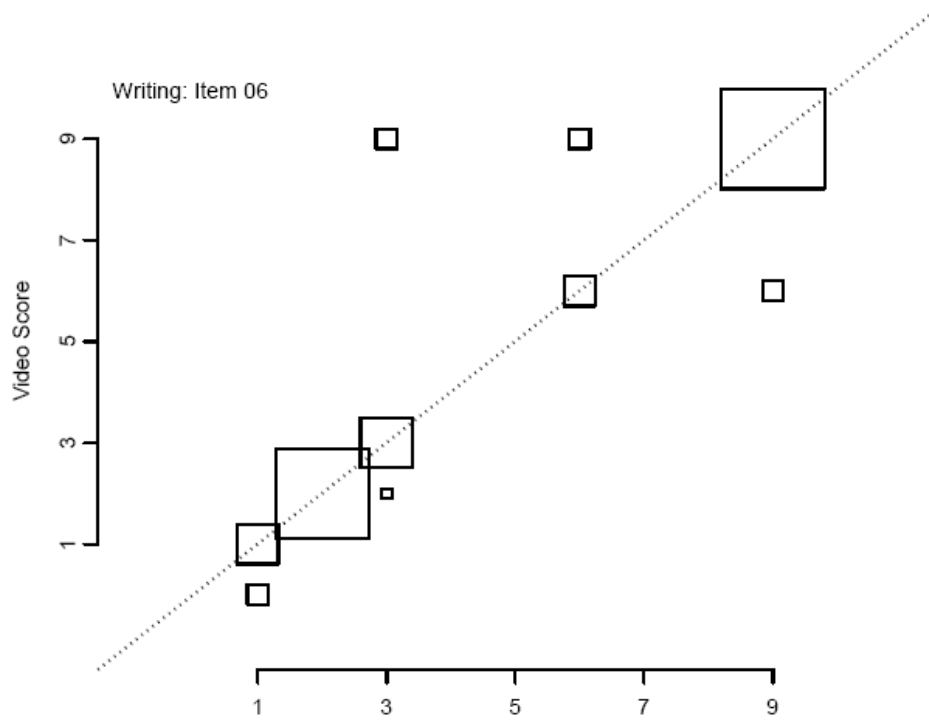
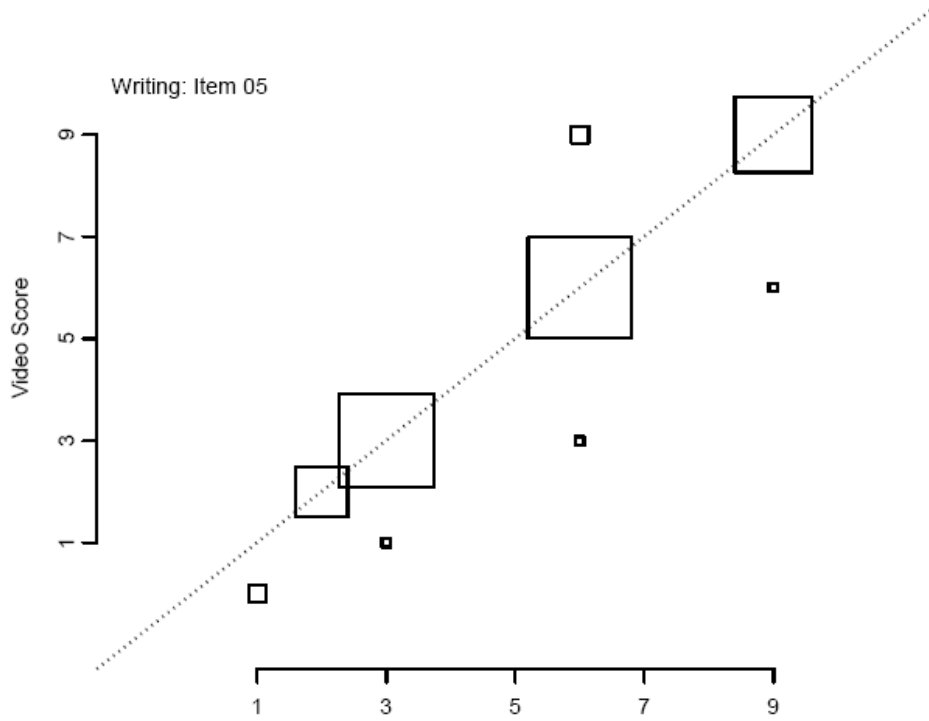


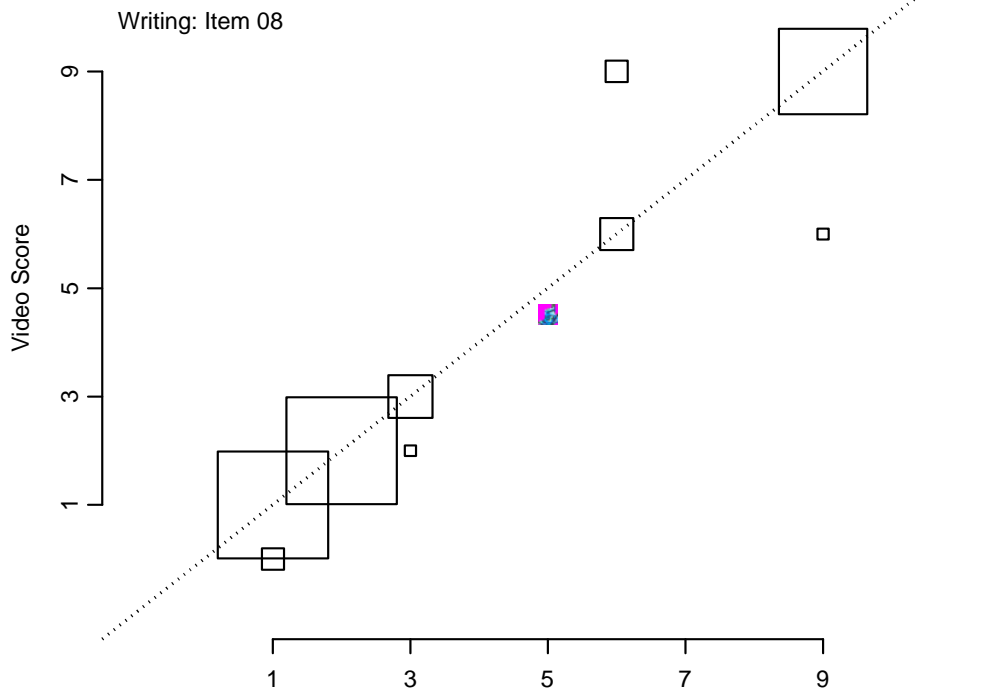
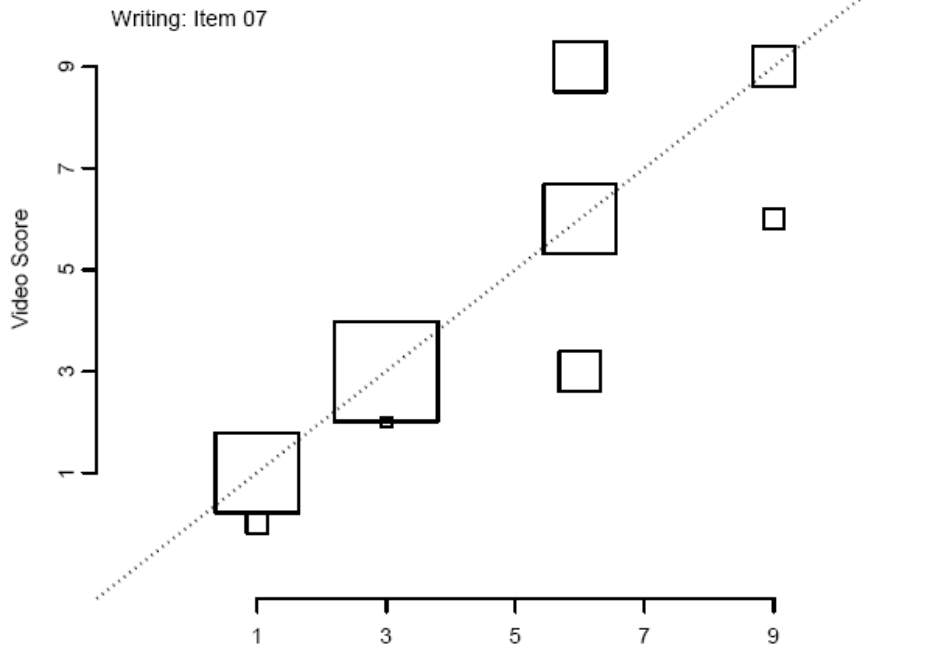


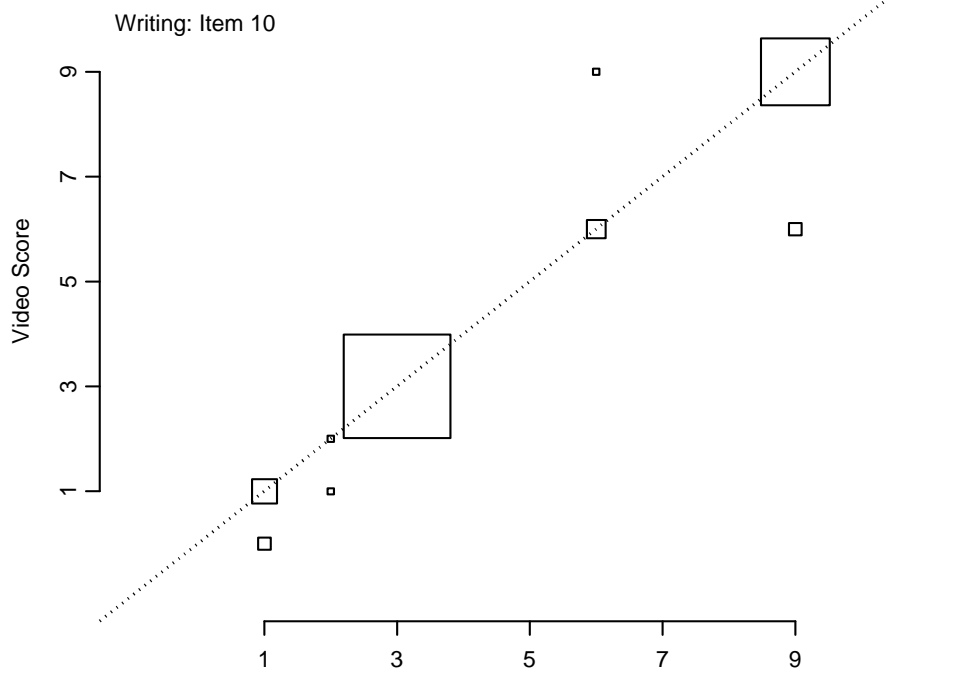
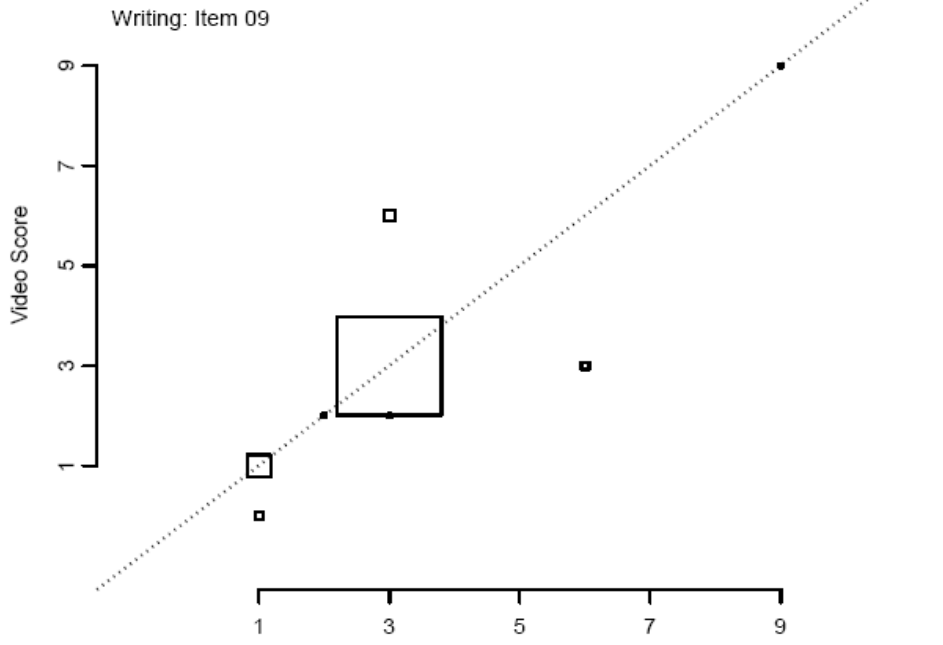


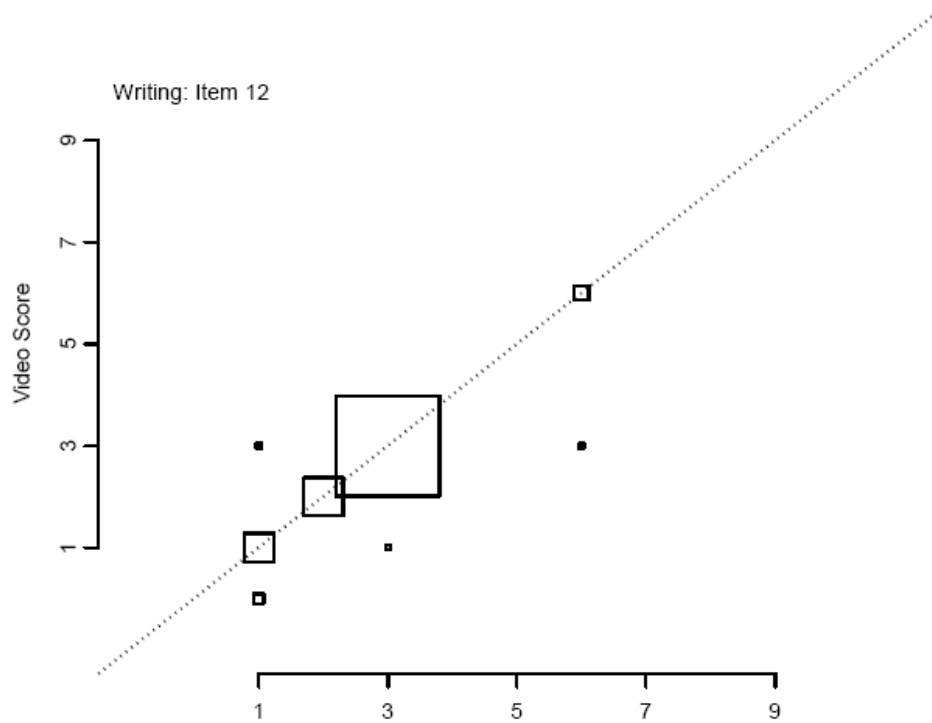
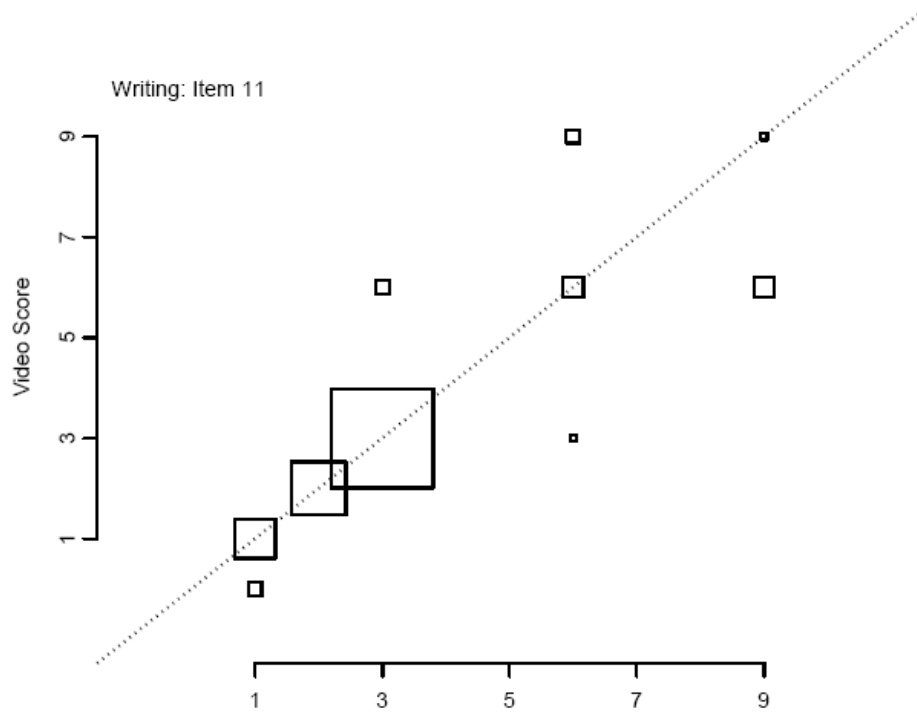


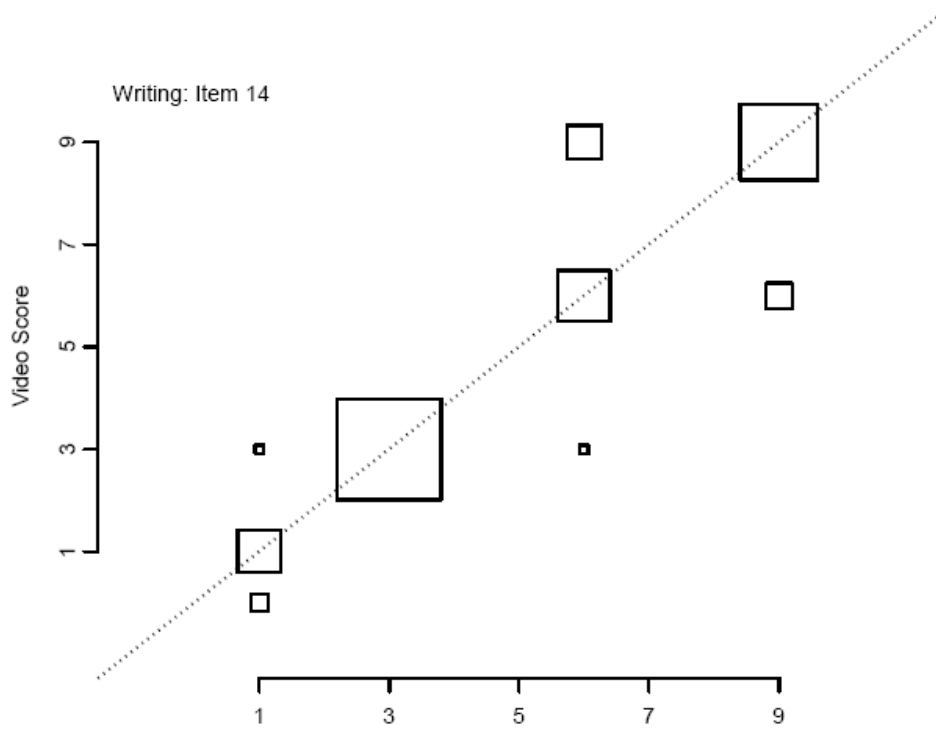
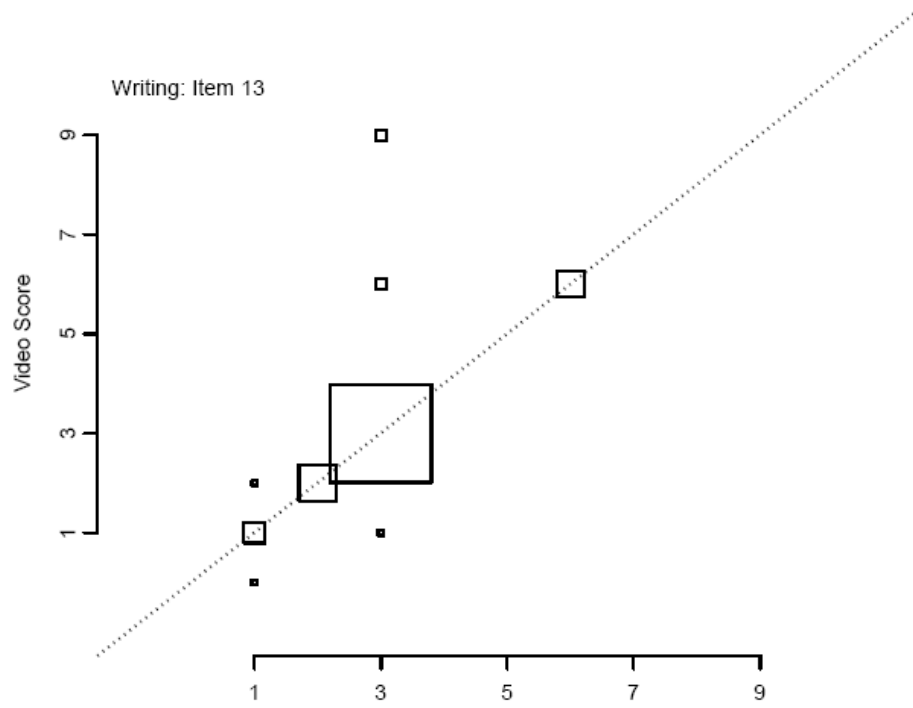


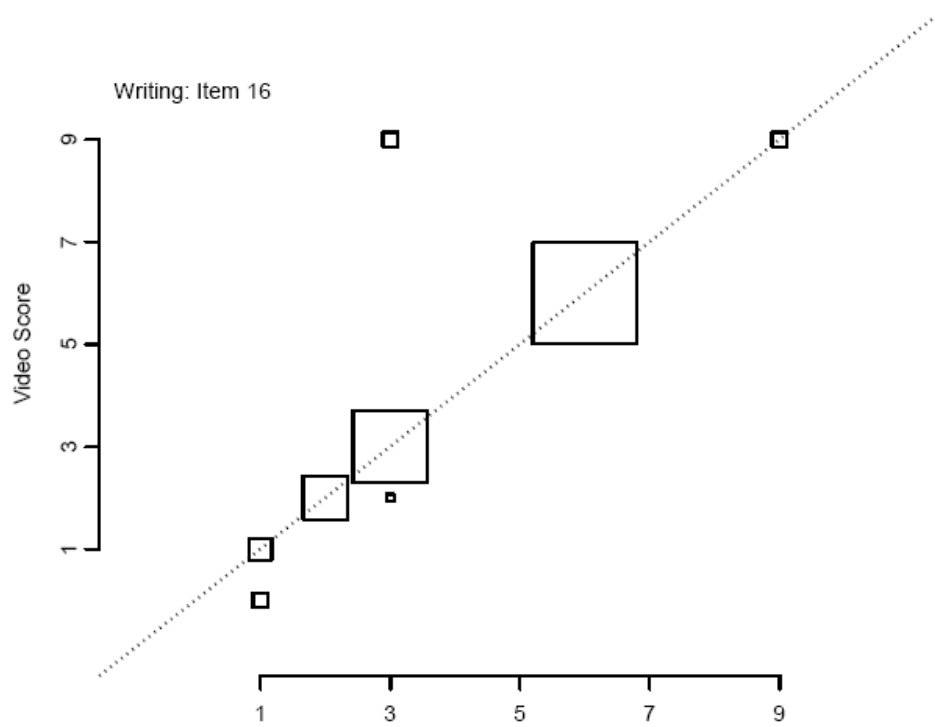
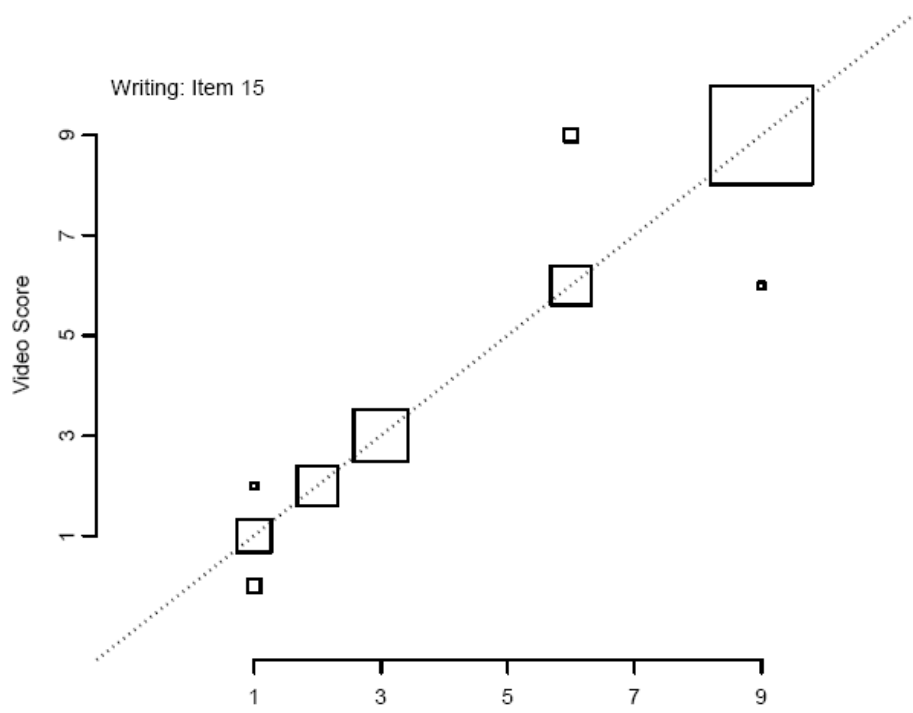


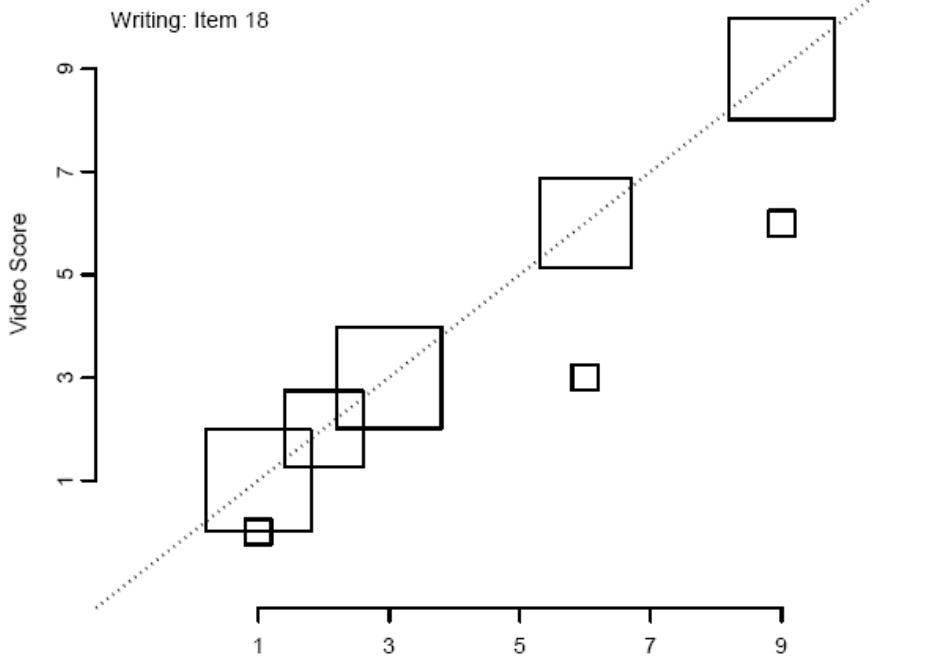
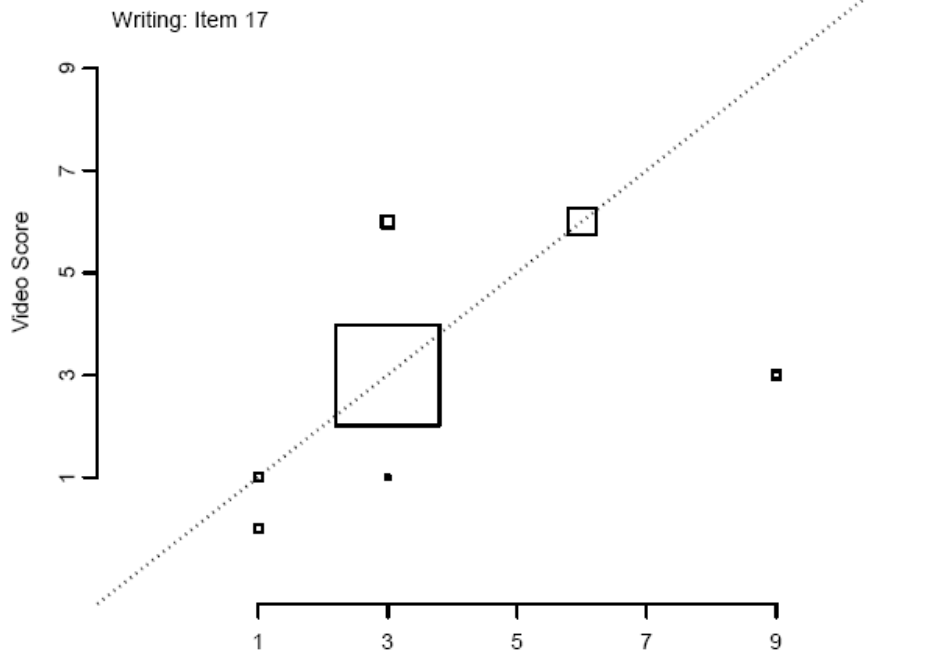


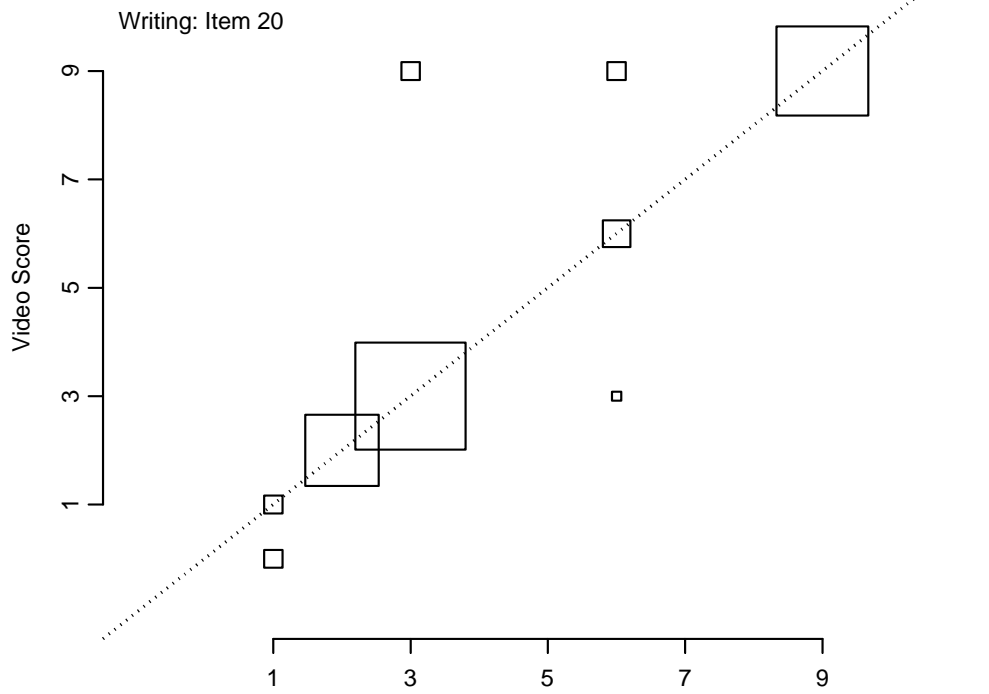
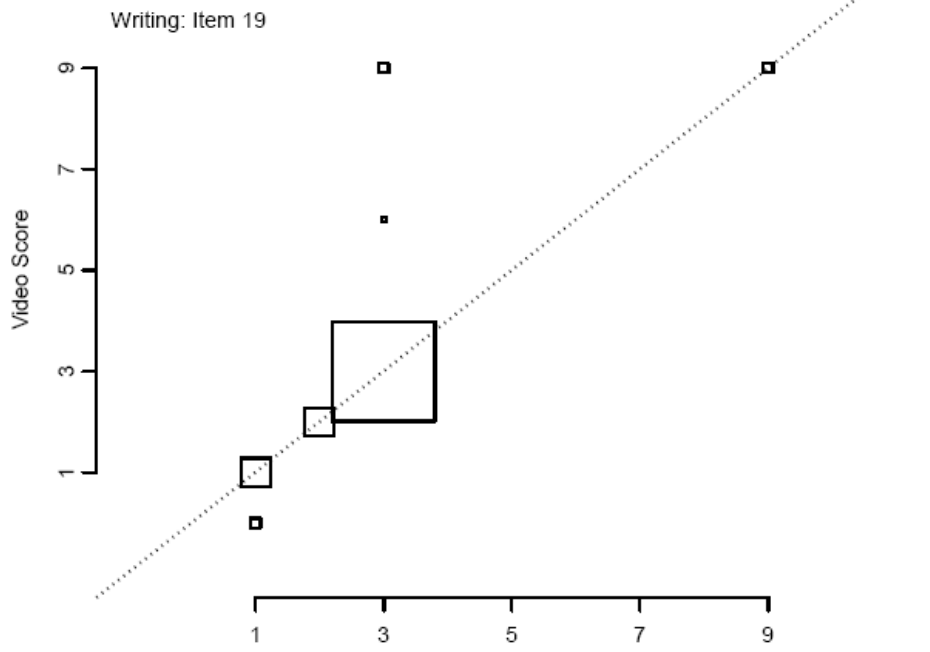












**Appendix E: Test-Retest Study
Comparison of Original and Retest Scores by Item**

**Comparison of Original and Retest Scores by Item
Mathematics Grade 5**

Item	Original Score	N	Retest Score							
			Participatory				Supported		Independent	
			0	1	2	3	0	6	0	9
1	0	6	67	33	0	0	100	0	100	0
	1	16	0	56	38	6	100	0	100	0
	2	11	9	18	55	18	91	9	82	18
	3	8	0	50	25	25	100	0	100	0
	6	0								
9	0									
2	0	6	67	33	0	0	100	0	100	0
	1	17	0	82	12	6	94	6	100	0
	2	13	0	62	15	23	85	15	100	0
	3	5	0	20	20	60	80	20	100	0
	6	0								
9	0									
3	0	3	100	0	0	0	100	0	100	0
	1	17	24	53	18	6	88	12	100	0
	2	8	0	75	0	25	75	25	75	25
	3	11	18	36	18	27	100	0	82	18
	6	2	0	0	50	50	50	50	100	0
9	0									
4	0	6	83	17	0	0	100	0	100	0
	1	17	12	59	24	6	94	6	100	0
	2	7	0	0	43	57	57	43	100	0
	3	10	10	40	20	30	90	10	90	10
	6	1	0	0	0	100	0	100	100	0
9	0									
5	0	3	100	0	0	0	100	0	100	0
	1	19	16	63	21	0	95	5	84	16
	2	11	18	18	9	55	100	0	82	18
	3	7	0	43	29	29	86	14	100	0
	6	1	0	0	0	100	100	0	100	0
9	0									
6	0	5	80	20	0	0	100	0	100	0
	1	27	7	41	22	30	85	15	85	15
	2	5	0	20	20	60	100	0	100	0
	3	3	0	67	0	33	100	0	67	33
	6	1	0	0	100	0	100	0	100	0
9	0									
7	0	5	80	0	20	0	100	0	100	0
	1	20	0	80	10	10	95	5	100	0
	2	7	0	43	43	14	86	14	100	0
	3	5	20	20	20	40	60	40	80	20
	6	4	0	50	25	25	50	50	100	0
9	0									
8	0	8	88	13	0	0	100	0	100	0
	1	20	10	75	15	0	100	0	85	15
	2	9	0	67	33	0	100	0	100	0
	3	4	25	50	0	25	100	0	100	0
	6	0								
9	0									

continued

**Comparison of Original and Retest Scores by Item
Mathematics Grade 5**

Item	Original Score	N	Retest Score							
			Participatory				Supported		Independent	
			0	1	2	3	0	6	0	9
9	0	6	83	17	0	0	100	0	100	0
	1	19	11	68	5	16	100	0	100	0
	2	7	0	29	29	43	100	0	86	14
	3	9	0	22	11	67	89	11	100	0
	6 9	0 0								
10	0	4	75	0	25	0	100	0	100	0
	1	19	16	53	21	11	95	5	100	0
	2	9	11	22	22	44	89	11	100	0
	3	8	0	0	25	75	88	13	100	0
	6 9	1 0	0 0	0 0	0 100		0 100		100 0	0 0
11	0	6	83	17	0	0	100	0	100	0
	1	17	12	41	41	6	100	0	94	6
	2	10	0	20	50	30	100	0	100	0
	3	7	0	43	0	57	100	0	86	14
	6 9	1 0	0 0	0 0	0 100		100 0		100 0	0 0
12	0	5	80	20	0	0	100	0	100	0
	1	17	18	53	12	18	100	0	94	6
	2	12	8	25	33	33	92	8	83	17
	3	6	0	17	50	33	100	0	100	0
	6 9	1 0	0 0	0 0	0 100		100 0		0 100	
13	0	7	71	29	0	0	100	0	100	0
	1	18	11	67	11	11	100	0	100	0
	2	9	0	22	33	44	100	0	100	0
	3	7	0	43	0	57	100	0	100	0
	6 9	0 0								
14	0	3	100	0	0	0	100	0	100	0
	1	21	24	57	5	14	95	5	95	5
	2	7	0	43	14	43	57	43	100	0
	3	7	0	29	29	43	71	29	100	0
	6 9	3 0	0 0	0 100			100 0		100 0	0 0
15	0	4	75	25	0	0	100	0	100	0
	1	14	14	64	7	14	100	0	86	14
	2	8	0	50	25	25	88	13	88	13
	3	14	0	43	14	43	100	0	93	7
	6 9	1 0	0 0	0 100	0 0	0 0	100 0	0 0	100 0	0 0
16	0	10	60	20	20	0	100	0	100	0
	1	20	5	60	20	15	95	5	85	15
	2	7	0	57	14	29	86	14	86	14
	3	4	0	50	25	25	100	0	100	0
	6 9	0 0								

continued

**Comparison of Original and Retest Scores by Item
Mathematics Grade 5**

Item	Original Score	N	Retest Score							
			Participatory				Supported		Independent	
			0	1	2	3	0	6	0	9
17	0	7	71	29	0	0	100	0	100	0
	1	19	16	53	16	16	84	16	89	11
	2	8	0	50	38	13	88	13	88	13
	3	3	0	33	0	67	100	0	67	33
	6	2	0	50	50	0	100	0	100	0
9	2	0	0	50	50	50	50	100	0	
18	0	5	80	20	0	0	100	0	100	0
	1	18	17	56	22	6	94	6	89	11
	2	6	0	50	33	17	100	0	100	0
	3	9	0	44	33	22	100	0	100	0
	6	2	0	50	50	0	50	50	100	0
9	1	0	0	0	100	100	0	100	0	
19	0	5	80	0	0	20	100	0	100	0
	1	18	17	72	6	6	100	0	100	0
	2	6	17	0	67	17	100	0	100	0
	3	9	0	0	44	56	89	11	89	11
	6	0								
9	3	0	0	67	33	33	67	33	67	
20	0	6	67	17	17	0	100	0	100	0
	1	19	11	68	0	21	89	11	100	0
	2	9	0	44	22	33	89	11	100	0
	3	4	25	25	25	25	75	25	100	0
	6	2	0	0	50	50	50	50	100	0
9	1	0	0	0	100	100	0	100	0	

**Comparison of Original and Retest Scores by Item
Reading Grade 8**

Item	Original Score	N	Retest Score							
			Participatory				Supported		Independent	
			0	1	2	3	0	6	0	9
1	0	2	50	0	0	50	100	0	100	0
	1	21	14	62	14	10	95	5	100	0
	2	7	0	29	14	57	86	14	86	14
	3	9	0	11	33	56	100	0	100	0
	6	2	0	0	100	0	100	0	100	0
2	0	2	50	50	0	0	100	0	100	0
	1	22	5	77	5	14	100	0	91	9
	2	8	0	50	25	25	100	0	88	13
	3	12	0	17	42	42	92	8	83	17
	6	0	0	0	0	100	100	0	100	0
3	0	3	67	33	0	0	100	0	100	0
	1	23	4	78	9	9	96	4	91	9
	2	13	0	46	46	8	77	23	92	8
	3	2	0	0	50	50	50	50	50	50
	6	3	0	0	67	33	67	33	67	33
4	0	3	33	0	33	33	67	33	100	0
	1	19	5	79	5	11	89	11	100	0
	2	9	0	44	22	33	100	0	100	0
	3	11	0	45	36	18	91	9	100	0
	6	2	0	0	0	100	0	100	100	0
5	0	2	50	50	0	0	100	0	100	0
	1	23	17	48	26	9	96	4	100	0
	2	7	0	0	43	57	86	14	100	0
	3	11	0	9	73	18	91	9	100	0
	6	0	0	0	50	50	100	0	100	0
6	0	4	50	50	0	0	100	0	100	0
	1	27	4	70	15	11	96	4	96	4
	2	6	0	50	33	17	100	0	100	0
	3	5	0	20	60	20	80	20	80	20
	6	3	0	33	0	67	100	0	67	33
7	0	3	33	67	0	0	67	33	100	0
	1	29	10	52	17	21	93	7	93	7
	2	5	0	40	0	60	100	0	100	0
	3	4	25	0	75	0	100	0	100	0
	6	1	0	0	0	100	100	0	100	0
8	0	5	20	60	0	20	80	20	100	0
	1	19	11	74	11	5	95	5	100	0
	2	7	0	71	0	29	100	0	100	0
	3	9	0	11	44	44	78	22	100	0
	6	4	0	0	50	50	25	75	75	25
9	1	0	0	0	100	100	0	100	0	

continued

**Comparison of Original and Retest Scores by Item
Reading Grade 8**

Item	Original Score	N	Retest Score							
			Participatory				Supported		Independent	
			0	1	2	3	0	6	0	9
9	0	3	33	33	33	0	100	0	100	0
	1	25	16	64	12	8	92	8	100	0
	2	6	0	50	17	33	100	0	100	0
	3	11	0	36	18	45	91	9	82	18
	6 9	0 0								
10	0	4	50	25	0	25	100	0	75	25
	1	28	25	50	18	7	93	7	100	0
	2	5	0	60	40	0	100	0	100	0
	3	7	14	29	0	57	71	29	100	0
	6 9	1 0	0 0	0 0	0 0	100	100	0	100	0
11	0	3	33	33	33	0	100	0	100	0
	1	31	13	74	10	3	94	6	97	3
	2	6	0	50	17	33	100	0	100	0
	3	4	0	0	75	25	75	25	100	0
	6 9	1 0	0 0	0 0	0 0	100	100	0	100	0
12	0	5	20	80	0	0	100	0	100	0
	1	23	13	70	9	9	96	4	100	0
	2	6	0	50	50	0	100	0	83	17
	3	10	0	50	10	40	80	20	90	10
	6 9	1 0	0 0	0 0	100	0	100	0	100	0
13	0	4	25	50	0	25	75	25	100	0
	1	27	15	59	15	11	96	4	96	4
	2	6	0	50	33	17	100	0	83	17
	3	4	0	0	0	100	100	0	100	0
	6 9	4 0	0 0	25 0	50 0	25	50	50	100	0
14	0	1	100	0	0	0	100	0	100	0
	1	31	13	55	23	10	100	0	100	0
	2	5	0	60	40	0	100	0	100	0
	3	8	0	25	38	38	88	13	88	13
	6 9	0 0								
15	0	4	25	50	25	0	100	0	100	0
	1	20	10	60	10	20	95	5	100	0
	2	11	0	45	36	18	100	0	100	0
	3	8	0	13	13	75	63	38	75	25
	6 9	1 1	0 0	0 0	100 100	0 0	0 100	100	100	0 0
16	0	5	20	60	0	20	100	0	100	0
	1	27	7	56	22	15	96	4	100	0
	2	5	0	20	20	60	80	20	80	20
	3	7	0	14	14	71	100	0	86	14
	6 9	1 0	0 0	100 0	0 0	0	100	0	100	0

continued

**Comparison of Original and Retest Scores by Item
Reading Grade 8**

Item	Original Score	N	Retest Score							
			Participatory				Supported		Independent	
			0	1	2	3	0	6	0	9
17	0	5	20	60	0	20	100	0	100	0
	1	24	8	67	13	13	100	0	96	4
	2	11	9	45	36	9	82	18	91	9
	3	5	0	20	40	40	100	0	100	0
	6 9	0 0								
18	0	5	20	80	0	0	100	0	100	0
	1	23	13	70	13	4	100	0	100	0
	2	11	0	45	27	27	82	18	91	9
	3	5	0	20	40	40	60	40	80	20
	6 9	1 0	0	0	0	100	100	0	100	0
19	0	3	67	33	0	0	67	33	100	0
	1	26	12	69	4	15	96	4	100	0
	2	6	0	33	67	0	83	17	100	0
	3	10	0	10	40	50	100	0	100	0
	6 9	0 0								
20	0	4	50	50	0	0	100	0	75	25
	1	25	16	64	8	12	100	0	100	0
	2	6	0	83	17	0	83	17	83	17
	3	9	0	44	22	33	89	11	89	11
	6 9	1 0	0	100	0	0	0	100	100	0