Longitudinal/Cross-Sectional Study of the Impact of *Mathematics in Context* on Student Performance

Interrater Reliability at the 1998–1999 Scoring Institutes

(Working Paper #22)

Lorene Folgert and Mary Shafer

University of Wisconsin-Madison

September 2001

Folgert, L., & Shafer, M. C. (2001). *Interrater Reliability at the 1998-1999 Scoring Institutes* (*Mathematics in Context* Longitudinal/Cross-Sectional Study Working Paper No. 22). Madison, WI: University of Wisconsin–Madison.

The research reported in this paper was supported in part by the National Science Foundation #REC-9553889. The views expressed here are those of the authors and do not necessarily reflect the views of the funding agency.

INTRODUCTION

The purposes of the longitudinal/cross-sectional study of the impact of *Mathematics in Context* (MiC; National Center for Research in Mathematical Sciences Education & Freudenthal Institute, 1997–1998) on student performance are (a) to determine the mathematical knowledge, understanding, attitudes, and levels of student performance as a consequence of studying MiC for over three years; and (b) to compare student knowledge, understanding, attitudes, and levels of performance of students using MiC with those using conventional mathematics curricula. The research model for this study is an adaptation of a structural model for monitoring changes in school mathematics (Romberg, 1987). For this study, information is being gathered on 14 variables over a 3-year period for three groups of students (those in Grades 6, 7, and 8 in 1998–1999). The variables have been organized in five categories (prior, independent, intervening, outcome, and consequent). (See Figure 1 for variables and hypothesized relationships.)

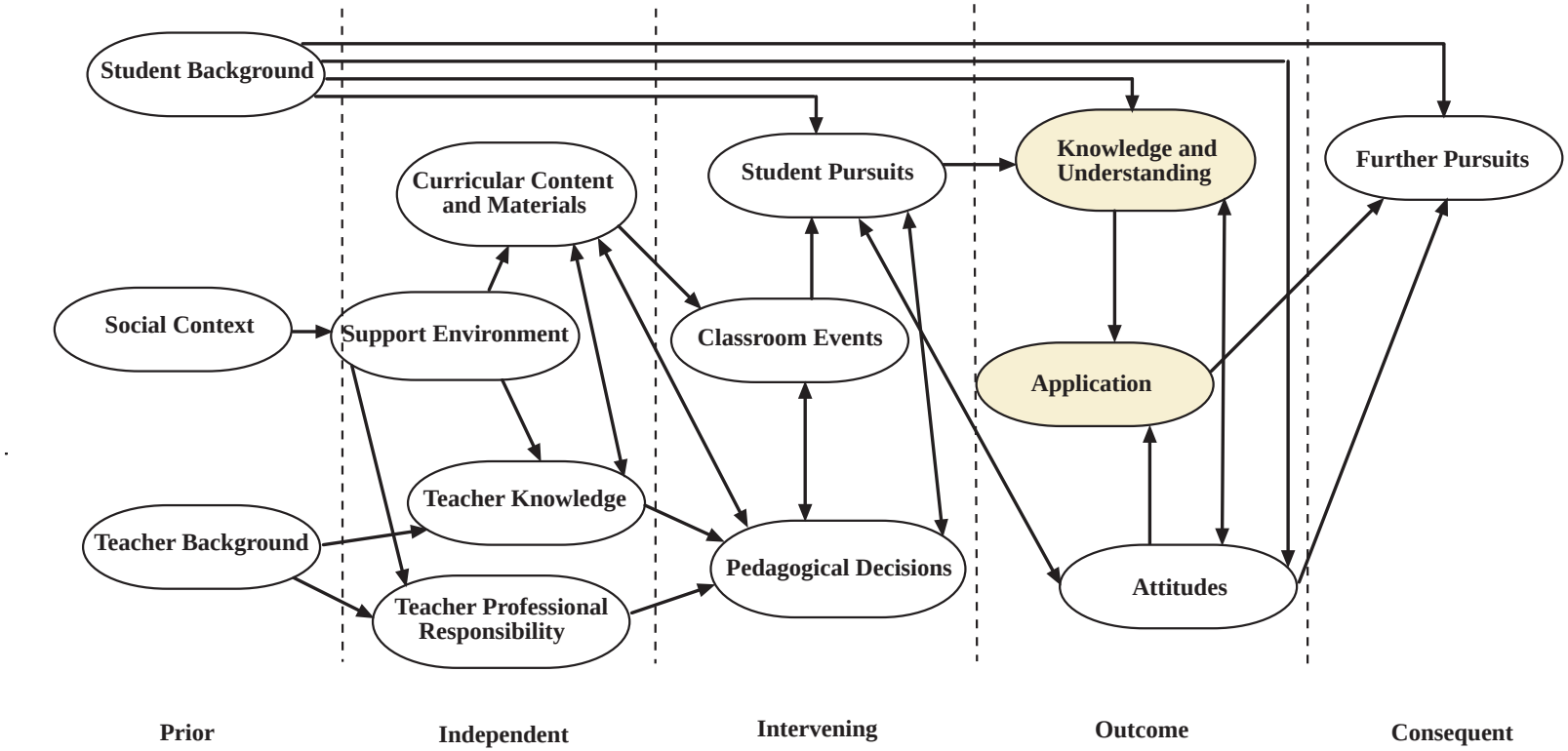

Figure 1. Revised model for the monitoring of school mathematics.

Interrater Reliability at the 1998-99 Scoring Institutes

Eight scoring institutes were held in 1999 to score the Problem Solving Assessments and External Assessments administered to sixth-, seventh-, and eighth-grade students in spring 1999 as part of the longitudinal/cross-sectional study. During the scoring institutes, each student response was scored by two raters who were experienced elementary- and middle-school teachers. Interrater reliability was calculated to assess the scoring procedure and the quality of the scoring. Interrater reliability is the frequency at which the two raters who scored a student response agreed with one another.[1] The purpose of this paper is to describe the scoring procedure at these scoring institutes, to summarize interrater reliability, and to report factors that influenced interrater reliability.

Problem Solving Assessments (PSAs) and External Assessments (EAs) were administered to all study students (Grades 6–8) in each of the four districts in the study. The first four scoring institutes, held in spring 1999, were conducted in the districts; study teachers were the raters, providing an opportunity for them to participate in the scoring process, learn about the assessments, and examine student work from a variety of teachers. Three more scoring institutes were held at the University of Wisconsin-Madison; teachers (Grades 3–12) from schools in the Madison area were the raters. The each of the three Madison institutes was week-long and was held in summer 1999.

The number of PSAs and EAs varied by the number of study students at each grade level and the number of absentees when the assessments were administered. The number of student assessments scored by grade level and type of assessment is summarized in Table 1.

Table 1
Problem Solving Assessments (PSAs) and External Assessments (EAs)Scored by Grade Level

Grade	Assessments	
	PSA *(N)*	EA *(N)*
6	688	713
7	825	810
8	503	446

[1] If there was a discrepancy between the scores, a third rater adjudicated. Occasionally more adjudications were necessary. When two raters agreed upon, it was considered final.

Assessments: Structure

Problem Solving Assessment System

The Problem Solving Assessment System is a set of grade-specific assessments composed of constructed-response items set in contexts. The number of items in each context varied depending on the mathematical content and level of reasoning assessed (see Figure 2). The PSA used 18 contexts, each of which was scored separately. PSA items examined students' application of mathematics and mathematical reasoning at three levels. Items designed to elicit reasoning at the second and third levels were more openended in nature and more complex to score. Partial-credit scoring rubrics were used to assign point values to student responses. Strategies students used in solving problems were also coded. Although scoring rubrics were prepared in advance of scoring, they evolved during the scoring process. As a result, some items of necessity were rescored at subsequent institutes.

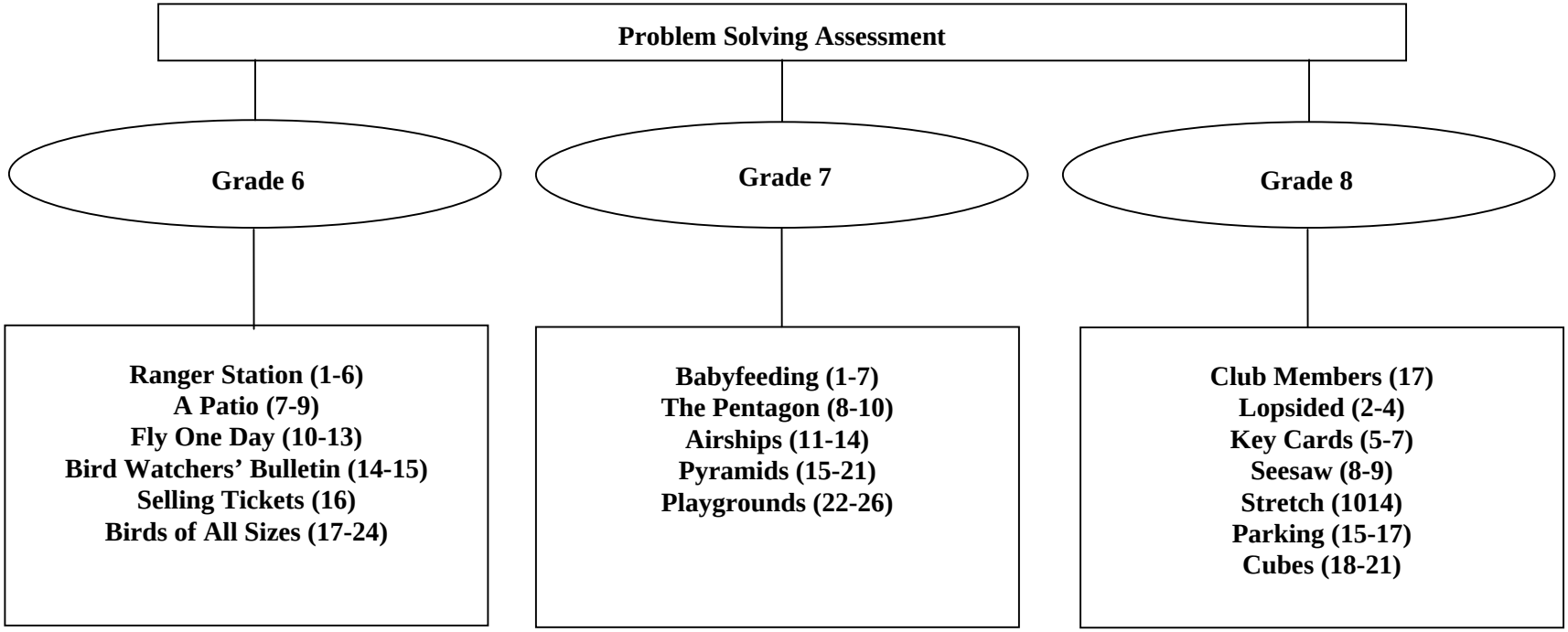

Figure 2. Contexts of the Grade 6, Grade 7, and Grade 8 Problem Solving Assessments.

External Assessment System

The External Assessment System is a set of grade-specific assessment composed of constructed-response and multiple-choice items from the National Assessment of Educational Progress (NAEP) and the Third International Mathematics and Science Study (TIMSS). In contrast to the PSA, six EA constructed-response anchor items were repeated on each grade-specific assessment. In addition, three other constructed-response items were scored (Contexts 7, 8, and 9). For purposes of scoring, each set of items was considered a context[2] (see Figure 3). The rubrics used in scoring EA items were identical to rubrics used in the NAEP and TIMSS assessments. Scoring involved assigning point scores (scoring for most contexts was based on partial-credit rubrics) and strategy codes when appropriate. Interrater reliability was determined only for point scores. In general, EA rubrics were less complicated than PSA rubrics, but, because these contexts involved anchor items repeated at each grade level in the study, in most cases, larger sets of assessments were scored for each EA context. (Multiple-choice items were also scored by two raters, but did not require analysis of interrater reliability.)

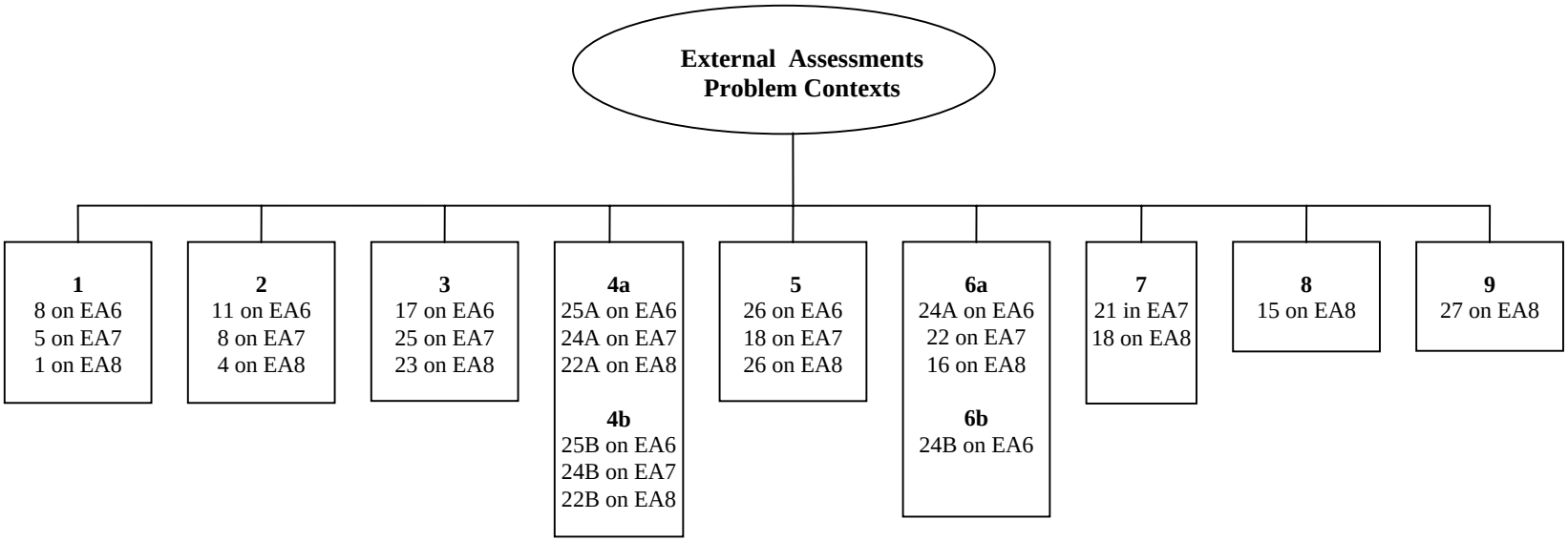

Figure 3. Constructed-Response Contexts of the Grade 6, Grade 7, and Grade 8 External Assessments

[2] The context numbers apply only to the context groupings in this paper. (The context groupings are numbered differently in the 1997-1998 and 1999-2000 Interrater Reliability papers since some of the items do not apply in the other papers.)

Assessments: Scoring Procedures

A total of 27 contexts on both assessments were scored. The number of contexts scored at each institute varied from 1 to 9 depending on the number of raters, the number of assessments, and the number of days the institute lasted. On average, two PSA or EA contexts were scored each day.

Preparation Prior to Scoring Institutes

To assure anonymity of students, teachers, and districts, names were removed from all student assessments and student scratch papers. At the district scoring institutes, assessments from the different schools and classes were mixed randomly; at the remaining institutes, assessments from different districts were mixed randomly. Assessments were separated into packets of 5–8 assessments, and each packet was scored by two raters. Each assessment contained two rating sheets. The second rating sheet had spaces for a third rating, if adjudication was necessary. Raters recorded their assigned codes on lines next to each context they scored. This procedure allowed us to track interrater reliability by rater and by institute. Raters were typically seated in groups of four.

Rater Training

On average, raters and adjudicators were trained 0.5 to 0.75 hour for each PSA context and 0.25 to 0.5 hour for each EA context. At all of the scoring institutes, the majority of the raters were veteran raters, permitting less training time compared to last year's institutes. The training included raters solving the problems in a particular context, presentation and discussion of the scoring rubric and strategy codes (if any) for that context, and examination of scored student work samples that clarified each portion of the rubric or each strategy code for each item. The context-specific training was followed by instruction on the general procedures for scoring (explained below). This context-specific training alternated with periods of scoring. For example, during a typical day at one of the Madison scoring institutes (June 15, 1999), all raters were trained in the scoring of "Ranger Station" (the first context, items 1-6, on the Grade 6 PSA). Test packets were randomly distributed. After raters finished rating this context, the coordinator randomly distributed the packet to different raters. Then each set of scores was compared and the assessments with discrepancies were routed to third raters (called adjudicators). When all of the Grade 6 PSAs were scored and adjudicated for this context, the raters were trained in the scoring of "A Patio" (the second context, items 7-9, on the Grade 6 PSA). Raters then scored "A Patio" in the same manner.

The Scoring Process

Each rater was given a packet of 5–8 student assessments to score. The rater scored the first assessment for a particular context and circled the score and strategy code (if applicable) on the Rater 1 Score Sheet. The rater then placed the Rater 1 Score Sheet at the back of the student assessment and placed the scored assessment at the bottom of the packet. Scoring continued until all student assessments in the packet were rated. The packet was handed to the coordinator, who in turn gave the rater another packet.[3] Scoring continued until all packets had been scored once.

Packets were then randomly distributed to different tables for the second round of rating. Raters used the same scoring process, but completed the Rater 2 Score Sheet for individual assessments. Scoring continued until all packets had been scored twice.

After each packet was scored a second time, the second rater compared both rating sheets for a given student assessment and marked scores and strategy codes (if applicable) that were not in agreement. These assessments were given to the coordinator to routed it to a third rater (called an adjudicator) for an additional rating. If agreement was reached between two of the now three raters, the agreed score or strategy code was used for the student response. If agreement was not reached, another adjudicator scored the response. If agreement was reached between two of the four raters, the agreed score or strategy code was used for the student response. The adjudication process continued until agreement was found between two of the raters.

This routing system allowed raters who worked faster to score more assessments than slower raters. One flaw in this system was that, since the second rater compared scores, s/he had an opportunity to change his/her score.

For EA multiple-choice items, packets were distributed in the same way as for PSA and EA contexts. Scoring, however, differed in that Rater 1 circled the letter selected by the student and the appropriate point value for the response (X for no response, 0 for an incorrect response, and 1 for a correct response). Rater 2 verified that the scoring was done correctly. Adjudication was unnecessary.

[3] Test packets were tracked by the coordinator using a color coding scheme to make sure different combinations of raters scored the different contexts on each test and so that no set of raters were paired too often.

Description of Rubrics

Item-specific scoring rubrics were used in the PSA. Scores ranged from X (no response) or 0 (incorrect response) to 4, depending on the complexity of the problem. Correct answers for less complex items, for example, were scored 1; answers for the most complex items could receive as many as 4 points (see Table 2). The complexity of these rubrics was reflected in the discussion during training. For many of these items, raters were also asked to determine the strategy evident in the student's solution from a predetermined list of codes specific to the item.

Some of the scoring rubrics evolved during the scoring process. Two factors influenced this development. First, PSA items were pilot-tested with groups of 75–100 students. Rubrics and strategy codes were created and revised based on those student samples. However, because during the study the PSA was administered to hundreds of students, additional types of student responses and solution strategies were detected. These newly discovered cases were integrated into existing scoring and coding schemes. Second, as a result of the pilot test, some items were rewritten, and new items were included. Because student work for changed and new items was unavailable prior to the administration of the assessments, rubrics and strategy codes were based on anticipated student responses. As student responses were examined during the rating process, rubrics and lists of strategy codes were refined to better represent the variety of responses actually demonstrated on specific items. When rubrics or lists of strategies were changed, items scored prior to the changes were rescored.

Table 2
Scoring Rubric for Problem Solving Assessment

Complexity	Scoring Scheme					
Less	X	0	1			
	X	0	1	2		
	X	0	1	2	3	
More	X	0	1	2	3	4

Item-specific rubrics were also used with the EA. These rubrics were generally less complex than PSA rubrics, and, because EAs were designed to yield comparisons with national and international samples of students involved in the NAEP and TIMSS, these rubrics could not be changed. All items, regardless of complexity, were assigned 1 point (see Table 3). Complexity of scoring is reflected in the breakdown of that point. Some items were scored with a fraction of a point. Some items also included codes for student strategy. Scoring, at times, was complex.

Table 3
Scoring Rubric for External Assessment

Complexity	Scoring Scheme					
Less	X	0	1			
	X	0	0.5	1		
	X	0	0.33	0.66	1	
More	X	0	0.25	0.5	0.75	1

Interrater Reliability by Scoring Institute

The eight scoring institutes were held on May 13–14 (2 institutes), May 19–20, May 25, June 14–18, July 12–16, July 26–30, and August 2–5 (see Appendix A1). The number of assessments rated at the institutes varied depending on which assessment was scored, the number of new assessments received, and number of assessments that needed to be rescored. The number of items per context varied from 1 to 7, and on average two PSA or EA contexts were scored each day. Interrater reliability was then calculated by scoring institute (see Table 4).

Table 4
1999 Interrater Reliability by Scoring Institute: Problem Solving Assessment and External Assessment

Institute	Rater *(N)*	Assessments Rated *(N)*	Contexts Rated *(N)*	Student Responses Rated *(N)*	Agreement *(N)*	Agreement *(%)*	Single Adjudication *(N)*	Single Adjudication *(%)*	Multiple Adjudication *(N)*	Multiple Adjudication *(%)*
1	7	154	3	2592	2069	79.82%	498	19.21%	25	0.96%
2	10	205	1	2505	2203	87.94%	284	11.34%	18	0.72%
3	14	130	4	2316	2012	86.87%	276	11.92%	28	1.21%
4	4	131	4	825	686	83.15%	137	16.61%	2	0.24%
5	12	700	6	25552	23408	91.61%	2072	8.11%	72	0.28%
6	17	826	4	35512	31968	90.02%	3285	9.25%	259	0.73%
7	18	3709	9	35002	32438	92.67%	2493	7.12%	71	0.20%
8	18	2376	8	29806	27598	92.59%	2140	7.18%	68	0.23%

The number of student responses given the same score points by two raters was determined and percentages were calculated. For example, of the 2592 student responses rated for the first scoring institute, 2069 student responses were assigned the same point scores by two raters. Therefore, the raters agreed on the point scores 79.82% of the time during the first scoring institute.

Interrater agreement was high for all eight scoring institutes, ranging from a low of 79.82% at the first institute to a high of 92.69% at the seventh institute. Interrater agreement tended to increase over time. This increase could be attributed to a high quality scoring procedure, the increasing experience of presenters and raters over time, the general movement from rating harder-to-score PSAs to easier-to-score EAs, and possibly, the manipulation of scores by the second rater when s/he compared scores.

Interrater Reliability by Rater

All raters at each institute were selected to calculate interrater reliability (see Table 5 and Table A1 in the Appendix). Agreement was then determined between ratings of the both raters on individual student responses, and percentages were calculated. For instance, Rater A agreed with a second rater on 365 of the 462 student responses or 79.00% of the time. (This includes when Rater A was the second rater.)

Interrater agreement was high for all raters, ranging from a low of 74.44% (Rater C) at the first institute to a high of 95.95% (Rater CD) at the seventh institute. Only six out of the 100 raters had less than 80% agreement with the second rater. More than half of the raters reached over 90% agreement with second raters. The factors that contribute to this high level of agreement can be attributed to clear rubrics, high quality presentation of rubrics and examples, and experience both over the course of each institute and over the set of institutes.

`

Table 5

Interrater Reliabiltiy by Rater

Institute (Student Responses Rated)	Rater	Student Responses Rated (N)	Agree-ment %	Single Adjudi-cation %	Multiple Adjudi-cation %	Institute (Student Responses Rated)	Rater	Student Responses Rated (N)	Agree-ment %	Single Adjudi-cation %	Multiple Adjudi-cation %	Institute (Student Responses Rated)	Rater	Student Responses Rated (N)	Agree-ment %	Single Adjudi-cation %	Multiple Adjudi-cation %
1	A	462	79.00%	19.48%	1.52%	**4**	AF	81	86.42%	13.58%	0.00%	**7**	BM	1784	91.59%	8.18%	0.22%
(2592)	B	260	82.69%	16.92%	0.38%	(825)	AG	114	87.72%	12.28%	0.00%	(35002)	BN	2138	91.96%	8.00%	0.05%
	C	446	74.44%	23.54%	2.02%		AH	409	81.42%	18.09%	0.49%		BO	1479	93.31%	6.49%	0.20%
	D	404	78.47%	20.54%	0.99%		AI	221	82.81%	17.19%	0.00%		BP	1809	90.49%	9.12%	0.39%
	E	214	78.50%	20.56%	0.93%								BQ	2203	92.42%	7.35%	0.23%
	F	476	84.66%	15.13%	0.21%	**5**	AJ	2143	92.44%	7.19%	0.37%		BR	2684	94.11%	5.77%	0.11%
	G	330	81.52%	18.18%	0.30%	(25552)	AK	2347	89.73%	9.97%	0.30%		BS	2275	95.12%	4.70%	0.18%
							AL	1932	93.63%	6.26%	0.10%		BT	1604	93.77%	6.05%	0.19%
2	H	114	84.21%	14.04%	1.75%		AM	2279	92.58%	7.11%	0.31%		BU	1922	92.98%	6.76%	0.26%
(2505)	I	150	88.00%	11.33%	0.67%		AN	2051	90.64%	8.97%	0.39%		BV	2523	93.30%	6.70%	0.00%
	J	195	81.03%	17.95%	1.03%		AO	2346	90.84%	8.87%	0.30%		BW	2595	92.87%	6.86%	0.27%
	K	285	89.12%	10.53%	0.35%		AP	2207	92.80%	6.98%	0.23%		BX	1741	87.65%	12.06%	0.29%
	L	324	91.05%	8.95%	0.00%		AQ	1455	92.78%	6.87%	0.34%		BY	1390	91.73%	7.77%	0.50%
	M	189	89.42%	10.58%	0.00%		AR	1999	92.75%	7.10%	0.15%		BZ	2422	92.20%	7.56%	0.25%
	N	342	89.47%	9.94%	0.58%		AS	1973	93.16%	6.69%	0.15%		CA	314	88.22%	11.78%	0.00%
	O	138	80.43%	15.94%	3.62%		AT	2696	90.58%	9.09%	0.33%		CB	2079	93.07%	6.69%	0.24%
	P	303	87.46%	11.88%	0.66%		AU	2124	88.51%	11.11%	0.38%		CC	2609	92.79%	7.13%	0.08%
	Q	465	89.68%	9.68%	0.65%	**6**	AV	2026	90.28%	8.79%	0.94%		CD	1431	95.95%	3.77%	0.28%
						(35512)	AW	1697	89.33%	9.49%	1.18%						
3	R	213	86.85%	11.74%	1.41%		AX	1579	89.68%	9.50%	0.82%	**8**	CE	1822	90.83%	8.95%	0.22%
(2316)	S	126	89.68%	10.32%	0.00%		AY	1574	86.98%	12.01%	1.02%	(29806)	CF	1646	92.71%	6.99%	0.30%
	T	222	90.99%	9.01%	0.00%		AZ	2566	90.88%	8.61%	0.51%		CG	1531	91.44%	8.23%	0.33%
	U	165	85.45%	12.73%	1.82%		BA	2821	90.43%	9.32%	0.25%		CH	2504	93.29%	6.47%	0.24%
	V	189	90.48%	7.94%	1.59%		BB	2628	92.09%	7.23%	0.68%		CI	2272	92.87%	7.09%	0.04%
	W	240	85.00%	13.75%	1.25%		BC	837	86.74%	12.31%	0.96%		CJ	1842	93.54%	6.19%	0.27%
	X	207	89.86%	9.66%	0.48%		BD	1724	92.05%	7.71%	0.23%		CK	1211	93.81%	6.11%	0.08%
	Y	186	80.11%	19.35%	0.54%		BE	1870	90.37%	8.72%	0.91%		CL	1564	92.26%	7.61%	0.13%
	Z	180	78.89%	17.78%	3.33%		BF	2909	90.62%	8.56%	0.83%		CM	1725	92.93%	6.90%	0.17%
	AA	147	91.16%	7.48%	1.36%		BG	2714	90.16%	9.32%	0.52%		CN	1729	94.74%	5.09%	0.17%
	AB	132	86.36%	12.88%	0.76%		BH	1720	88.90%	10.35%	0.76%		CO	1382	90.30%	9.41%	0.29%
	AC	222	90.54%	9.46%	0.00%		BI	2588	89.18%	10.16%	0.66%		CP	1393	91.67%	7.90%	0.43%
	AD	51	74.51%	15.69%	9.80%		BJ	1467	89.78%	9.20%	1.02%		CQ	1793	93.14%	6.64%	0.22%
	AE	36	88.89%	11.11%	0.00%		BK	1349	88.14%	10.60%	1.26%		CR	1274	90.66%	8.87%	0.47%
							BL	3443	90.21%	9.09%	0.70%		CS	2236	92.98%	6.89%	0.13%
													CT	1917	92.91%	7.04%	0.05%
													CW	214	86.92%	12.62%	0.47%
													CX	1751	93.20%	6.34%	0.46%

Conclusion

The high interrater agreement at all the 1999 scoring institutes indicates that a high quality procedure was used for scoring. The extensive training proved worthwhile because it reduced questions during scoring and lessened the need to adjudicate. As experienced elementary- and middle-school teachers, raters provided valuable input for clarifying PSA rubrics and identifying different categories of student responses and solution strategies. Through this process, rubrics became user-friendly, which in turn increased interrater reliability. The scoring institutes also provided a significant professional development opportunity for teacher-raters who commented that they would make changes in their pedagogy to emphasize mathematical communication, include lessons that promoted more complex reasoning, and integrate various types of problems designed to elicit student thinking at more complex levels in their classroom assessment practice.

Interrater Reliability on Problem Solving Assessments

Problem Solving Assessments (PSA) were scored at 7 scoring institutes in 1999 (see Table A1 in the Appendix). The number of assessments varied at each institute depending on number of contexts covered, number of new assessments received, and number of assessments rescored. The number of items in each context varied depending on the mathematical content and level of reasoning assessed. The PSA used 18 contexts, each of which was scored separately. The number of items per context varied from 1 to 7. On average, two contexts were scored each day. PSA items examined students' application of mathematics and mathematical reasoning at three levels. Items designed to elicit reasoning at the second and third levels were more open-ended in nature and more complex to score**.** In this section, interrater reliability is determined for each Problem Solving Assessment by grade and context in three ways: (a) overall, (b) by districts, and (c) by program (conventional curricula or *Mathematics in Context*).

Grade 6

Overall Interrater Reliability

The interrater agreement on the Grade 6 Problem Solving Assessment was high (90.56%; see Figure 4 and Appendix B1). Interrater agreement was over 80% on all contexts and 90% or more on two-thirds of the contexts. The interrater agreement ranged from a low of 84.43% on the "Selling Tickets" context (Item 6) to a high of 94.29% on the "Fly One Day" context (Items 10–13).

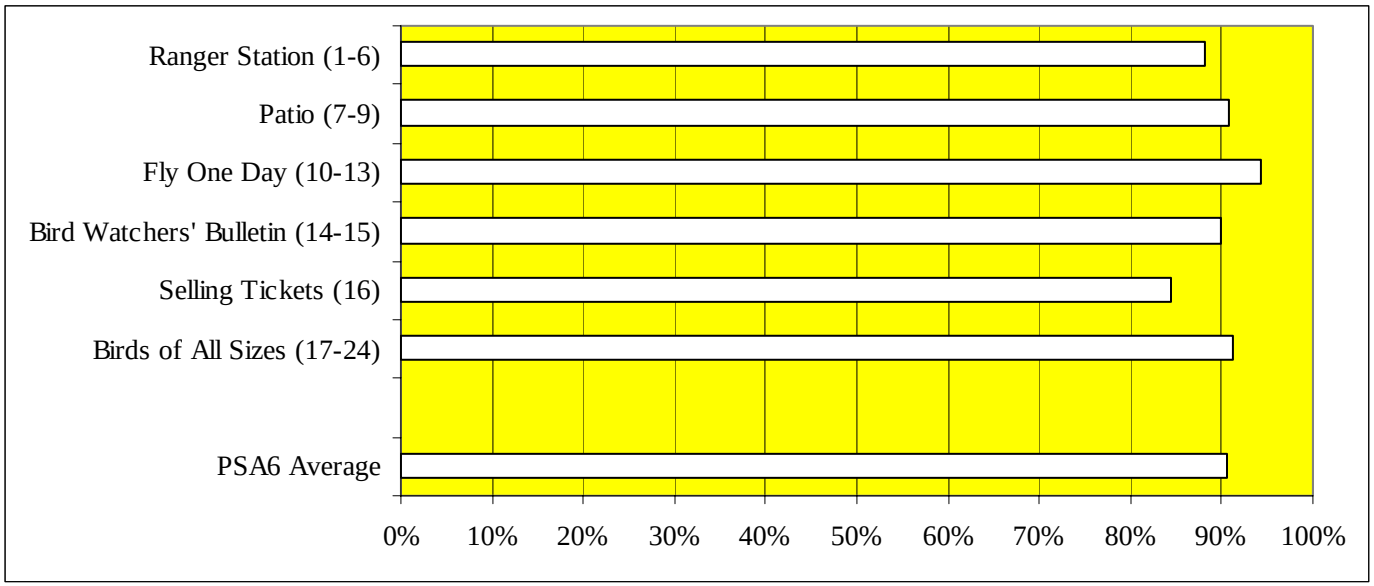

Figure 4. Interrater agreement on Grade 6 Problem Solving Assessment, by context.

All but one of the individual items had interrater agreement over 80%, and more than a third of the items had agreement over 90% (see Figure 5 and Table B1 in the Appendix). The interrater agreement on individual items ranged from a low of 79.43% on Item 9 from "A Patio" context to a high of 98.43% on two items (Item 7 from the "A Patio" context).

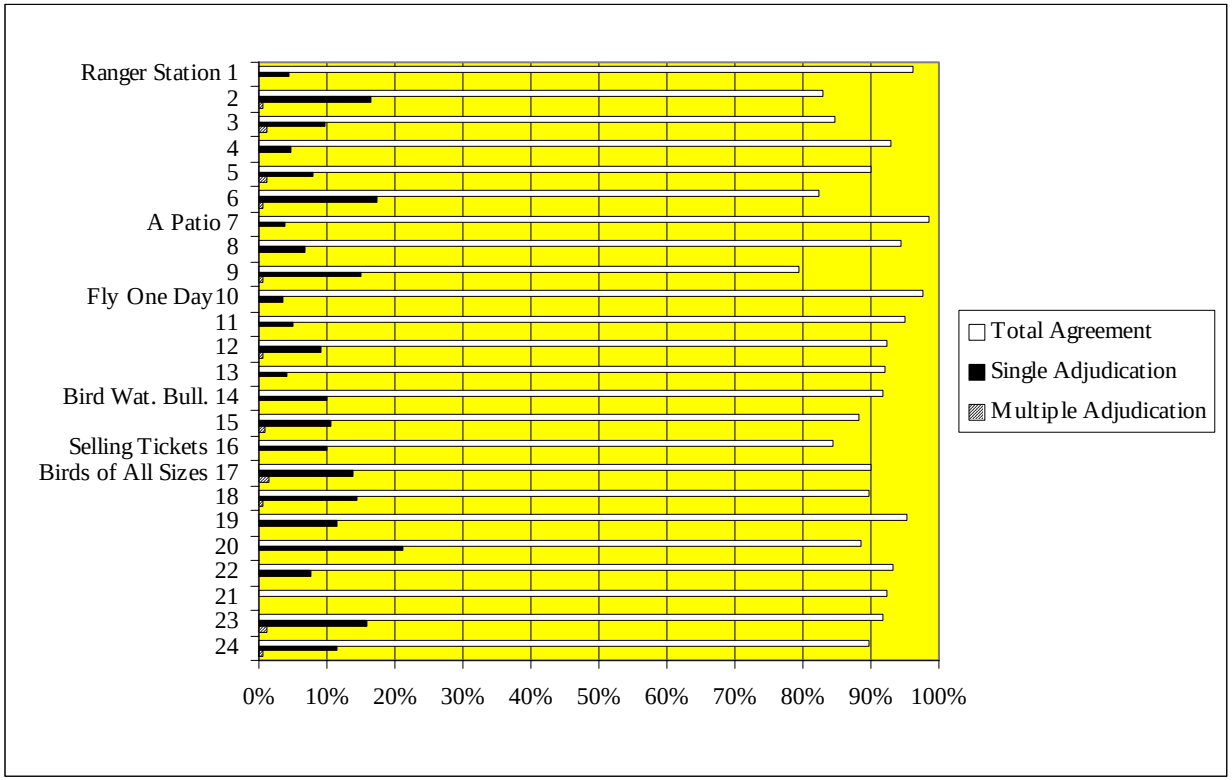

Figure 5. Interrater agreement on Grade 6 Problem Solving Assessment, by item.

The incidence of single adjudication was inversely proportional to the incidence of total agreement (see again Figure 5 and Table B1 in the Appendix). The percentage of single adjudication ranged from a low of 1.57% on Item 7 from the "A Patio" context to a high of 19.14% on Item 9 from the "A Patio" context.

The incidence of multiple adjudication was very low. It ranged from 0% on 6 items (Item 1 from "Ranger Station", Items 7 and 8 from the "A Patio" context, Item 10 from the "Fly One Day" context, and Items 20 and 22 from the "Birds of All Sizes" context) to a high of 1.57% on Item 6 from the "Ranger Station" context.

Factors that contributed to the high interrater agreement (and low adjudication) include (a) high quality training for raters; (b) well-defined and clarified rubrics; (c) effective scoring procedures; (d) lowest level of reasoning required in student responses (e.g., Item 7); and (e) proportion of student nonresponses (e.g., Item 4). The factor that contributed to lower interrater agreement (and higher adjudication) was raters at different sites may have been more perplexed with scoring certain items which were clarified at later scoring institutes (Item 9).[4]

[4] The "A Patio" context, including Item 9, was scored at the four on-side scoring institutes.

Interrater Reliability by Districts

District 1. In District 1, the interrater agreement on the Grade 6 Problem Solving Assessment was high (89.69%; see Figure 6 and Table B2 in the Appendix). Interrater agreement was over 80% on all contexts and half of the contexts were over 90%. The interrater agreement ranged from a low of 85.20% on the "Selling Tickets" context (Items 16) to a high of 94.39% on the "Fly One Day" context (Items 10–13).

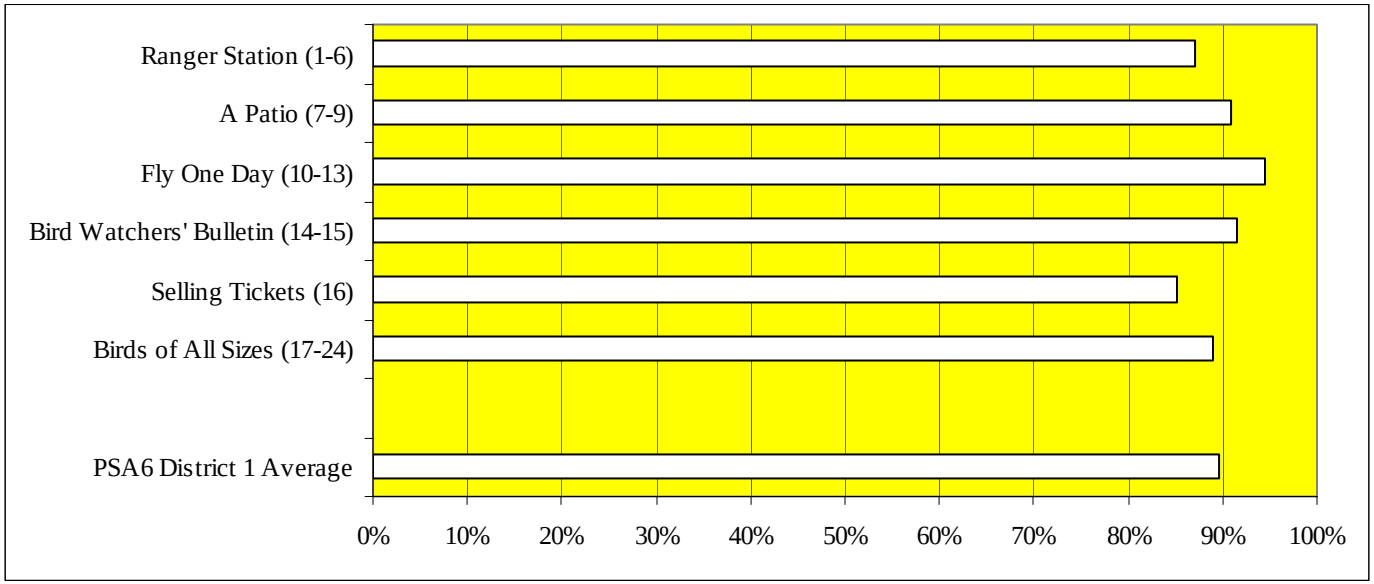

Figure 6. District 1 interrater agreement on Grade 6 Problem Solving Assessment, by context.

All of the individual items had interrater agreement over 80%, and almost half of the items had agreement over 90% (see Figure 7 and Table B2 in the Appendix). The interrater agreement on individual items ranged from a low of 80.10% on Item 9 from "A Patio" to a high of 98.98% on Item 10 from the "Fly One Day" context.

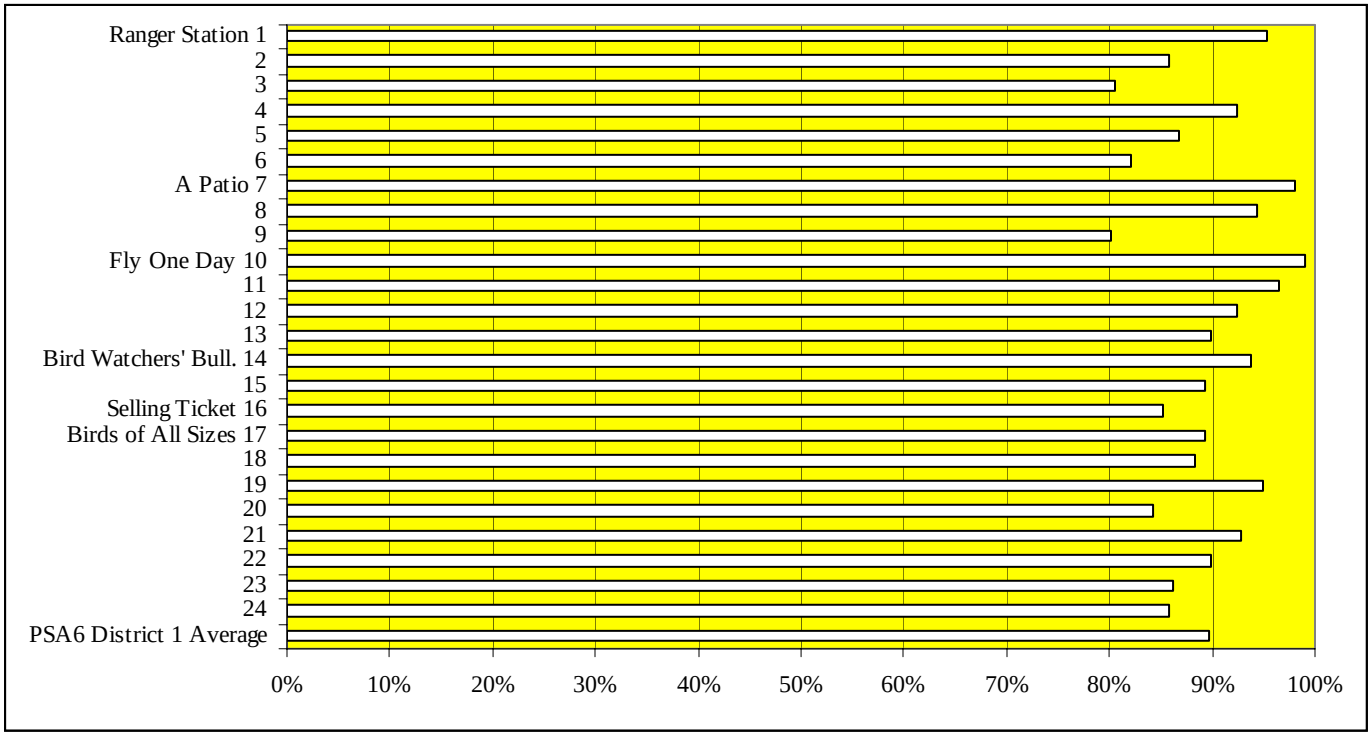

Figure 7. District 1 interrater agreement on Grade 6 Problem Solving Assessment, by item.

District 2. In District 2, the interrater agreement on the Grade 6 Problem Solving Assessment was high (91.92%; see Figure 8 and Table B2 in the Appendix). Interrater agreement was over 80% on all contexts and over 90% on four out of the six contexts. The interrater agreement ranged from a low of 87.45% on the "Bird Watchers' Bulletin" context (Item 14–15) to a high of 94.91% on the "Fly One Day" context (Items 7–9).

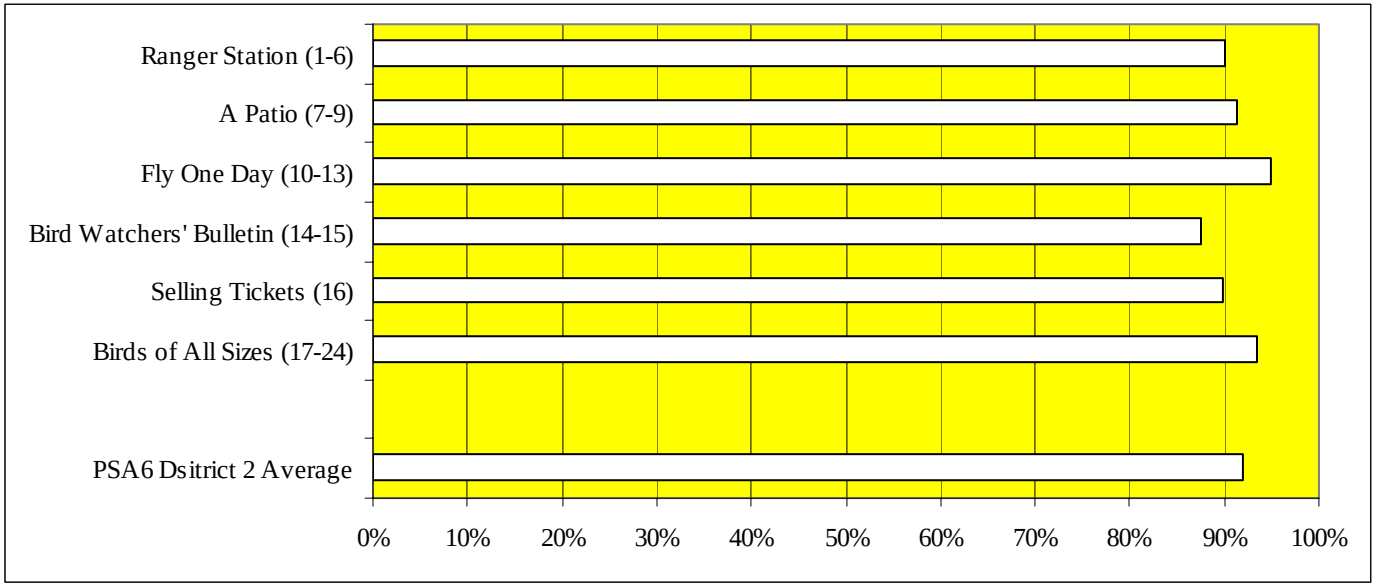

Figure 8. District 2 interrater agreement on Grade 6 Problem Solving Assessment, by context.

All of the individual items had interrater agreement over 80%, and three-quarters of the items had agreement over 90% (see Figure 9 and Table B2 in the Appendix). The interrater agreement on individual items ranged from a low of 80.36% on Item 9 from "A Patio" to a high of 99.27% on Item 7 from the "A Patio" context.

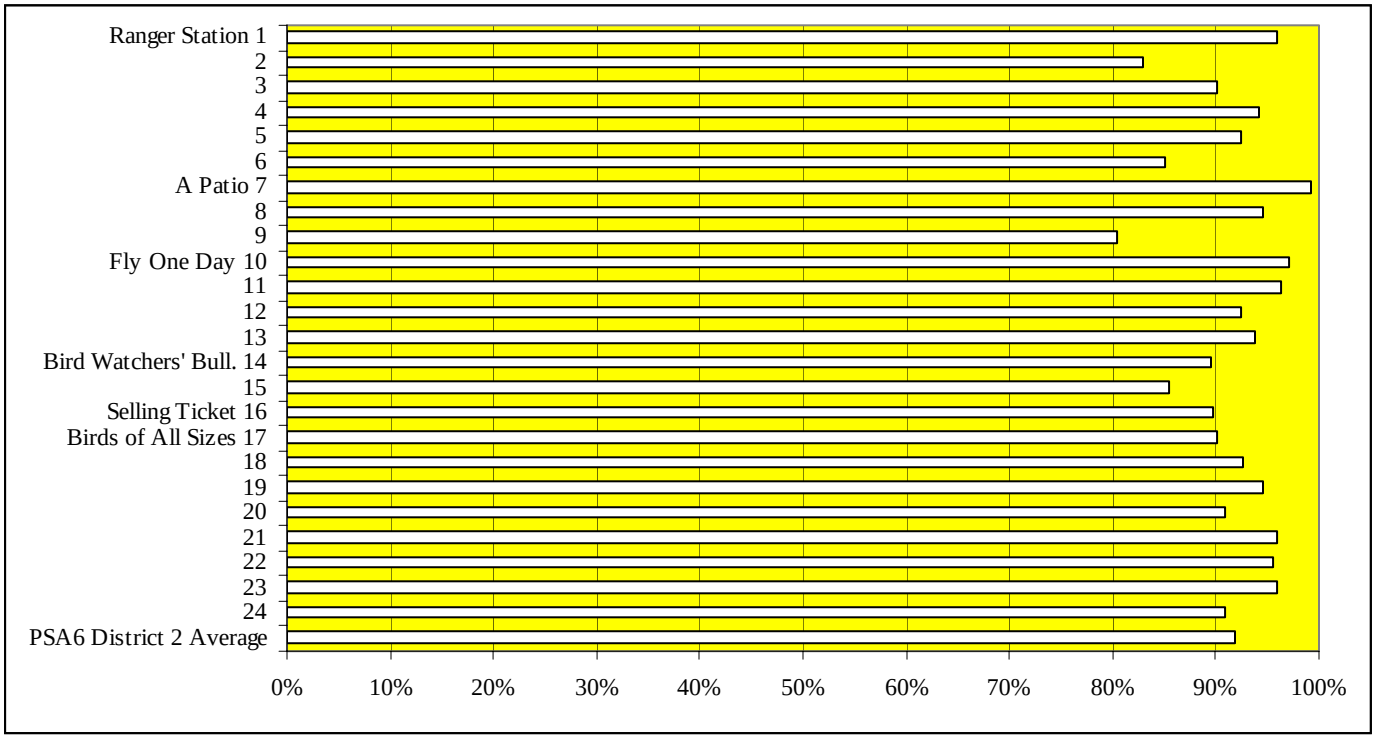

Figure 9. District 2 interrater agreement on Grade 6 Problem Solving Assessment, by item.

District 3. In District 3, the interrater agreement on the Grade 6 Problem Solving Assessment was high (86.71%; see Figure 10 and Table B2 in the Appendix). Interrater agreement was over 80% on five out of the six contexts, but only one context had agreement over 90%. The interrater agreement ranged from a low of 67.21% on the "Selling Tickets" context (Items 16) to a high of 92.21% on the "Fly One Day" context (Items 10–13).

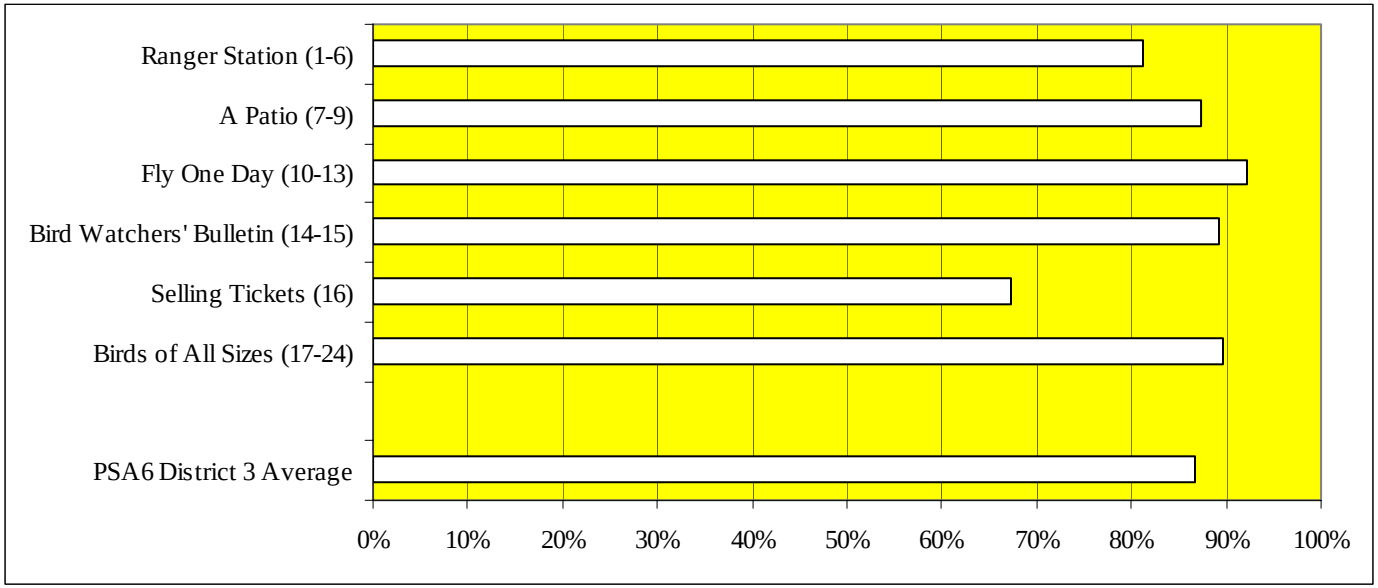

Figure 10. District 3 interrater agreement on Grade 6 Problem Solving Assessment, by context.

All but two of the individual items had interrater agreement over 80%, and about a third of the items had agreement over 90% (see Figure 11 and Table B2 in the Appendix). The interrater agreement on individual items ranged from a low of 67.21% on Item 16 from the "Selling Tickets" to a high of 99.18% on 2 items (Item 1 from the "Ranger Station" context and Item 10 from the "Fly One Day" context). Other individual items with low interrater agreement are Item 2 at 68.85%, Item 3 at 73.77%, and Item 6 at 71.31% from the "Ranger Station" context; Item 9 from the "A Patio" context at 71.31%.

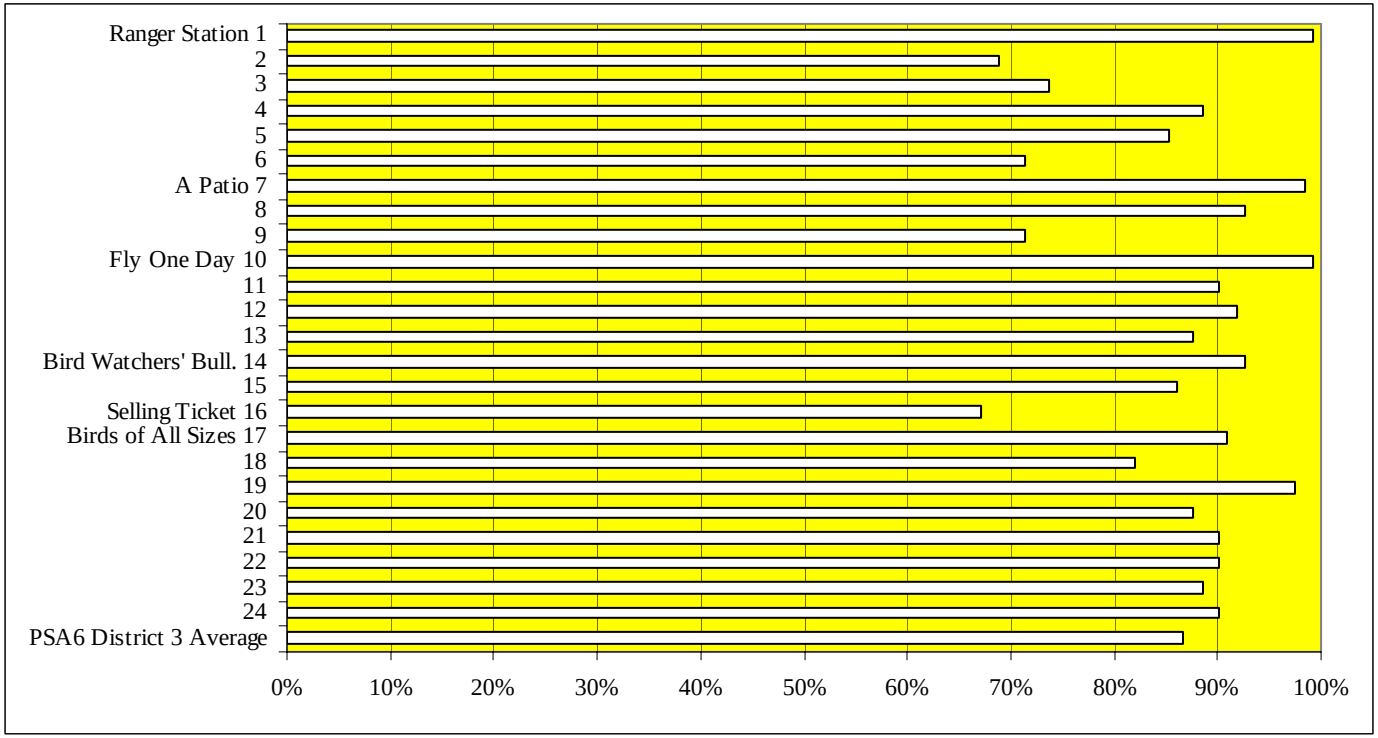

Figure 11. District 3 interrater agreement on Grade 6 Problem Solving Assessment, by item.

District 4. In District 4, the interrater agreement on the Grade 6 Problem Solving Assessment was high (93.03%; see Figure 12 and Table B2 in the Appendix). Interrater agreement was over 80% on all contexts, and over 90% on all but one of the contexts. The interrater agreement ranged from a low of 88.79% on the "Selling Tickets" context (Item 16) to a high of 94.86% on the "Fly One Day" context (Items 10–13).

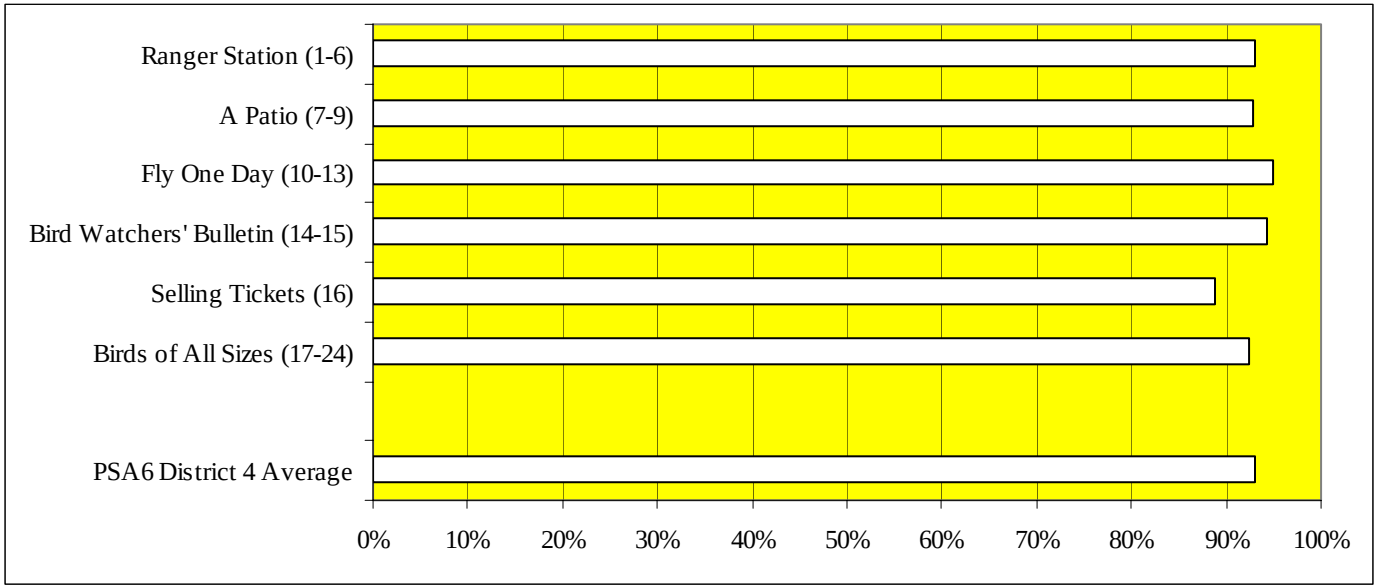

Figure 12. District 4 interrater agreement on Grade 6 Problem Solving Assessment, by context.

All of the individual items had interrater agreement over 80%, and four-fifths the items had agreement over 90% (see Figure 13 and Table B2 in the Appendix). The interrater agreement on individual items ranged from a low of 85.05% on Item 9 from the "A Patio" to a high of 97.20% on 2 items (Item 7 from the "A Patio" context and Item 13 from the "Fly One Day" context).

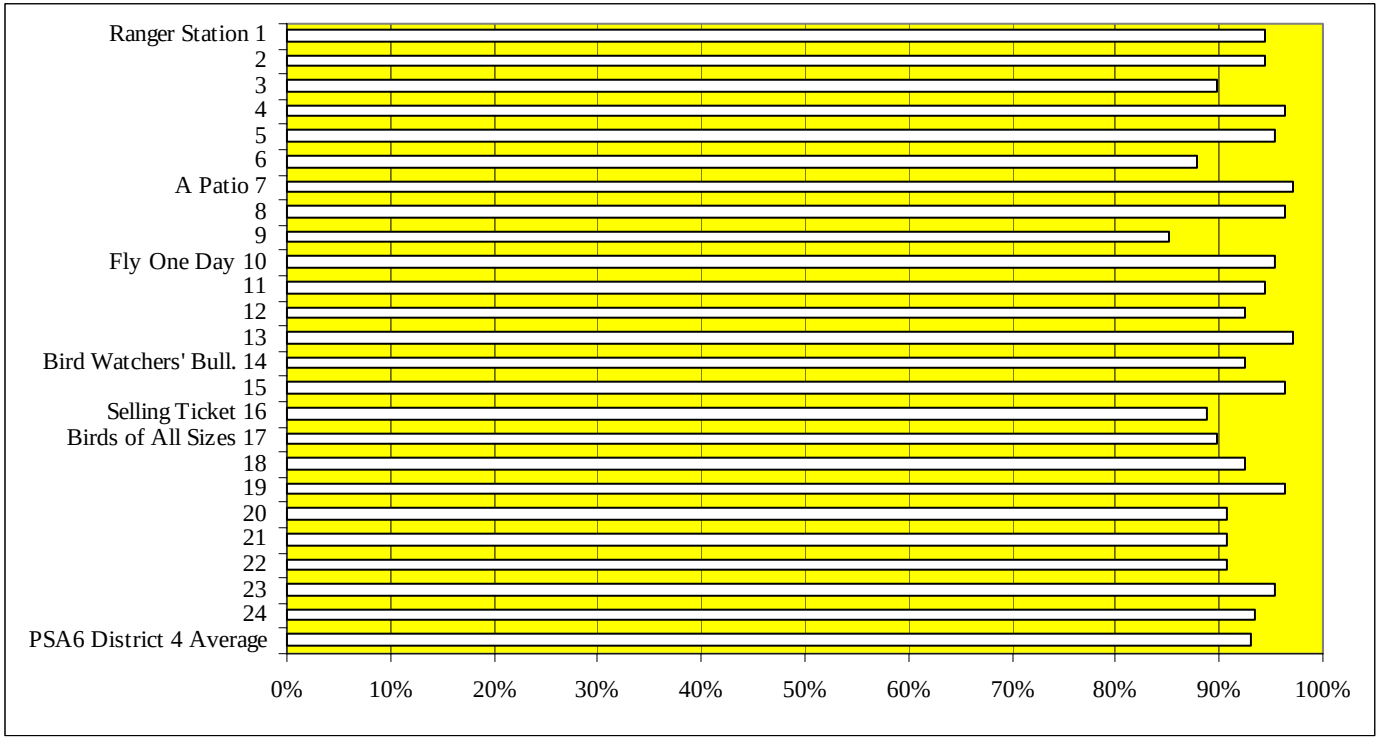

Figure 13. District 4 interrater agreement on Grade 6 Problem Solving Assessment, by item.

Across districts. There were some large differences (5% or greater) in interrater agreement across the districts (see Figure 14 and Table B2 in the Appendix). In District 1, there were no significant differences in interrater agreement compared to the other districts. In District 2, interrater agreement was higher than other districts on the "Selling Tickets." In District 3, interrater agreement was lower than other districts on the "Ranger Station," context and much lower on the "Selling Tickets" context. In District 4, interrater agreement was high on the "Selling Tickets" context.

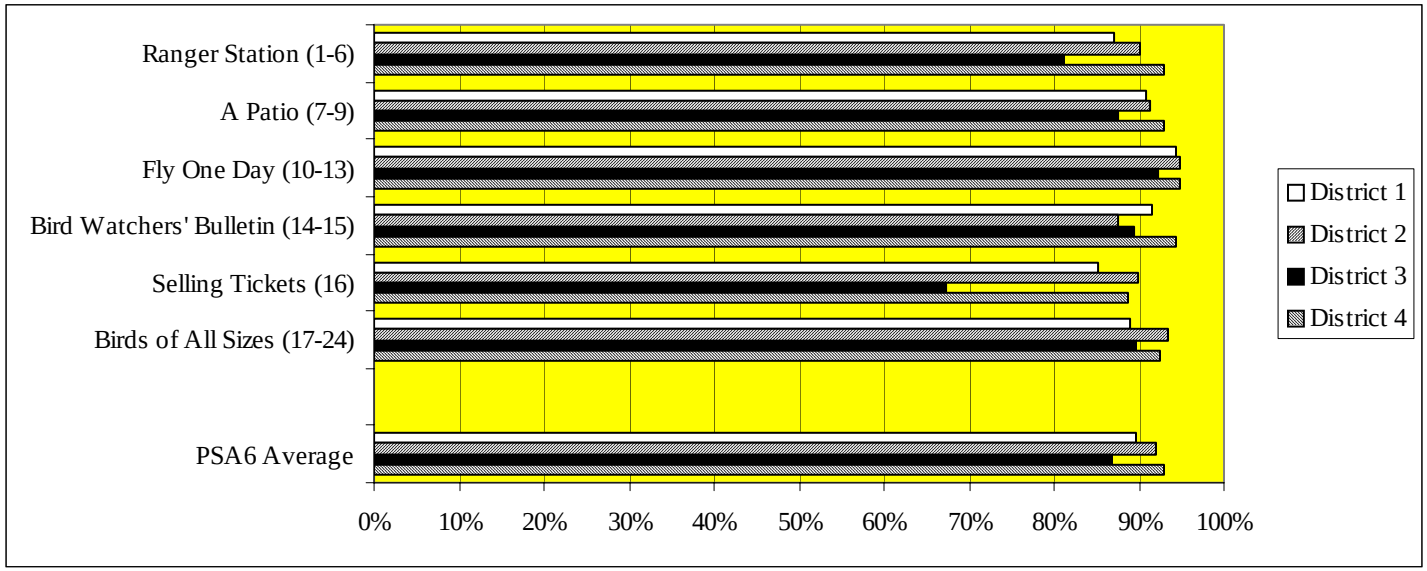

Figure 14. Across district interrater agreement on Grade 6 Problem Solving Assessment, by context.

Some individual items from each district had large (5% or greater) differences in interrater agreement (see Table 6 and Table B2 in the Appendix). In District 1, interrater agreement was lower than other in districts on Items 20 and 24, and low on Items 13 and 23. In District 2, interrater agreement was higher than other districts on Items 3, 21, 22, and 23. In District 3, interrater agreement was much lower than in other districts on Items 2, 3, 6, 9, 16, and 18; and lower on Items 4, 11, and 13. In District 4, interrater agreement was much higher than in other districts on Items 2, 9, 13, and 15; and high on Items 4 and 24.

Table 6

Interrater Agreement on Grade 6 Problem Solving Assessment by Item in all Districts

Context	Item Number	District 1	District 2	District 3	District 4
Ranger Station	1	95.41%	96.00%	99.18%	94.39%
	2	85.71%	82.91%	***68.85%*** [5]	**94.39%** [6]
	3	80.61%	**90.18%**	***73.77%***	89.72%
	4	92.35%	94.18%	***88.52%***	**96.26%**
	5	86.73%	92.36%	85.25%	95.33%
	6	82.14%	85.09%	***71.31%***	87.85%
A Patio	7	97.96%	99.27%	98.36%	97.20%
	8	94.39%	94.55%	92.62%	96.26%
	9	80.10%	80.36%	***71.31%***	**85.05%**
Fly One Day	10	98.98%	97.09%	99.18%	95.33%
	11	96.43%	96.36%	***90.16%***	94.39%
	12	92.35%	92.36%	91.80%	92.52%
	13	***89.80%***	93.82%	***87.70%***	**97.20%**
Bird Watchers' Bulletin	14	93.85%	89.45%	92.62%	92.52%
	15	89.29%	85.45%	86.07%	**96.26%**
Selling Tickets	16	85.20%	89.82%	***67.21%***	88.79%
Birds of All Sizes	17	89.29%	90.18%	90.98%	89.72%
	18	88.27%	92.73%	***81.97%***	92.52%
	19	94.90%	94.55%	97.54%	96.26%
	20	***84.18%***	90.91%	87.70%	90.65%
	21	92.86%	**96.00%**	90.16%	90.65%
	22	89.80%	**95.64%**	90.16%	90.65%
	23	***86.22%***	**96.00%**	88.52%	95.33%
	24	***85.71%***	90.91%	90.16%	**93.46%**
Average		89.69%	91.92%	86.71%	93.03%

[5] Percentage in bold with italics indicates lower differences (5% or greater) in interrater agreement.

[6] Percentage in bold indicates higher differences (5% or greater) in interrater agreement.

The large differences in interrater agreement across districts, when they occurred, were most likely due to differences in (a) presentation and interpretation of rubrics during each scoring institute[7]; (b) initial item scored at each institute; (c) content study teachers taught; (d) time of day items were scored; and (e) items eliciting a higher level of reasoning being left blank (District 3 had far fewer nonresponses, and District 4 more, than the other districts). In addition, teachers in District 3 did not teach as many sixth-grade units, teaching instead a combination of fifth and sixth-grade units. That might have affected interrater agreement.

[7] These assessments were scored at some sites simultaneously. The scoring institutes were conducted by different presenters and had different sets of study teachers as raters. Later institutes, with different raters, were held to score newly received assessments and to rescore items with improved rubrics. As a result, several different presenters and several different sets of raters scored these assessments.

Interrater Reliability by Program (Conventional Curricula or Mathematics in Context Classes)

Conventional curricula. The interrater agreement on the Grade 6 Problem Solving Assessment from classes that studied conventional curricula was high (89.58%; see Figure 15 and Table B3 in the Appendix). Interrater agreement was over 80% on all contexts and over 90% on one of the contexts. The interrater agreement ranged from a low of 85.71% on the "Selling Tickets" context (Items 16) to a high of 97.32% on the "Fly One Day" context (Items 10-13).

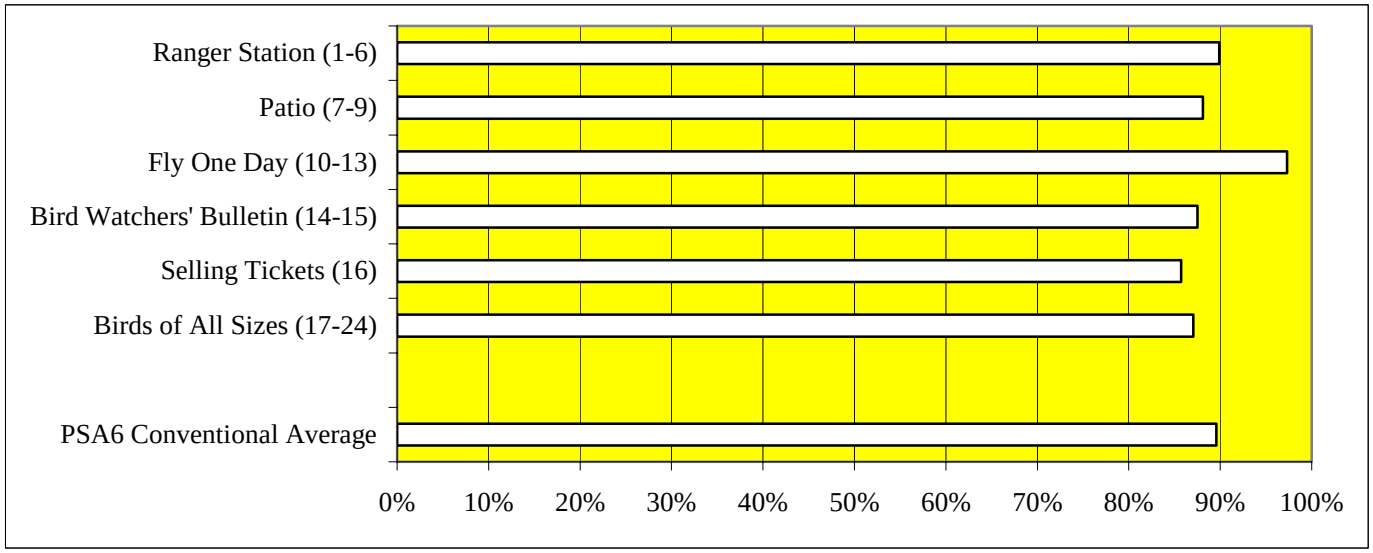

Figure 15. Interrater agreement on Grade 6 Problem Solving Assessment, by context: Conventional curricula.

All but two of the individual items had interrater agreement over 80%, and about half of the items had agreement over 90% (see Figure 16 and Table B3 in the Appendix). The interrater agreement on individual items ranged from a low of 78.57% on Item 9 from the "A Patio" and Item 20, the "Birds of All Sizes" context to a high of 100.00% on Item 13 from the "Fly One Day" context).

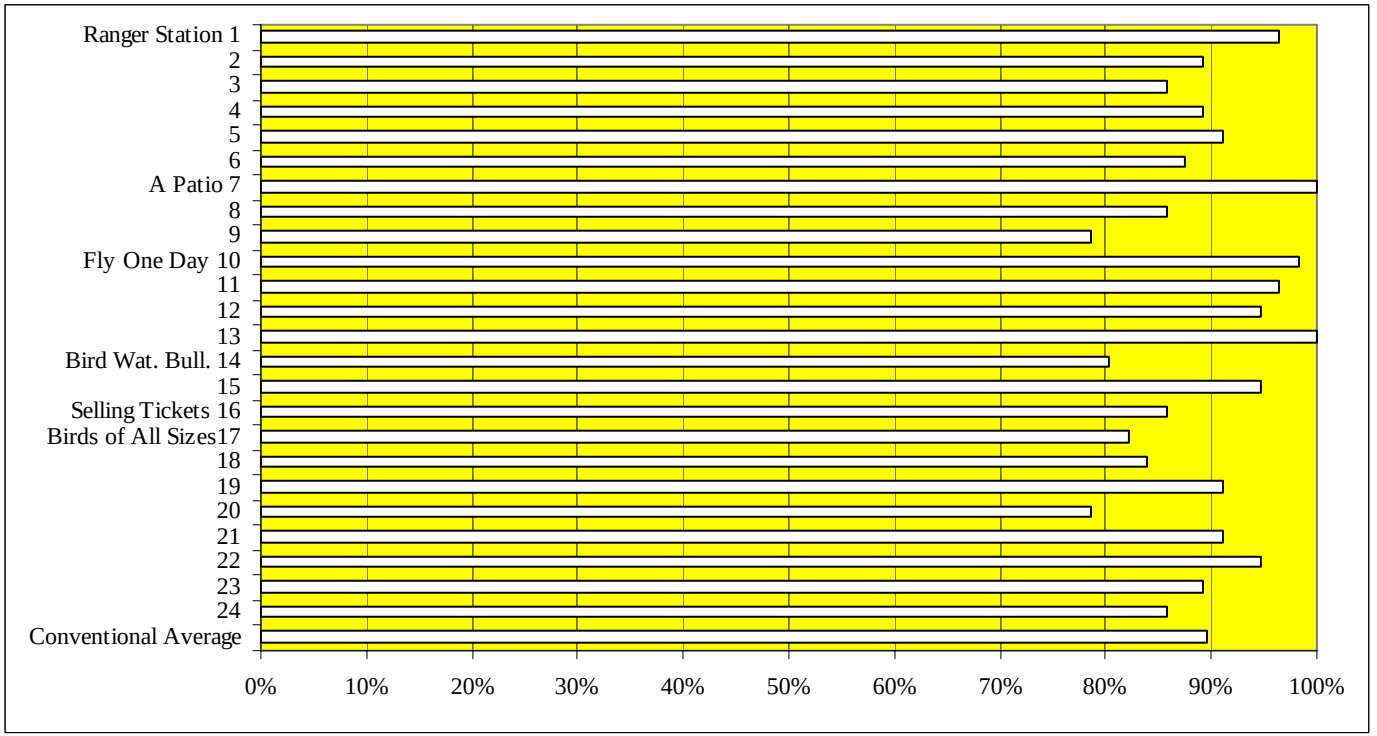

Figure 16. Interrater agreement on Grade 6 Problem Solving Assessment, by item: Conventional curricula.

Mathematics in Context classes. The interrater agreement on the Grade 6 Problem Solving Assessment from *Mathematics in Context* classes was high (90.64%; see Figure 17 and Table B3 in the Appendix). Interrater agreement was over 80% on all contexts and over 90% on two-thirds of the contexts. The interrater agreement ranged from a low of 84.32% on the "Selling Tickets" context (Item 16) to a high of 94.02% on the "Fly One Day" context (Items 10–13).

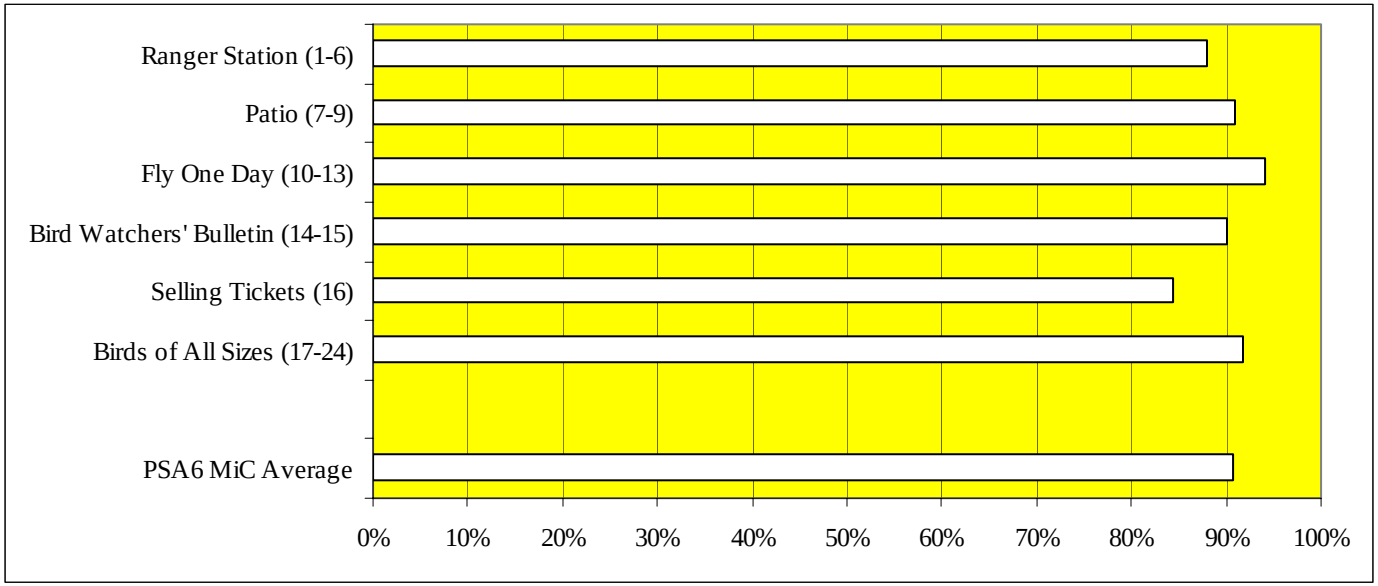

Figure 17. Interrater agreement on Grade 6 Problem Solving Assessment, by context: *Mathematics in Context* classes.

All but one of the individual items had interrater agreement over 80%, and two-thirds of the items had agreement over 90% (see Figure 18 and Table B3 in the Appendix). The interrater agreement on individual items ranged from a low of 79.50% on Item 9 from the "A Patio" context to a high of 98.29% on Item 7 from the "A Patio" context.

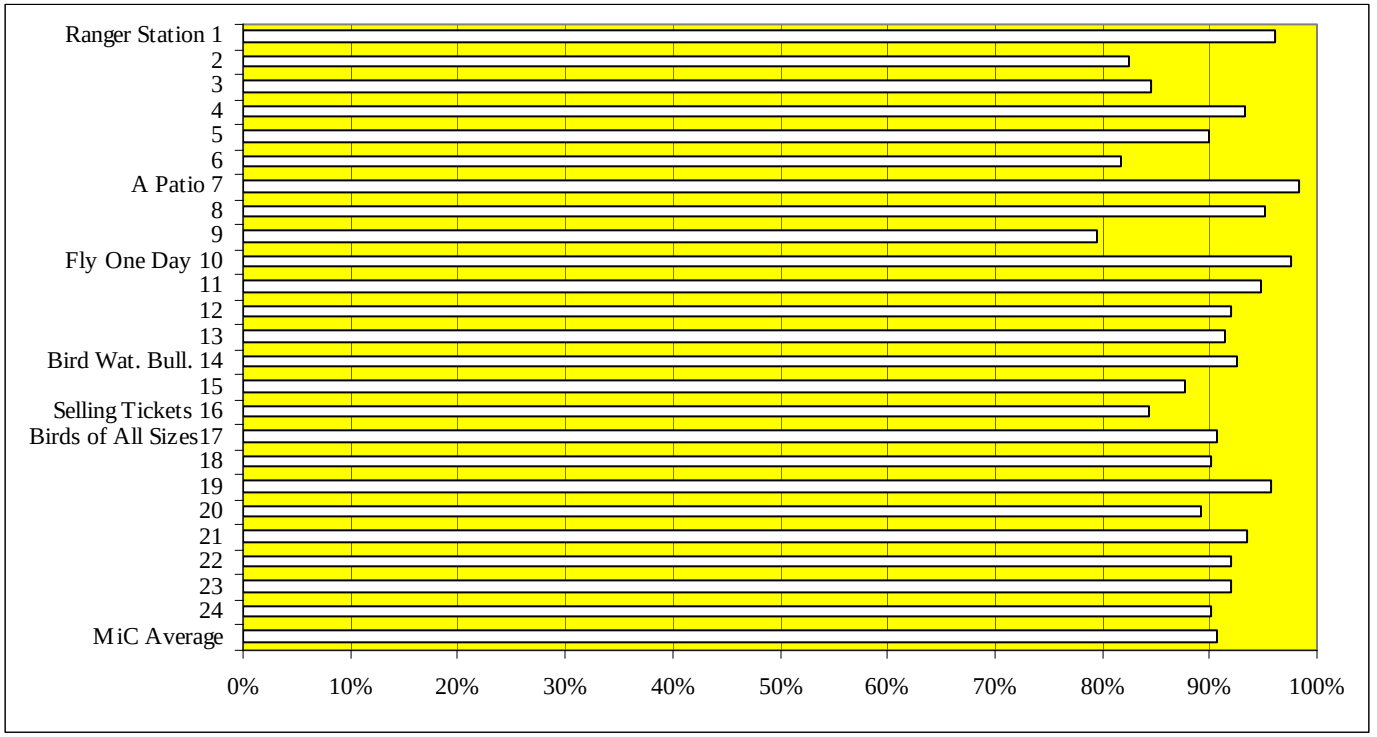

Figure 18. Interrater agreement on Grade 6 Problem Solving Assessment, by item: *Mathematics in Context* classes.

Across programs. Overall, the interrater agreement in conventional curricula and *Mathematics in Context* classes was similar (see Figure 19 and Table B3 in the Appendix). The average interrater agreement for conventional curricula was 89.58% and 90.64% for *Mathematics in Context* classrooms.

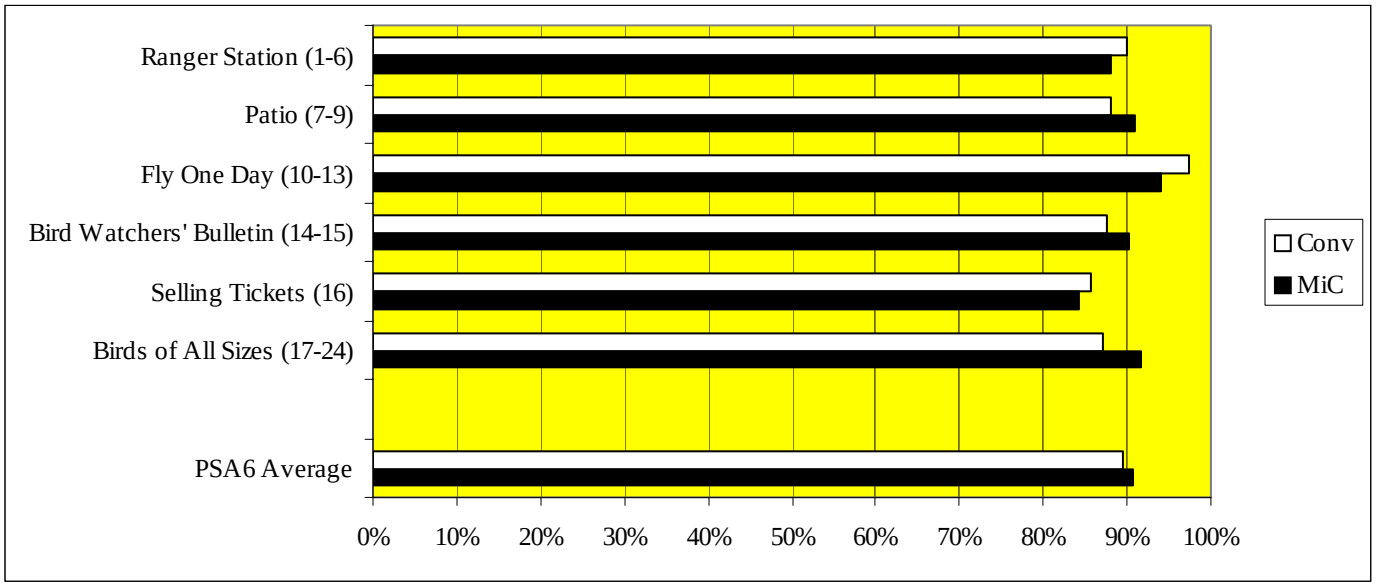

Figure 19. Interrater agreement on Grade 6 Problem Solving Assessment, by context: Conventional curricula and *Mathematics in Context* classes.

The interrater agreement on some individual items also revealed a discrepancy between conventional curricula and *Mathematics in Context* classes (see Figure 20 and Table B3 in the Appendix). Assessments from conventional curricula had much higher agreement (5% or greater) on Items 2 and 6 from the "Ranger Station" context, Item 13 from the "Fly One Day" context, and Items 15 from the "Bird Watchers' Bulletin" context. However, assessments from *Mathematics in Context* classes had much higher agreement (5% or greater) on Item 8 from the "A Patio" context, Item 14 from the "Bird Watchers' Bulletin" context, and Items 17, 18, and 20 from the "Birds of All Sizes" context.

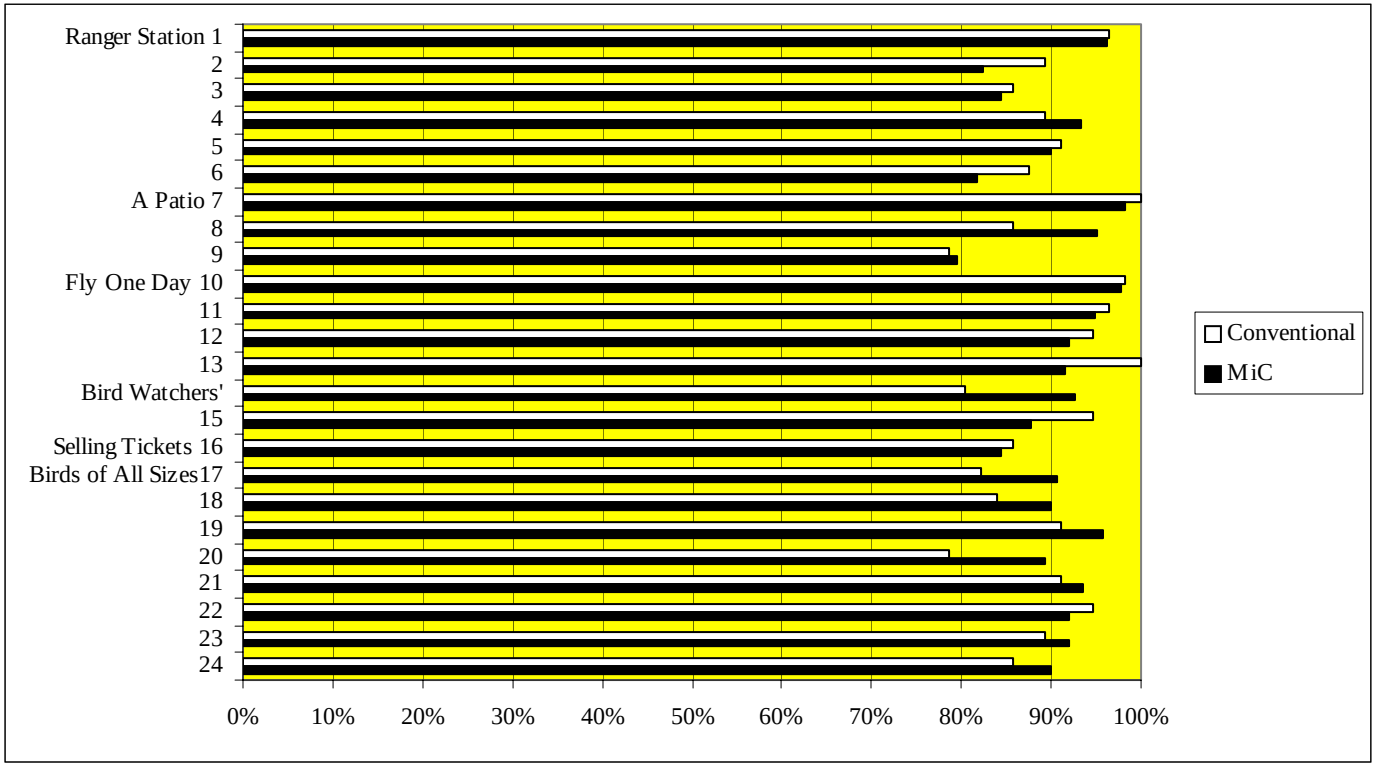

Figure 20. Interrater agreement on Grade 6 Problem Solving Assessment, by item: Conventional curricula and *Mathematics in Context* classes.

The large differences in interrater agreement were most likely due to differences in (a) initial item scored at each institute; (b) content study teachers taught; (c) time of day items were scored; (d) raters' interpretation of student work; and (e) proportion of student nonresponses at the end of section for the day.

Grade 7

Overall Interrater Reliability

The interrater agreement on the Grade 7 Problem Solving Assessment was high (90.40%; see Figure 21 and Appendix B4). Interrater agreement was over 80% on all contexts and over 90% in three-fifths of the contexts. The interrater agreement ranged from a low of 82.81% on "The Pentagon" context (Items 8–10) to a high of 94.89% on the "Playgrounds" context (Items 22–26).

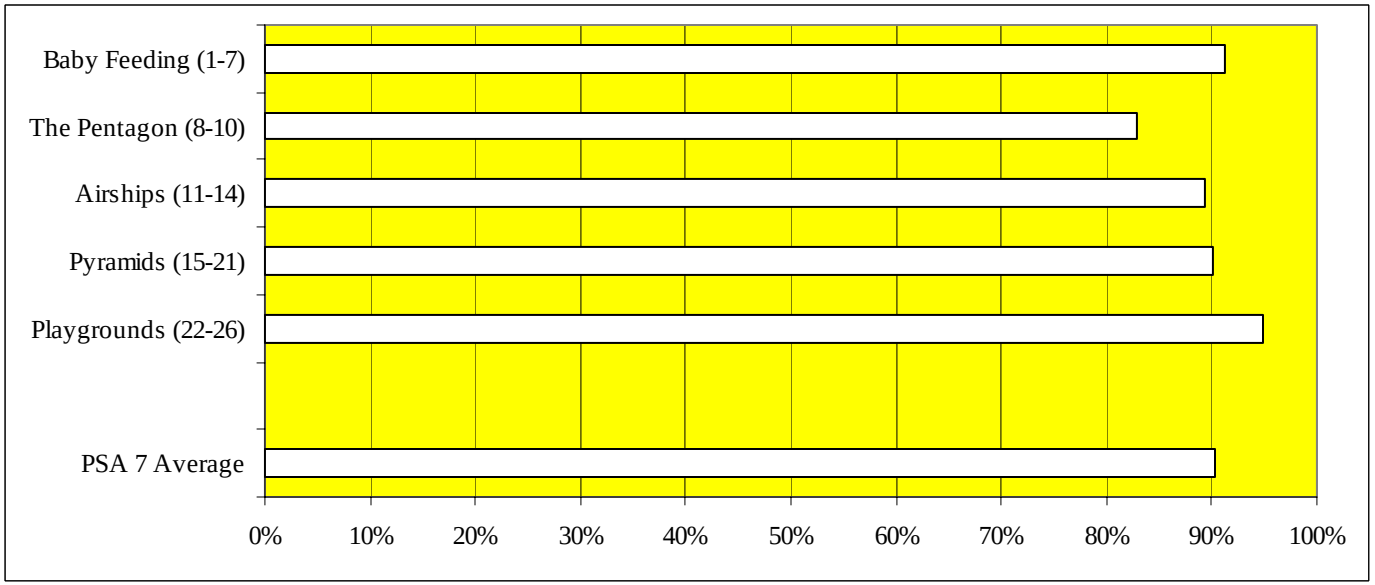

Figure 21. Interrater agreement on Grade 7 Problem Solving Assessment, by context.

All but two of the individual items had interrater agreement over 80%, and about two-thirds of the items had agreement over 90% (see Figure 22 and Table B4 in the Appendix). The interrater agreement on individual items ranged from a low of 69.61% on Item 12 from "Airships" context to a high of 97.94% on Item 11 from the "Airships" context. The other item with lower interrater agreement was Item 9 from "The Pentagon" context at 70.70%.

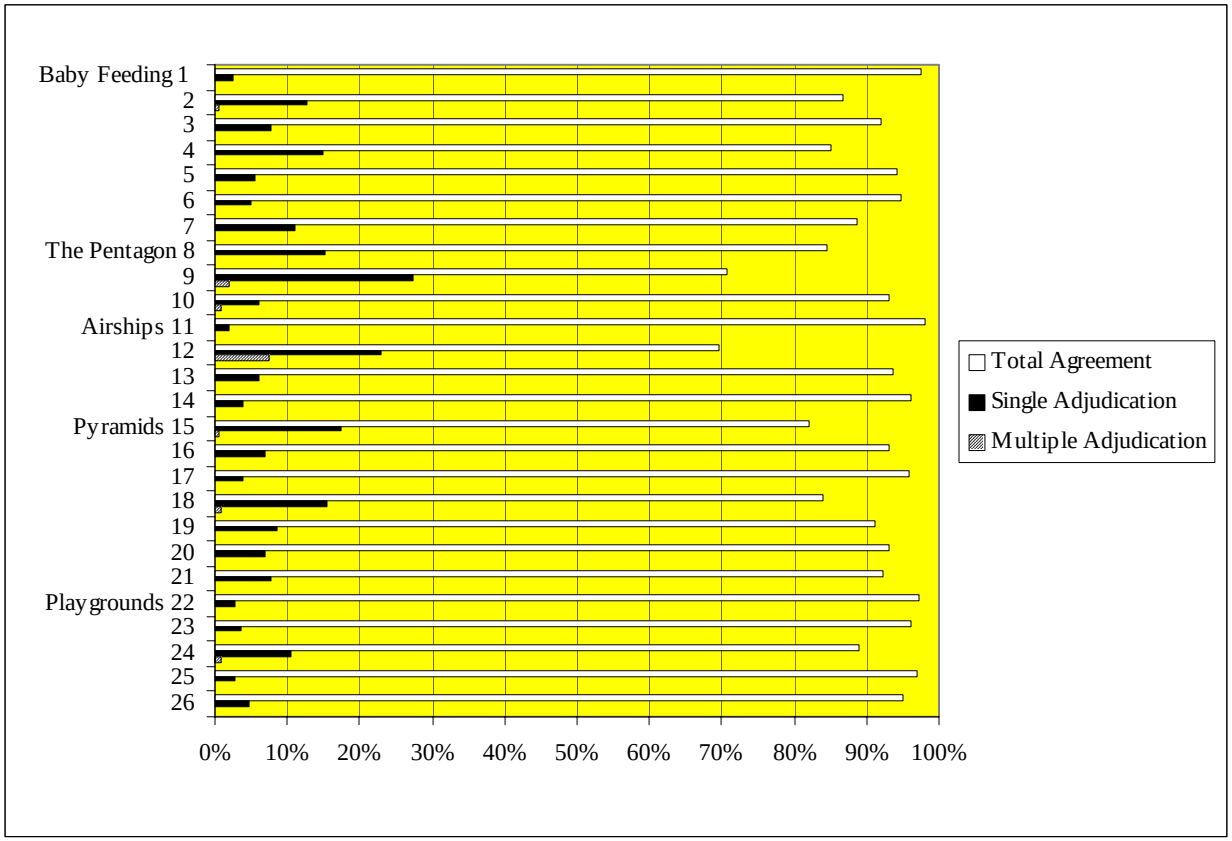

Figure 22. Interrater agreement on Grade 7 Problem Solving Assessment, by item.

The incidence of single adjudication was inversely proportional to the incidence of total agreement (see again Figure 22 and Table B4 in the Appendix). The percentage of single adjudication ranged from a low of 1.94% on Item 11 from the "Airships" context to a high of 27.36% on Item 9 from the "The Pentagon" context. Item 12 from the "Airships" context also had a very high incidence of single adjudication at 23.00%.

The incidence of multiple adjudication was very low. It ranged from 0% on 4 items (Item 1 from the "Baby Feeding" context, Item 14 from the "Airships" context, Item 20 from the "Pyramids" context, and Item 22 from the "Playgrounds" context) to a high of 7.38% on Item 12 from the "Airships" context.

The factor that contributed to the lower interrater agreement and higher adjudication was multiple and detailed criteria in scoring graphs (Item 12) and drawings (Item 9) which students produced.

Factors that contributed to the high interrater agreement and low adjudication include (a) high quality training for raters; (b) well-defined and clarified rubrics; (c) effective scoring procedures; (d) lowest level of reasoning required in student responses (e.g., Item 7); and (e) proportion of student nonresponses and incorrect answers (Items 11 and 25).

Interrater Reliability by Districts

District 1. In District 1, the interrater agreement on the Grade 7 Problem Solving Assessment was high (90.62%; see Figure 23 and Table B5 in the Appendix). Interrater agreement was over 80% on all contexts and three-fifths of the contexts were over 90%. The interrater agreement ranged from a low of 80.99% on the "The Pentagon" context (Items 8–10) to a high of 94.46% on the "Playgrounds" context (Items 22–26).

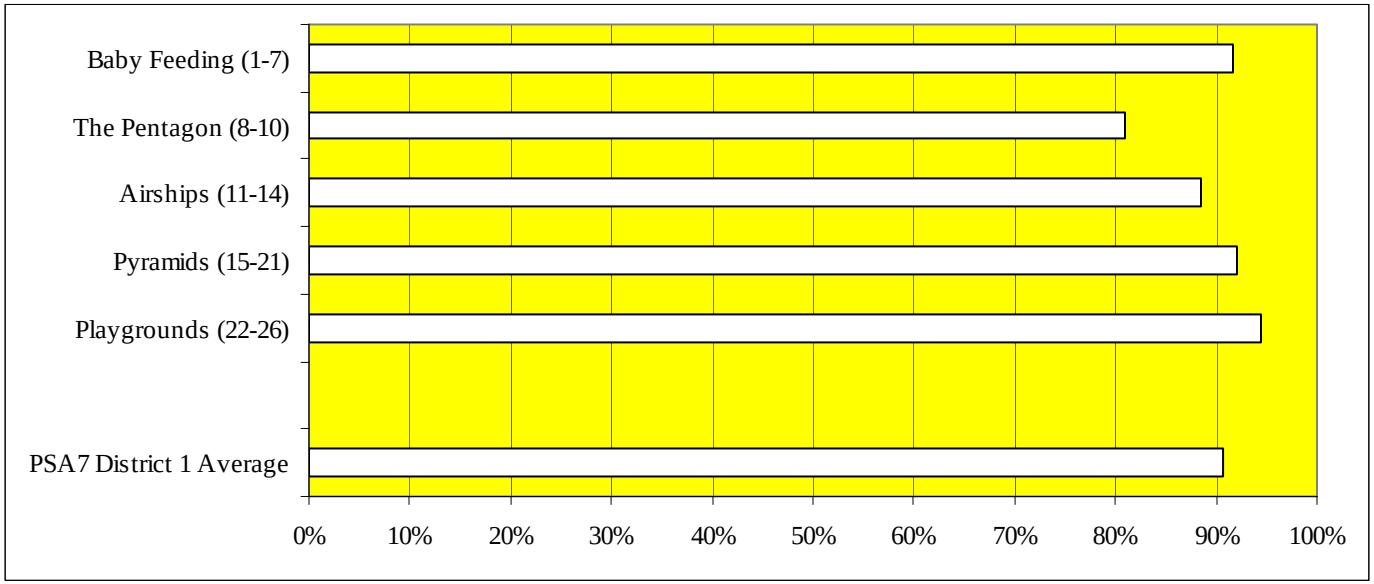

Figure 23. District 1 interrater agreement on Grade 7 Problem Solving Assessment, by context.

All but two of the individual items had interrater agreement over 80%, and almost three-quarters of the items had agreement over 90% (see Figure 24 and Table B5 in the Appendix). The interrater agreement on individual items ranged from a low of 68.67% on Item 12 from "Airships" to a high of 99.20% on Item 1 from the "Baby Feeding" context. The other context with low interrater agreement was Item 9 from "The Pentagon" context at 69.08%

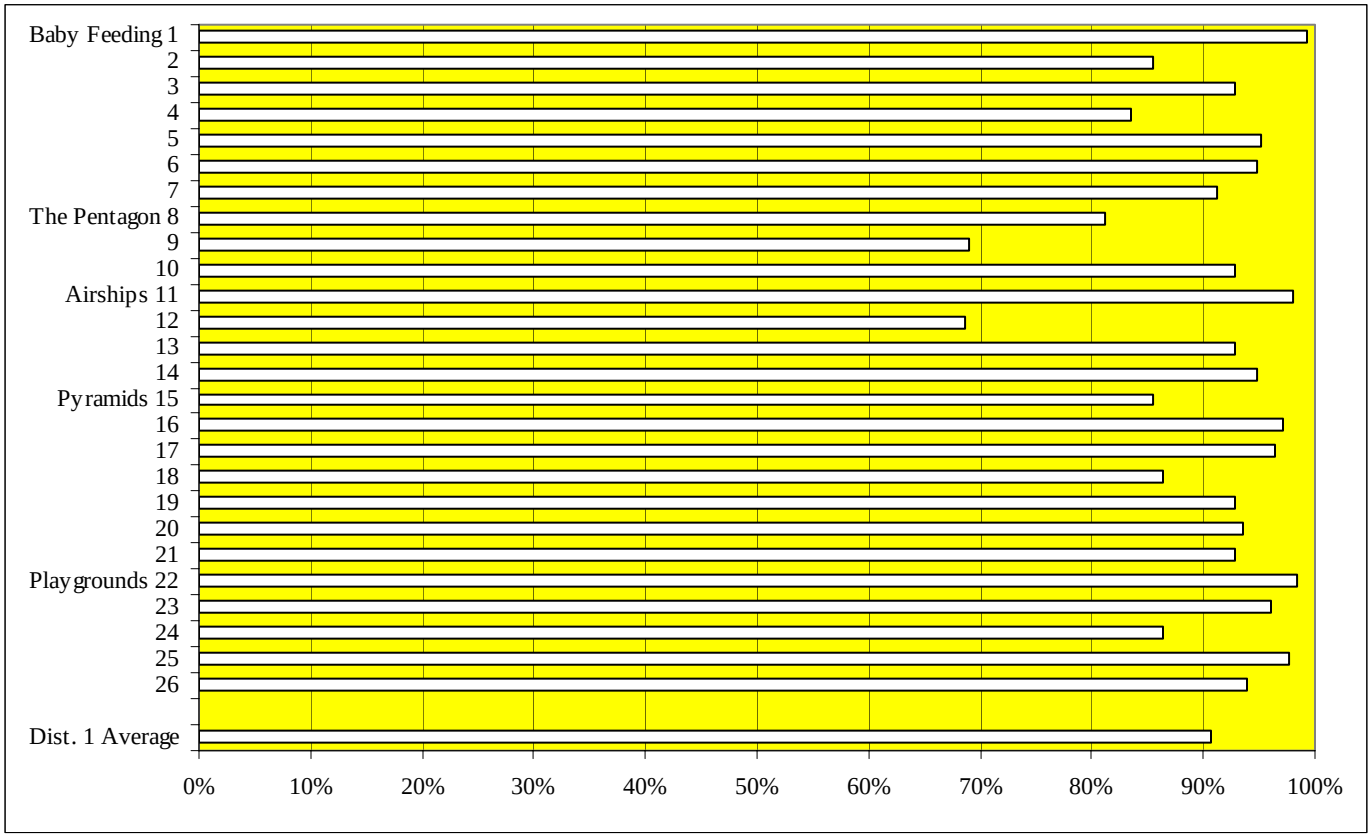

Figure 24. District 1 interrater agreement on Grade 7 Problem Solving Assessment, by item.

District 2. In District 2, the interrater agreement on the Grade 7 Problem Solving Assessment was high (90.45%; see Figure 25 and Table B5 in the Appendix). Interrater agreement was over 80% on all contexts and over 90% on two out of the five contexts. The interrater agreement ranged from a low of 83.47% on the "The Pentagon" context (Item 8–10) to a high of 95.52% on the "Playgrounds" context (Items 22–26).

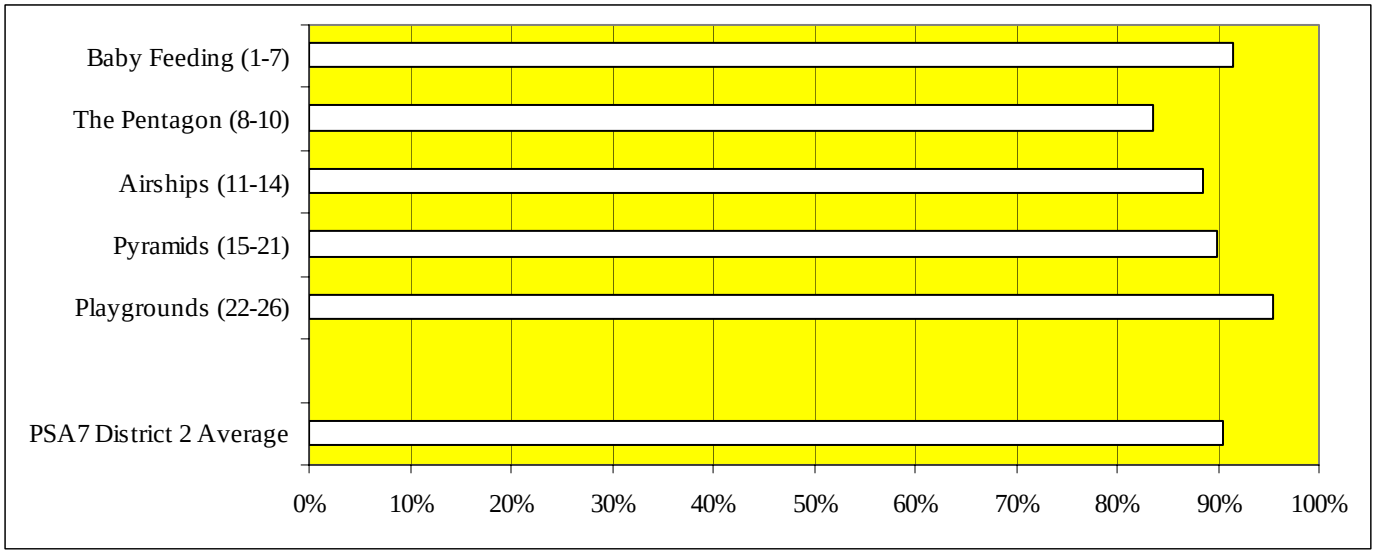

Figure 25. District 2 interrater agreement on Grade 7 Problem Solving Assessment, by context.

All but two of the individual items had interrater agreement over 80%, and about two-thirds of the items had agreement over 90% (see Figure 26 and Table B5 in the Appendix). The interrater agreement on individual items ranged from a low of 65.20% on Item 12 from "Airships" to a high of 98.00% on Item 11 from the "Airships" context. The other item with low interrater agreement was Item 9 from "The Pentagon" context at 70.80%.

Figure 26. District 2 interrater agreement on Grade 7 Problem Solving Assessment, by item.

District 3. In District 3, the interrater agreement on the Grade 7 Problem Solving Assessment was high (87.74%; see Figure 27 and Table B5 in the Appendix). Interrater agreement was over 80% on four out of the five contexts. The interrater agreement ranged from a low of 78.02% on the "The Pentagon" context (Items 7–10) to a high of 93.19% on the "Playgrounds" context (Items 22–26).

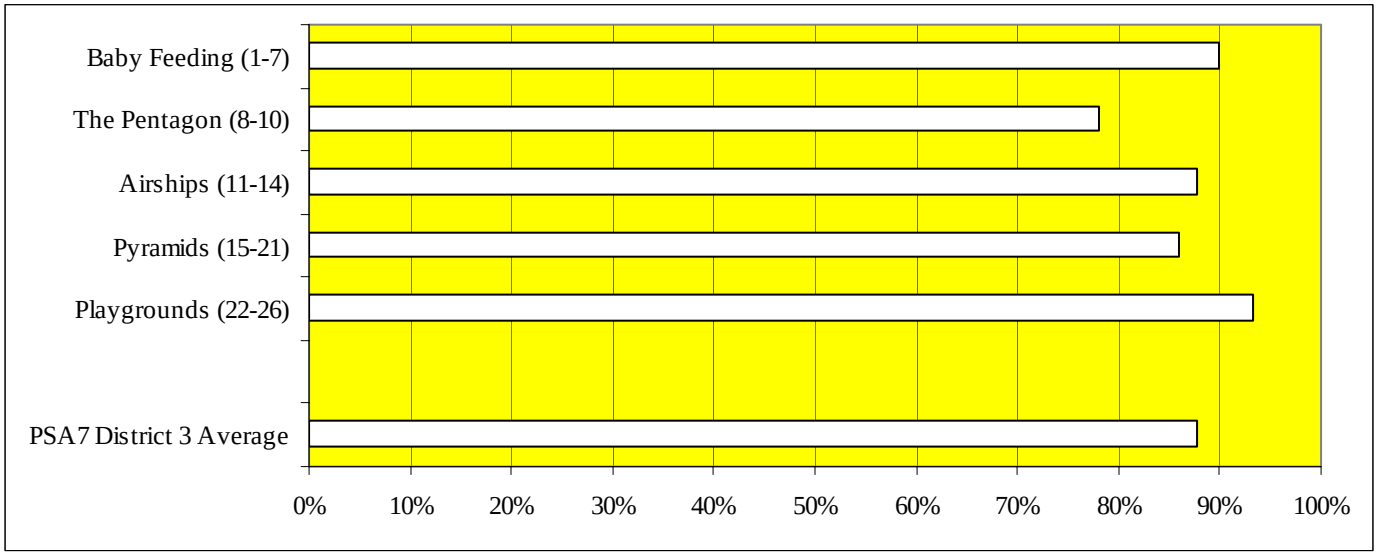

Figure 27. District 3 interrater agreement on Grade 7 Problem Solving Assessment, by context.

About five-sixths of the individual items had interrater agreement over 80%, and about half of the items had agreement over 90% (see Figure 28 and Table B5 in the Appendix). The interrater agreement on individual items ranged from a low of 57.97% on Item 9 from the "The Pentagon" to a high of 97.83% on Item 1 from the "Baby Feeding" context. Other individual items with low interrater agreement are Item 12 from the "Airships" context at 69.57%, Item 18 from the "Pyramids" context at 77.54%, and Item 15 from the "Pyramids" context at 78.99%.

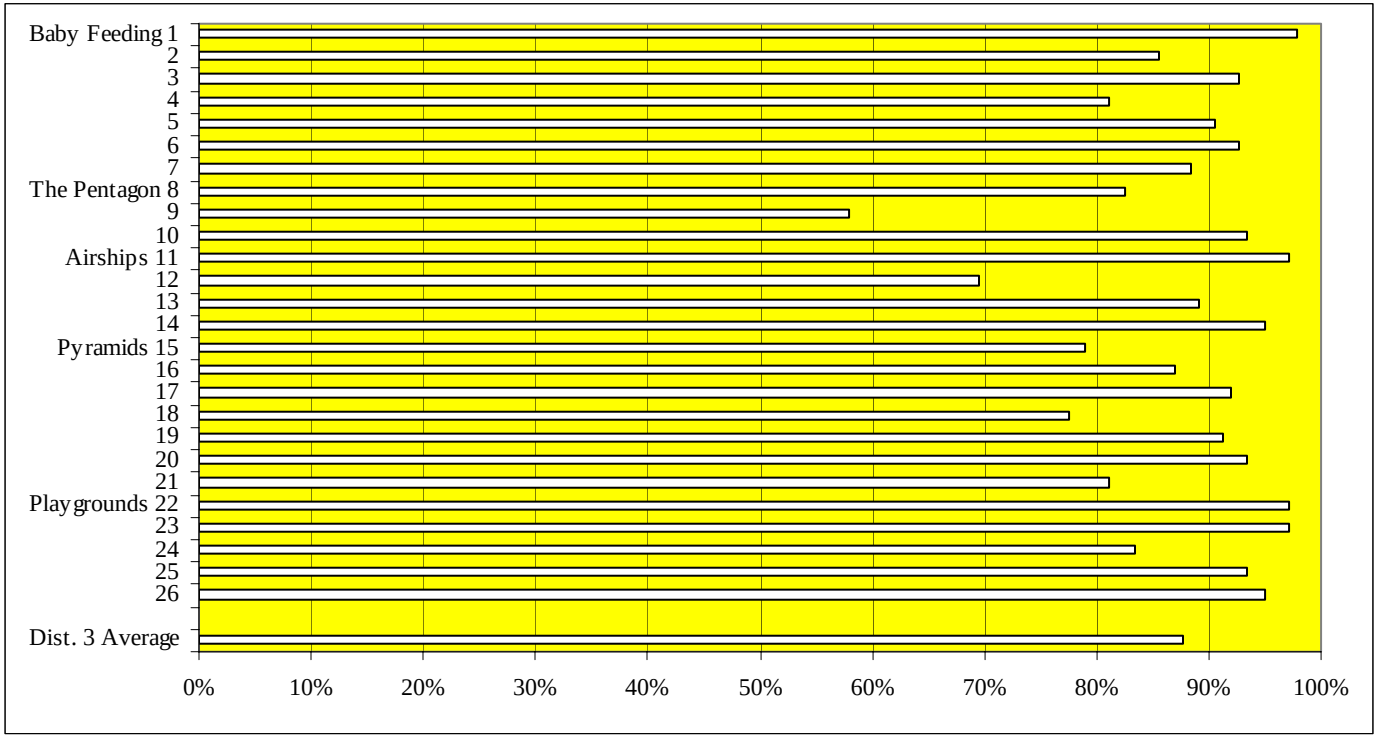

Figure 28. District 3 interrater agreement on Grade 7 Problem Solving Assessment, by item.

District 4. In District 4, the interrater agreement on the Grade 7 Problem Solving Assessment was high (93.03%; see Figure 29 and Table B5 in the Appendix). Interrater agreement was over 80% on all contexts, and four-fifths of the contexts were over 90%. The interrater agreement ranged from a low of 87.83% on the "The Pentagon" context (Items 7–10) to a high of 95.87% on the "Playgrounds" context (Items 22–26).

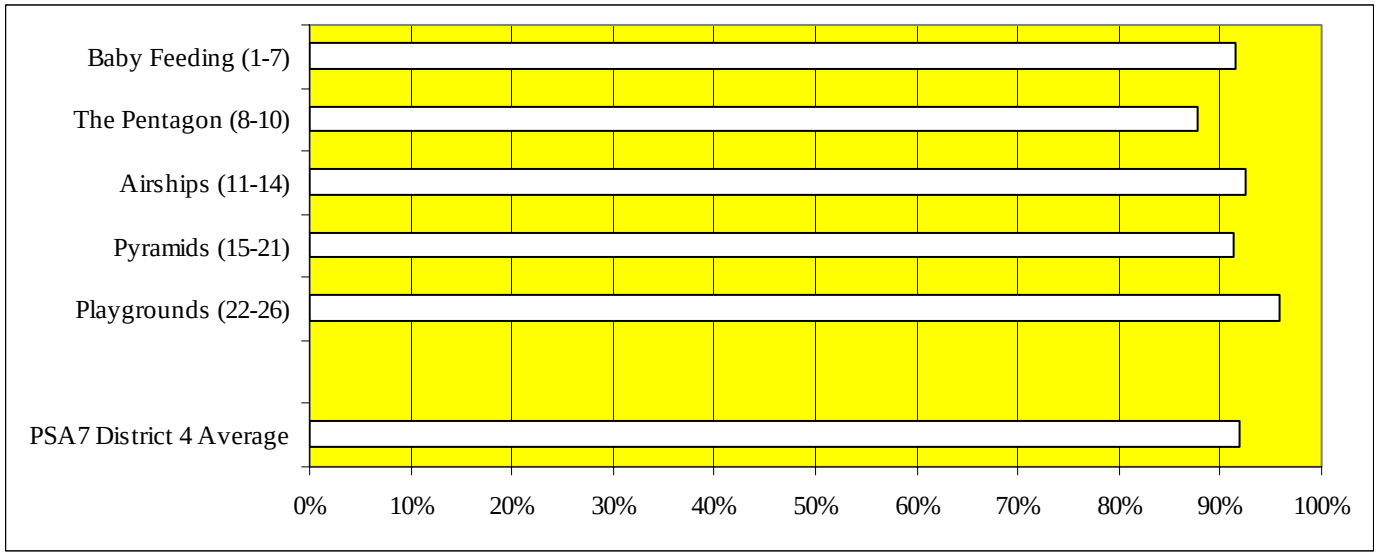

Figure 29. District 4 interrater agreement on Grade 7 Problem Solving Assessment, by context.

All but one of the individual items had interrater agreement over 80%, and two-thirds of the items had agreement over 90% (see Figure 30 and Table B5 in the Appendix). The interrater agreement on individual items ranged from a low of 76.72% on Item 12 from the "Airships" to a high of 98.94% on Item 22 from the "Playgrounds" context.

Figure 30. District 4 interrater agreement on Grade 7 Problem Solving Assessment, by item.

Across districts. There were some large differences (5% or greater) in interrater agreement across the districts (see Figure 31 and Table B5 in the Appendix). Districts 1 and 2 had no large differences in interrater agreement. In District 3, interrater agreement was lower than other districts overall and on "The Pentagon," and the "Pyramids" context. In District 4, interrater agreement was higher than the other districts on "The Pentagon" context.

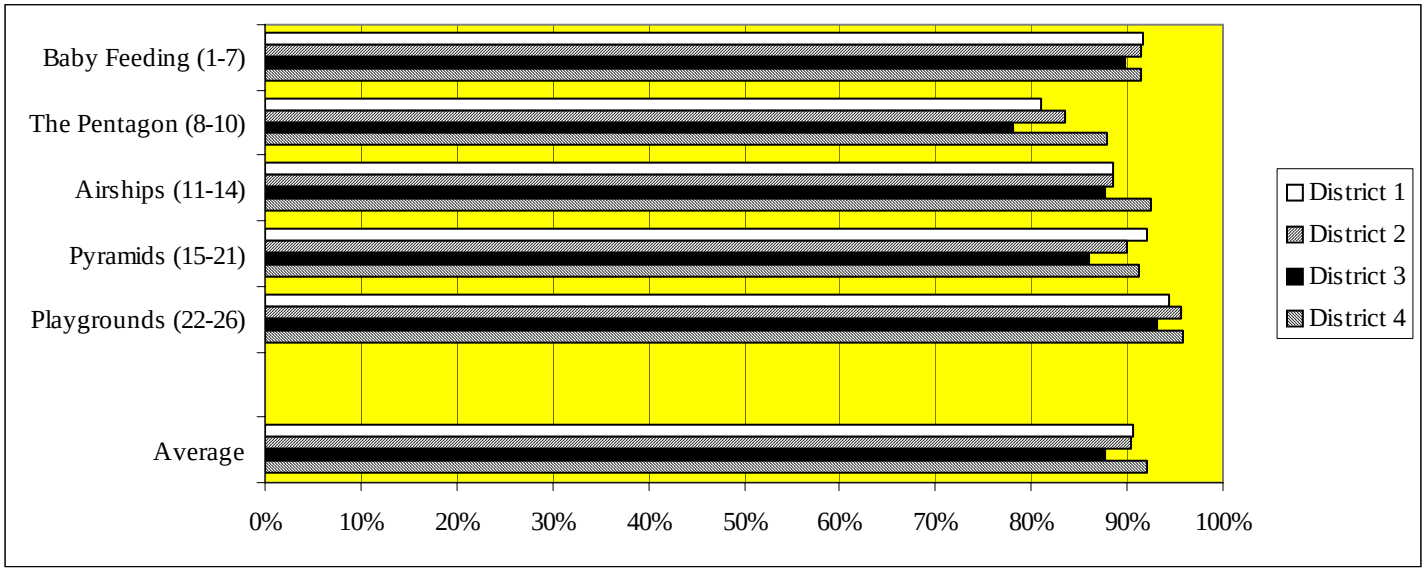

Figure 31. Across district interrater agreement on Grade 7 Problem Solving Assessment, by context.

Some individual items from each district had large (5% or greater) differences in interrater agreement (see Table 7 and Table B5 in the Appendix). In District 1, interrater agreement was higher than other in districts on Items 7 from the "Baby Feeding" context and on Items 15 and 16, from the "Pyramids" context and high on Item 8 from "The Pentagon" context and Item 24 from the "Playgrounds" context. In District 2, interrater agreement was lower than the other districts on Item 19 from the "Pyramids" context and high on Item 8 from "The Pentagon" context. In District 3, interrater agreement was low on Items 4 and 5 from the "Baby Feeding" context, Item 8 from "the Pentagon" context, Items 16, 17, 18, and 21 from the "Pyramids" context, Item 24 from the "Playgrounds" context, and very low on Item 9 from "The Pentagon" context. In District 4, interrater agreement was low on Item 7 from the "Baby Feeding" context, high on Item 4 from the "Baby Feeding" context, Items 12 and 13 from the "Airships" context, Item 21 from the "Pyramids" context, and very high on Item 9 from "The Pentagon" context.

Table 7

Interrater Agreement on Grade 7 Problem Solving Assessment by Item in all Districts

Context	Item Number	District 1	District 2	District 3	District 4
Baby Feeding	1	99.20%	96.80%	97.83%	96.30%
	2	85.54%	86.40%	85.51%	89.42%
	3	92.77%	93.20%	92.75%	88.89%
	4	83.53%	85.60%	***81.16%*** [8]	**88.89%** [9]
	5	95.18%	94.00%	***90.58%***	95.77%
	6	94.78%	95.20%	92.75%	95.77%
	7	**91.16%**	89.20%	88.41%	***85.19%***
The Pentagon	8	***81.12%***	**86.80%**	***82.61%***	**87.30%**
	9	69.08%	70.80%	***57.97%***	**82.01%**
	10	92.77%	92.80%	93.48%	94.18%
Airships	11	97.99%	98.00%	97.10%	98.41%
	12	68.67%	65.20%	69.57%	**76.72%**
	13	92.77%	93.60%	89.13%	**97.88%**
	14	94.78%	97.20%	94.93%	96.83%
Pyramids	15	**85.54%**	81.60%	78.99%	80.42%
	16	**97.19%**	94.00%	***86.96%***	90.48%
	17	96.39%	96.00%	***92.03%***	97.88%
	18	86.35%	84.00%	***77.54%***	85.19%
	19	92.77%	***87.60%***	91.30%	93.65%
	20	93.57%	92.00%	93.48%	94.18%
	21	92.77%	94.00%	***81.16%***	**96.83%**
Playgrounds	22	98.39%	94.80%	97.10%	98.94%
	23	95.98%	96.40%	97.10%	95.77%
	24	***86.35%***	92.80%	***83.33%***	91.01%
	25	97.59%	97.60%	93.48%	98.41%
	26	93.98%	96.00%	94.93%	95.24%
Average		90.62%	90.45%	87.74%	91.98%

[8] Percentage in bold with italics indicates lower differences (5% or greater) in interrater agreement.

[9] Percentage in bold indicates higher differences (5% or greater) in interrater agreement.

The large differences in interrater agreement across districts were most likely due to differences in (a) presentation and interpretation of rubrics during each scoring institute; (b) initial item scored at each institute; (c) content study teachers taught; (d) time of day items were scored; and (e) items eliciting a higher level of reasoning being left blank (District 3 had far fewer nonresponses, and District 4 more than the other districts). In addition, teachers in District 3 did not teach as many seventh-grade units, teaching instead a combination of sixth- and seventh-grade units, which might have affected interrater agreement.

Interrater Reliability by Program (Conventional Curricula or Mathematics in Context Classes)

Conventional curricula. The interrater agreement on the Grade 7 Problem Solving Assessment from conventional curricula was high (91.44%; see Figure 32 and Table B6 in the Appendix). Interrater agreement was over 80% on all contexts and over 90% on three-fifths of the contexts. The interrater agreement ranged from a low of 82.35% on "The Pentagon" context (Items 8–10) to a high of 95.63% on the "Playgrounds" context (Items 22–26).

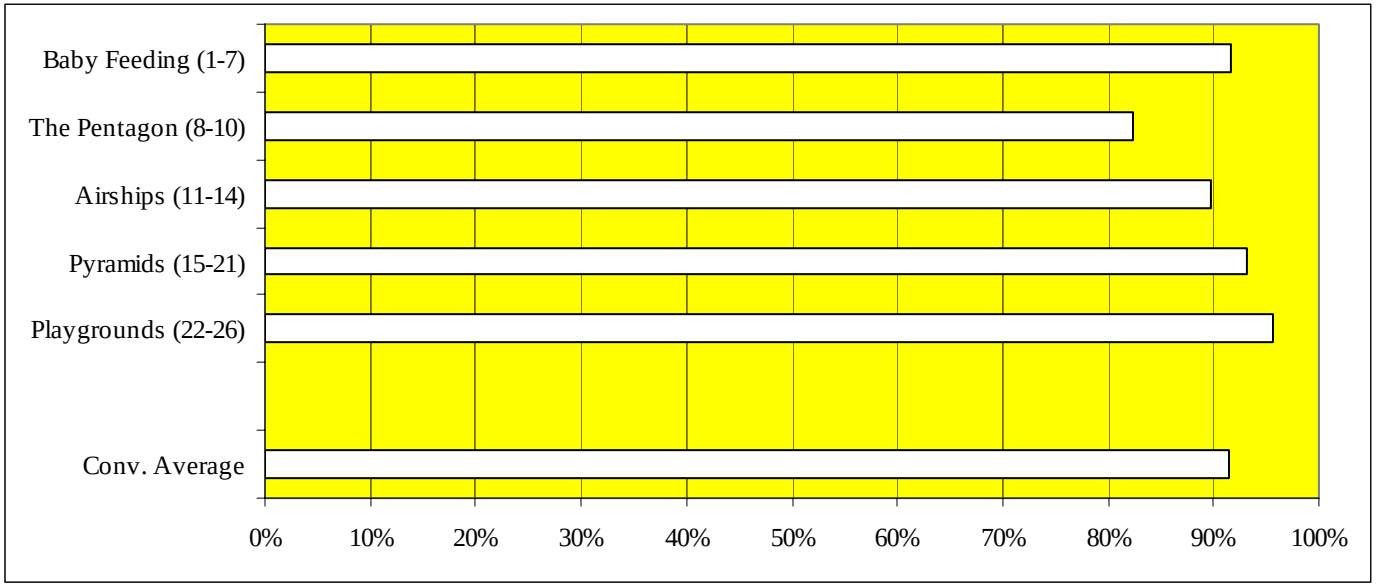

Figure 32. Interrater agreement on Grade 7 Problem Solving Assessment, by context: Conventional curricula.

All but two of the individual items had interrater agreement over 80%, and about two-thirds of the items had agreement over 90% (see Figure 33 and Table B6 in the Appendix). The interrater agreement on individual items ranged from a low of 68.07% on Item 9, "The Pentagon" context to a high of 100.00% on 2 items (Item 17 from the "Pyramids" context and Item 25 from the "Playgrounds" context). The other item with low interrater agreement was Item 12 from the "Airships" context at 76.47%.

Figure 33. Interrater agreement on Grade 7 Problem Solving Assessment, by item: Conventional curricula.

Mathematics in Context classes. The interrater agreement on the Grade 7 Problem Solving Assessment from *Mathematics in Context* classes was high (90.22%; see Figure 34 and Table B6 in the Appendix). Interrater agreement was over 80% on all contexts and over 90% on two-fifths of the contexts. The interrater agreement ranged from a low of 82.89% on "The Pentagon" context (Items 8–10) to a high of 94.77% on the "Playgrounds" context (Items 22–26).

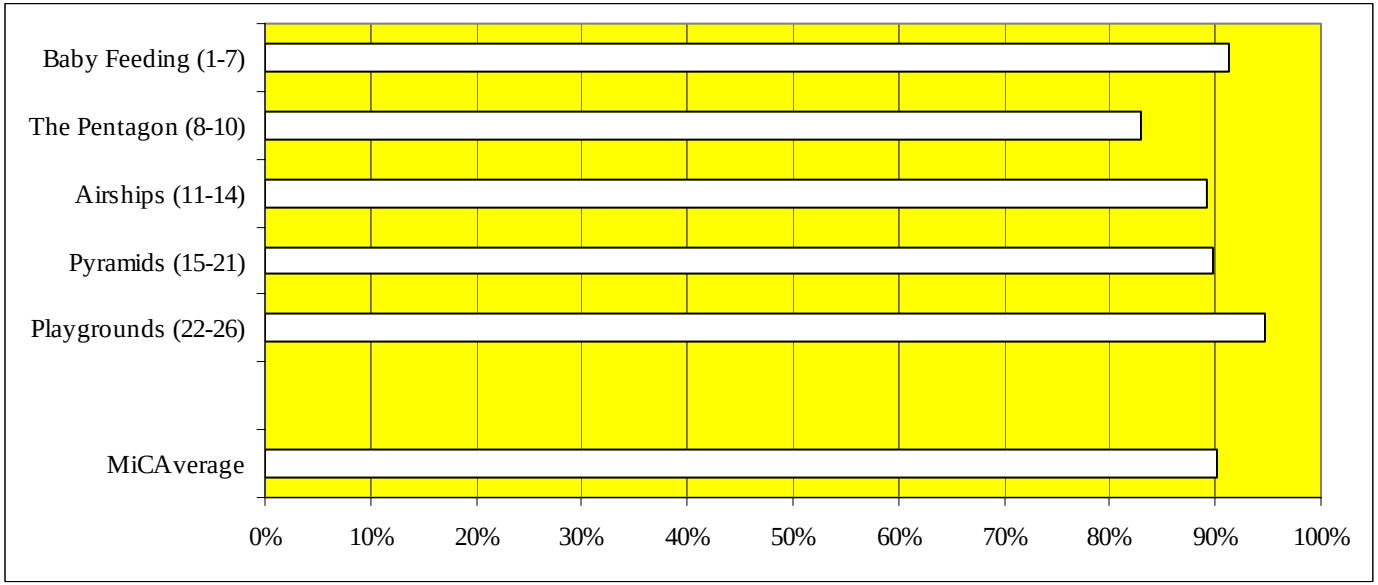

Figure 34. Interrater agreement on Grade 7 Problem Solving Assessment, by context: *Mathematics in Context* classes.

All but two of the individual items had interrater agreement over 80%, and about two-thirds of the items had agreement over 90% (see Figure 35 and Table B6 in the Appendix). The interrater agreement on individual items ranged from a low of 68.46% on Item 12 from the "Airships" context to a high of 98.02% on Item 11 from the "Airships" context.

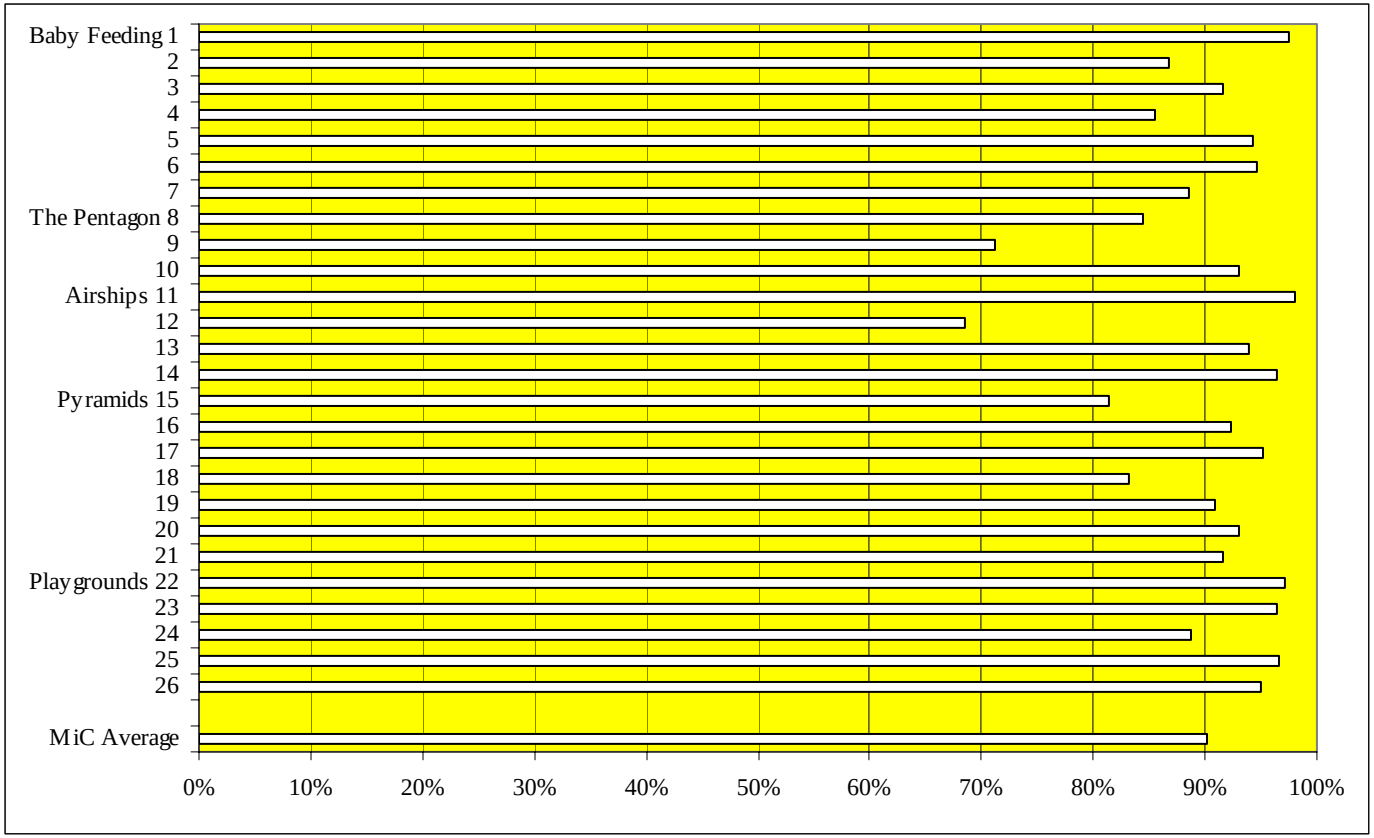

Figure 35. Interrater agreement on Grade 7 Problem Solving Assessment, by item: *Mathematics in Context* classes.

Across programs. Overall, the interrater agreement in conventional curricula and *Mathematics in Context* classes was similar (see Figure 36 and Table B6 in the Appendix). The average interrater agreement for conventional curricula was 91.44% and for *Mathematics in Context* classrooms 90.22%. The interrater agreement was similar across programs on all contexts.

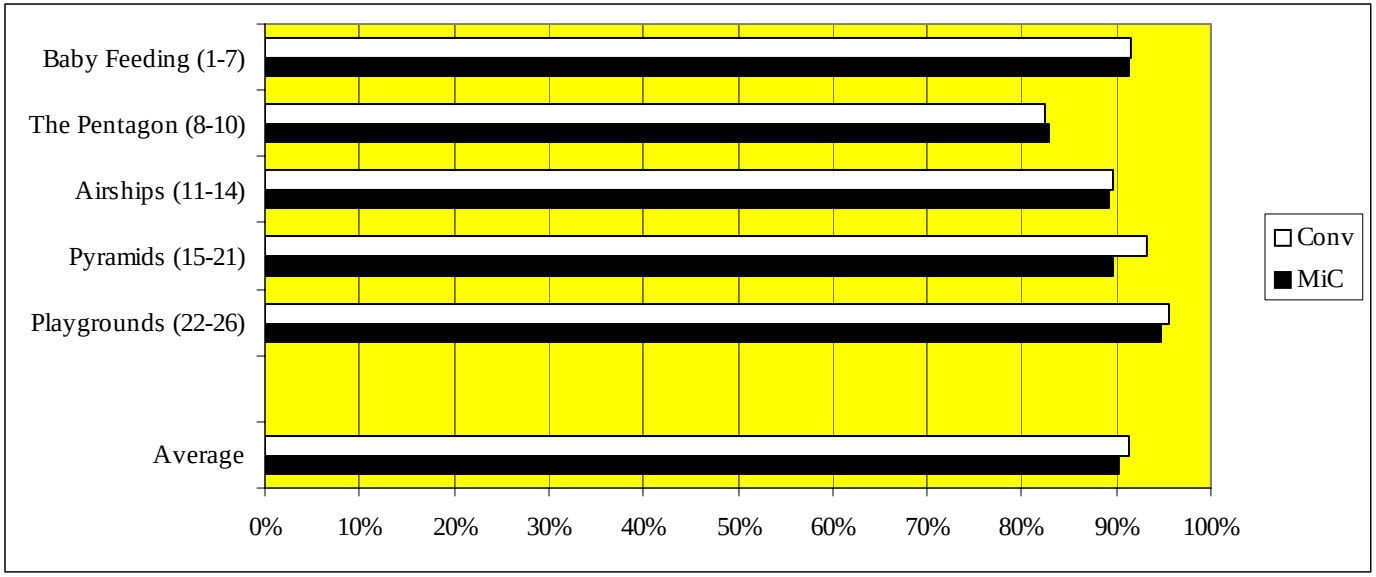

Figure 36. Interrater agreement on Grade 7 Problem Solving Assessment, by context: Conventional curricula and *Mathematics in Context* classes.

The interrater agreement on some individual items also revealed a discrepancy between conventional curricula and *Mathematics in Context* classes (see Figure 37 and Table B6 in the Appendix). Assessments from conventional curricula had much higher agreement (5% or greater) on Item 12 from the "Airships" context and Items 18 from the "Pyramids" context.

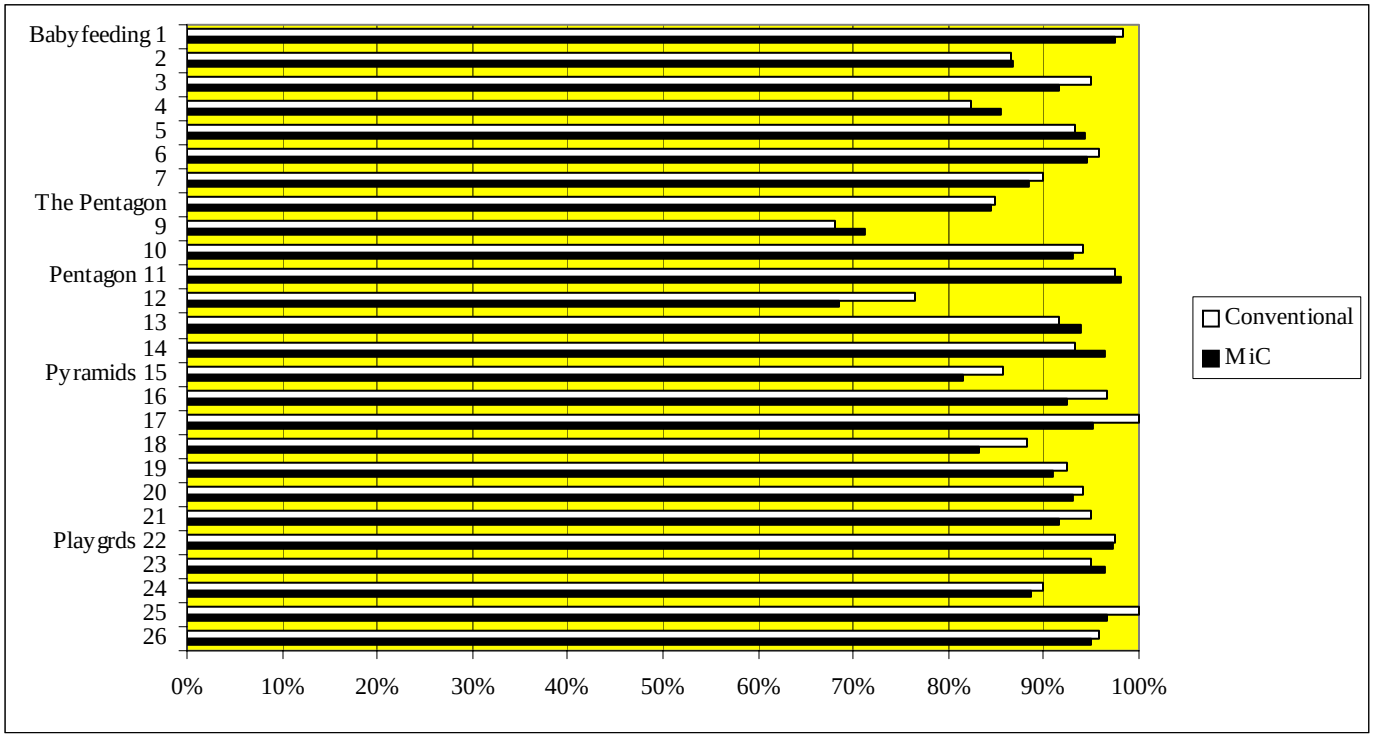

Figure 37. Interrater agreement on Grade 7 Problem Solving Assessment, by item: Conventional curricula and *Mathematics in Context* classes.

The large differences in interrater agreement were most likely due to differences in (a) initial item scored at each institute; (b) content study teachers taught; (c) time of day items were scored; (d) raters' interpretation of student work; and (e) proportion of student nonresponses at the end of section for the day.[10]

[10] Students were given Items 1–13 the first day and 14–24 the second day. Many students left Items 11–13 and 22–24 blank.

Grade 8

Overall Interrater Reliability

The interrater agreement on the Grade 8 Problem Solving Assessment was high (92.90% see Figure 38 and Appendix B7). Interrater agreement was over 80% on all contexts and over 90% on five out of the seven contexts. The interrater agreement ranged from a low of 82.45% on the "Club Members" context (Item 1) to a high of 96.35% on the "Seesaw" context (Items 8–9).

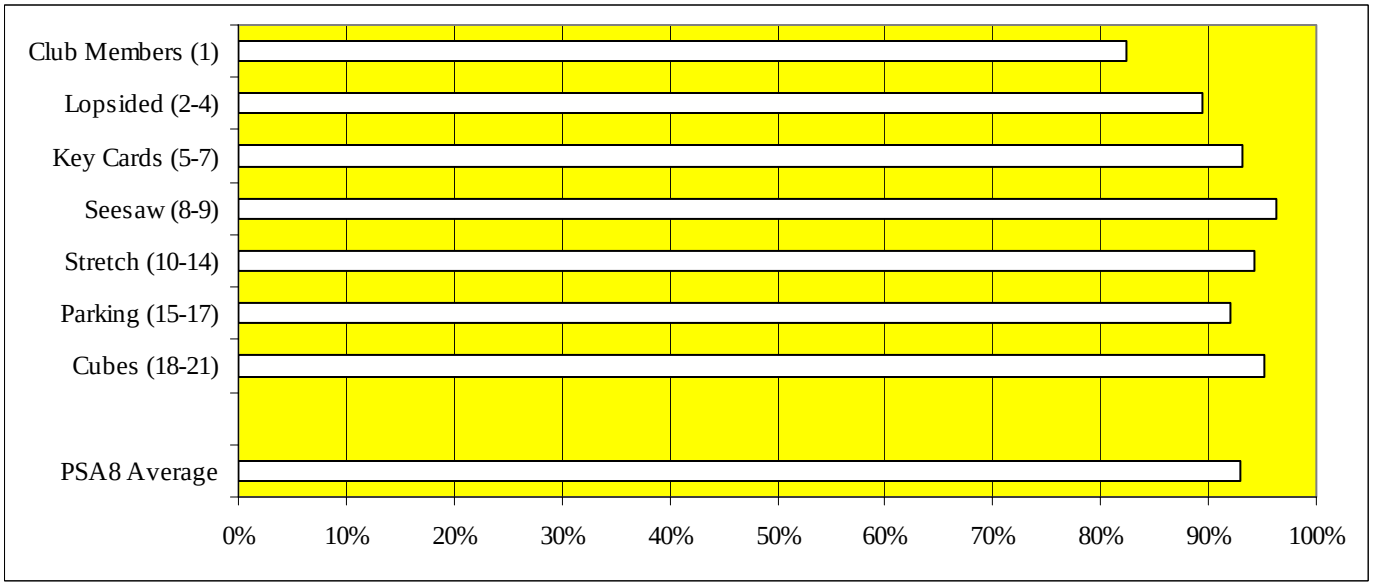

Figure 38. Interrater agreement on Grade 8 Problem Solving Assessment, by context.

All of the individual items had interrater agreement over 80%, and about three-quarters the items had agreement over 90% (see Figure 39 and Table B7 in the Appendix). The interrater agreement on individual items ranged from a low of 82.45% on Item 1 from "Club Members" context to a high of 99.01% on Item 12 from the "Stretch" context.

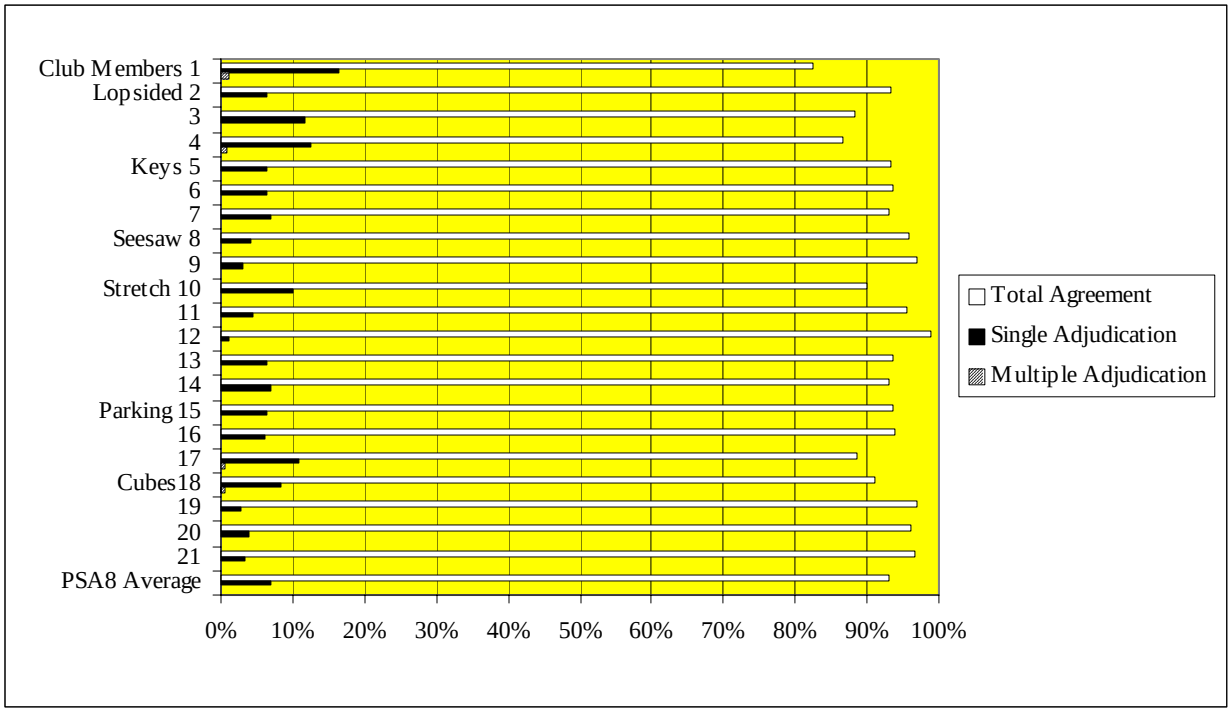

Figure 39. Interrater agreement on Grade 8 Problem Solving Assessment, by item.

The incidence of single adjudication was inversely proportional to the incidence of total agreement (see Figure 39 and Table B7 in the Appendix). The percentage of single adjudication ranged from a low of 0.99% on Item 12 from the "Stretch" context to a high of 16.57% on Item 1 from "Club Members" context. The incidence of multiple adjudication was very low. It ranged from 0% on 12 items (Item 3 from the "Lopsided" context, Items 6 and 7 from the "Key Cards" context, Item 8 from the "Seesaw" context, Items 10, 11, 12, and 13 from the "Stretch" context, Items 15 and 16 from the "Parking" context, and Items 20 and 21 from the "Cubes" context) to a high of 16.72% on Item 12 from the "Airships" context.

Factors that contributed to the high interrater agreement and low adjudication include (a) high quality training for raters, (b) rater experience, (c) well-defined and clarified rubrics, (d) effective scoring procedures; and (e) many items with nonresponses or incorrect responses (Items 12, 16, 19, and 20). The factor that contributed to the lower interrater agreement (and higher adjudication) was subtleties in graphs/figures which students may not have marked clearly and which some raters did not recognize in the first round of scoring (Item 1 from the "Club Members" context).

Interrater Reliability by Districts

District 1. In District 1, the interrater agreement on the Grade 8 Problem Solving Assessment was high (93.82%; see Figure 40 and Table B8 in the Appendix). Interrater agreement was over 80% on all contexts and over 90% on six out of the seven contexts. The interrater agreement ranged from a low of 85.71% on "Club Members" context (Item 1) to a high of 95.68% on the "Cubes" context (Items 18–21).

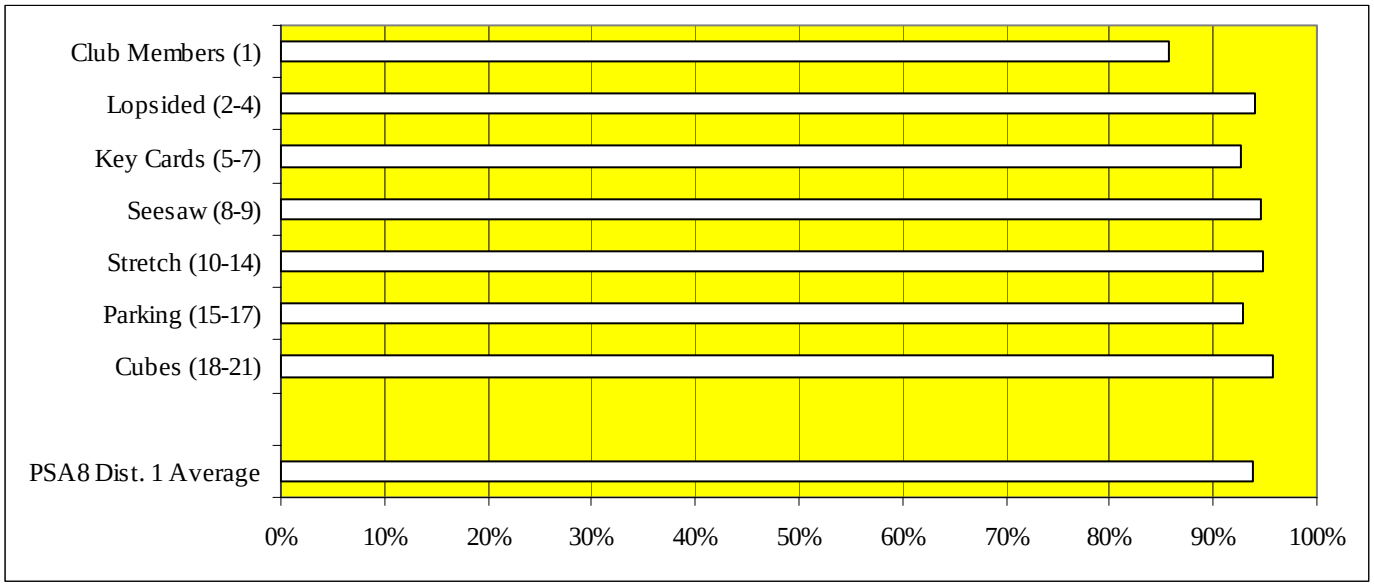

Figure 40. District 1 interrater agreement on Grade 8 Problem Solving Assessment, by context.

All of the individual items had interrater agreement over 80%, and all but two of the items had agreement over 90% (see Figure 41 and Table B8 in the Appendix). The interrater agreement on individual items ranged from a low of 85.71% on Item 1 from "Club Members" to a high of 99.40% on Item 12 from the "Stretch" context.

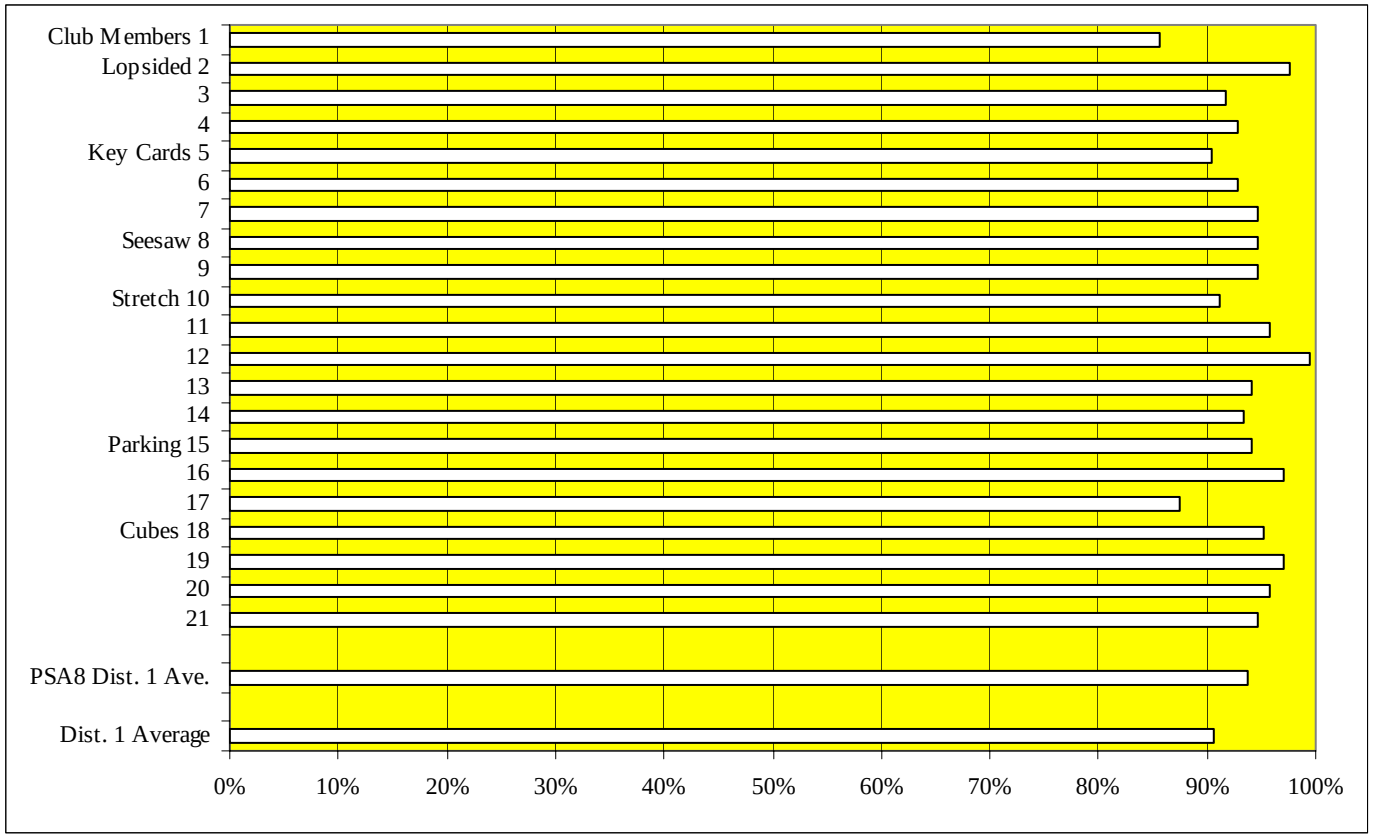

Figure 41. District 1 *i*nterrater agreement on Grade 8 Problem Solving Assessment, by item.

District 2. In District 2, the interrater agreement on the Grade 8 Problem Solving Assessment was high (92.55%; see Figure 42 and Table B8 in the Appendix). Interrater agreement was over 80% on all contexts, and it was over 90% on five out of the seven of the contexts. The interrater agreement ranged from a low of 82.86% on the "Club Members" context (Item 1) to a high of 98.57% on the "Seesaw" context (Items 8–9).

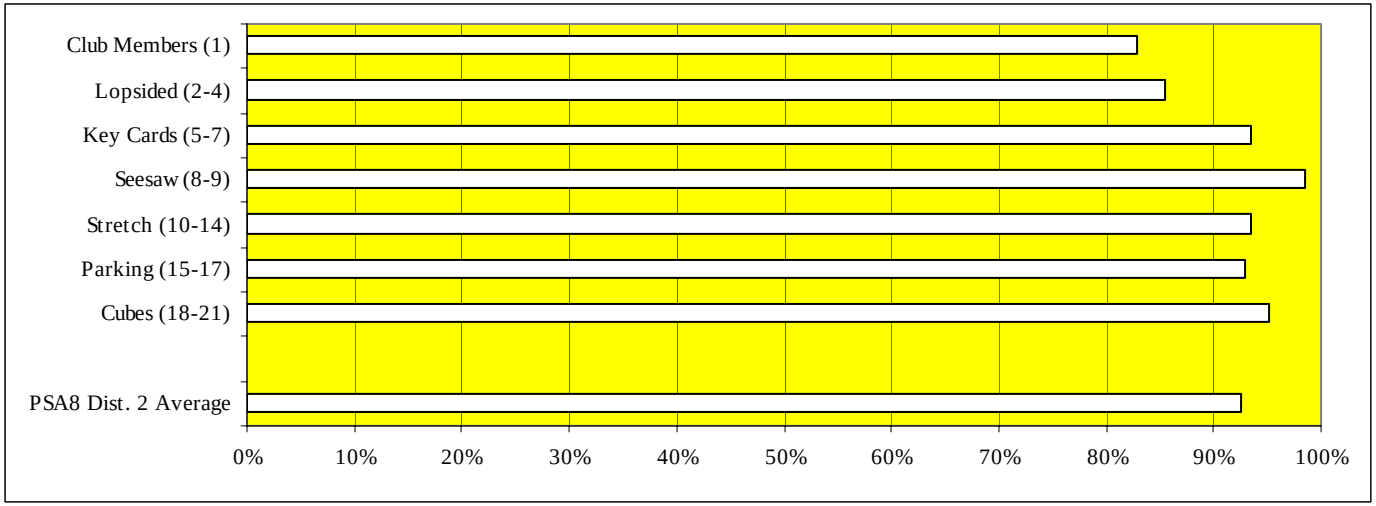

Figure 42. District 2 interrater agreement on Grade 8 Problem Solving Assessment, by context.

All of the individual items had interrater agreement 80% or higher, and almost three-quarters of the items had agreement over 90% (see Figure 43 and Table B8 in the Appendix). The interrater agreement on individual items ranged from a low of 80.00% on Item 4 from the "Lopsided" to a high of 100.00% on Item 12 from the "Stretch" context.

Figure 43. District 2 interrater agreement on Grade 8 Problem Solving Assessment, by item.

District 3. In District 3, the interrater agreement on the Grade 8 Problem Solving Assessment was high (89.74%; see Figure 44 and Table B8 in the Appendix). Interrater agreement was over 80% on all of the contexts and over 90% on three out of the seven contexts. The interrater agreement ranged from a low of 81.69% on the "Lopsided" context (Items 2–4) to a high of 93.66% on two contexts, "Seesaw" context (Items 8–9) and "Cubes" (Items 18–21).

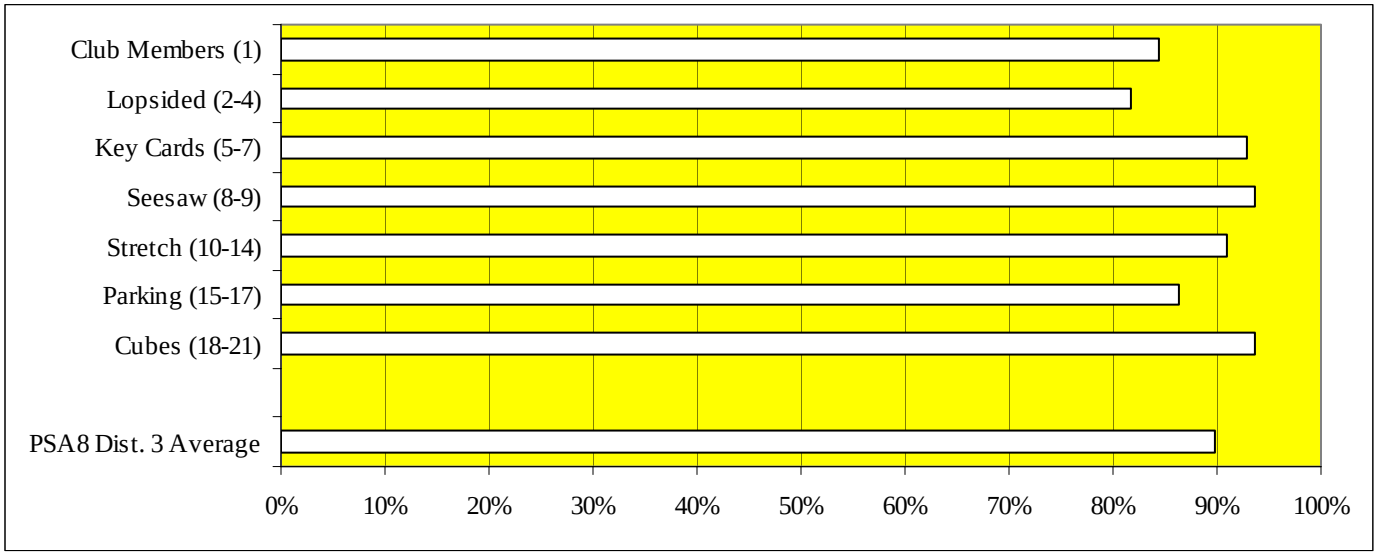

Figure 44. District 3 interrater agreement on Grade 8 Problem Solving Assessment, by context.

All but one of the individual items had interrater agreement over 80%, and more than half of the items had agreement over 90% (see Figure 45 and Table B8 in the Appendix). The interrater agreement on individual items ranged from a low of 76.06% on Item 4 from the "Lopsided" to a high of 98.59% on two items, Item 5 from the "Key Cards" context and Item 20 from the "Cubes" context.

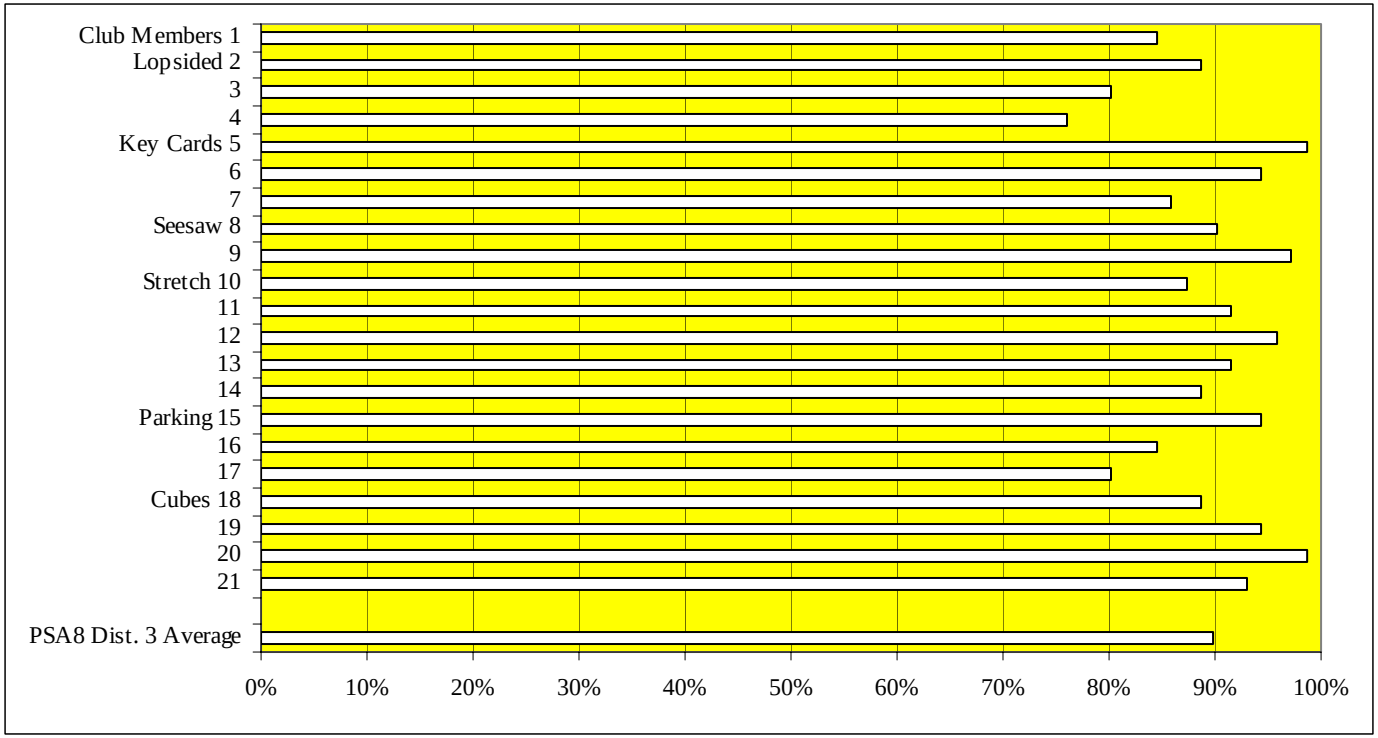

Figure 45. District 3 interrater agreement on Grade 8 Problem Solving Assessment, by item.

District 4. In District 4, the interrater agreement on the Grade 8 Problem Solving Assessment was high (93.82%; see Figure 46 and Table B8 in the Appendix). Interrater agreement was over 80% on all but 1 context, and it was over 90% on six out of the seven of the contexts. The interrater agreement ranged from a low of 76.56% on the "Club Members" context (Item 1) to a high of 97.66% on the "Seesaw" context (Items 8–9).

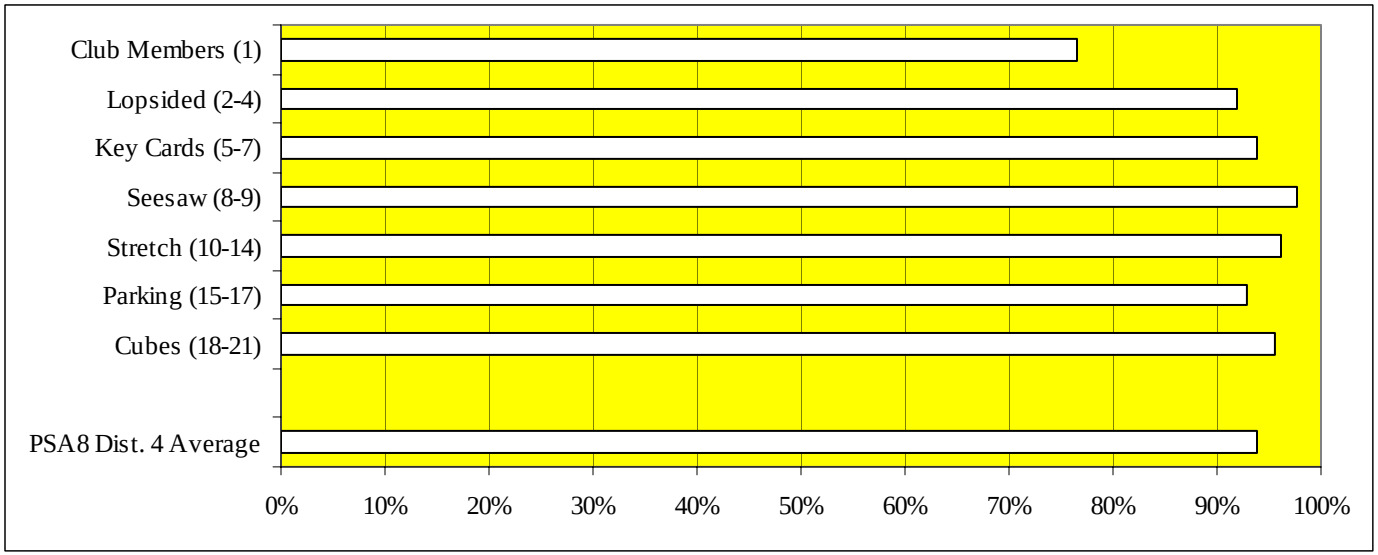

Figure 46. District 4 interrater agreement on Grade 8 Problem Solving Assessment, by context.

All but one of the individual items had interrater agreement over 80%, and four-fifths the items had agreement over 90% (see Figure 47 and Table B8 in the Appendix). The interrater agreement on individual items ranged from a low of 76.56% on Item 1 from the "Club Members" to a high of 99.22% on 2 items (Item 12 from the "Stretch" context and Item 21 from the "Cubes" context).

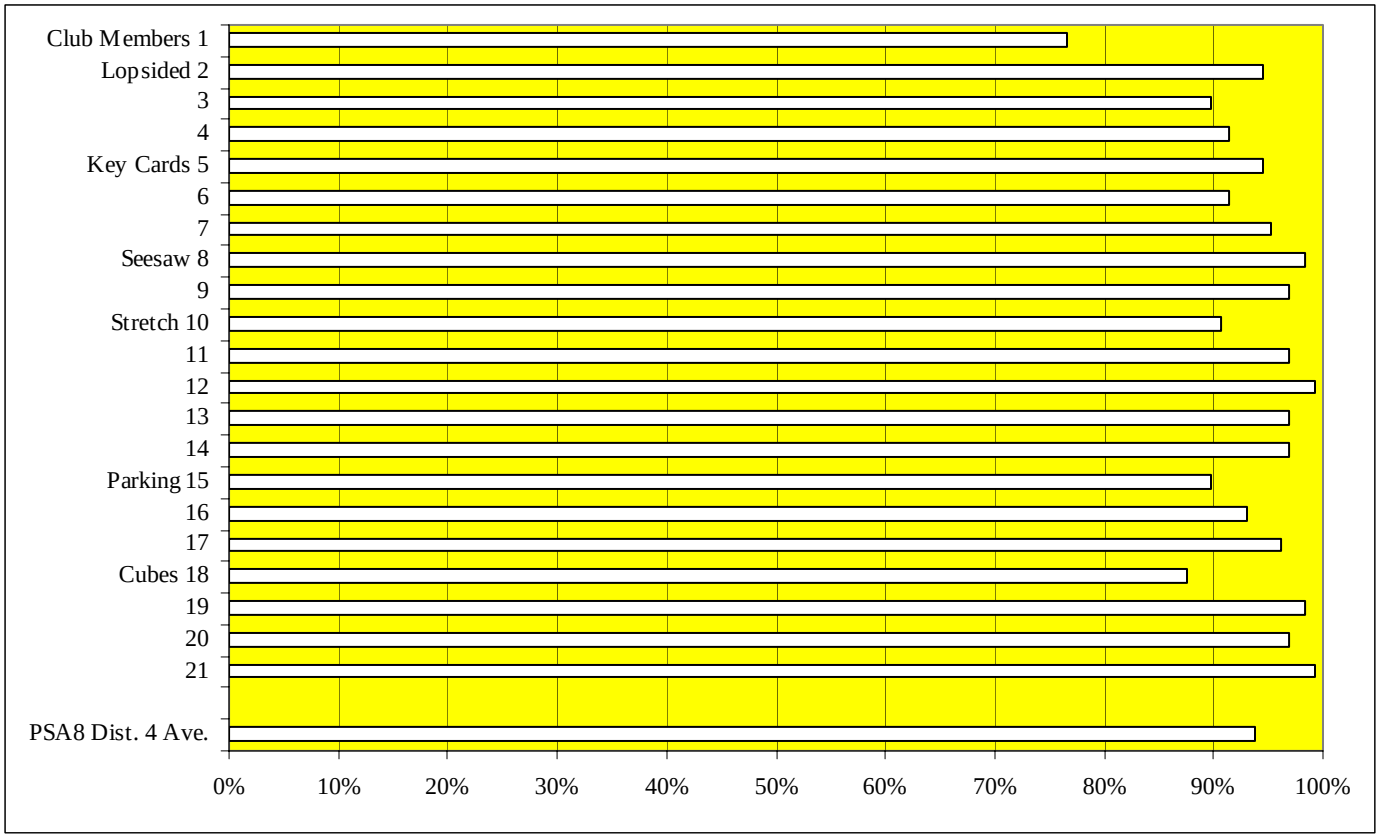

Figure 47. District 4 interrater agreement on Grade 8 Problem Solving Assessment, by item.

Across districts. There were some large differences (5% or greater) in interrater agreement across the districts (see Figure 48 and Table B8 in the Appendix). In District 1, interrater agreement was higher than other districts on the "Lopsided" context. In District 2, interrater agreement was high on the "Seesaw" context. In District 3, interrater agreement was lower than the other districts on the "Lopsided," "Stretch," and "Parking" contexts. In District 4, interrater agreement was much lower than the other districts on the "Club Members" context.

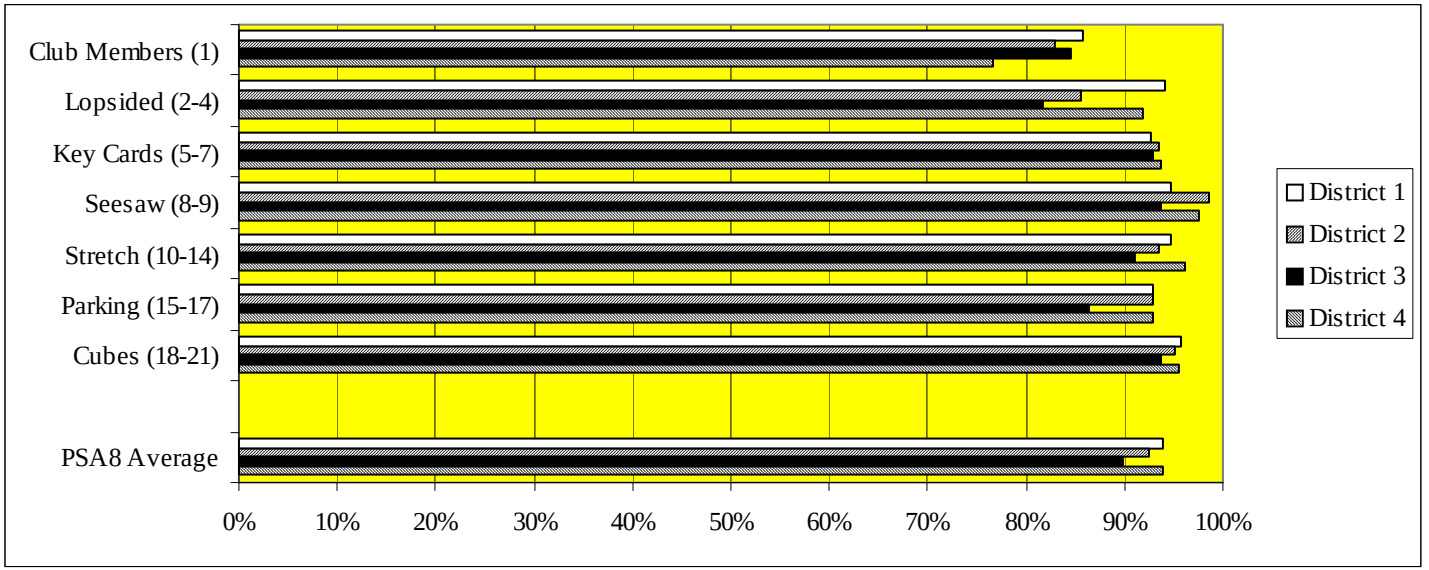

Figure 48. Across district interrater agreement on Grade 8 Problem Solving Assessment, by context.

Some individual items from each district have large (5% or greater) differences in interrater agreement (see Table 8 and Table B8 in the Appendix). In District 1, interrater agreement was higher than other districts on Item 18 from the "Cubes" context and high on Items 2 and 4 from the "Lopsided" context and Item 16 on the "Parking" context. In District 2, interrater agreement was lower than the other districts on Item 2 from the "Lopsided" context. In District 3, interrater agreement was much lower than other districts on Items 3 and 4 from the "Lopsided" context, Item 7 from the "Key Cards" context, Item 8 from the "Seesaw" context, Items 11 and 14 from the "Stretch" context, Items 16 and 17 from the "Parking" context; low on Item 2 from the "Lopsided" context and Item 21 from the "Cubes" context; and high on Item 5 from the "Key Cards" context. District 4, interrater agreement was lower than other districts on Item 1 from the "Club Members" context and Item 15 from the "Parking" context; higher than the other districts on Item 17 from the "Parking" context; and high on Items 2 and 4 from the "Lopsided" context, Item 8 from the "Seesaw" context, and Items 13 or 14 on the "Stretch" context.

Table 8
Interrater Agreement on Grade 7 Problem Solving Assessment by Item in all Districts

Context	Item Number	District 1	District 2	District 3	District 4
Club Members	1	85.71%	82.86%	84.51%	***76.56%*** [11]
Lopsided	2	**97.62%** [12]	***89.29%***	***88.73%***	**94.53%**
	3	91.67%	87.14%	***80.28%***	89.84%
	4	**92.86%**	80.00%	***76.06%***	**91.41%**
Key Cards	5	90.48%	92.86%	**98.59%**	94.53%
	6	92.86%	95.71%	94.37%	91.41%
	7	94.64%	92.14%	***85.92%***	95.31%
Seesaw	8	94.64%	**97.86%**	***90.14%***	**98.44%**
	9	94.64%	99.29%	97.18%	96.88%
Stretch	10	91.07%	89.29%	87.32%	90.63%
	11	95.83%	96.43%	***91.55%***	96.88%
	12	99.40%	100.00%	95.77%	99.22%
	13	94.05%	90.71%	91.55%	**96.88%**
	14	93.45%	90.71%	***88.73%***	**96.88%**
Parking	15	94.05%	95.71%	94.37%	***89.84%***
	16	**97.02%**	95.71%	***84.51%***	92.97%
	17	87.50%	87.14%	***80.28%***	**96.09%**
Cubes	18	**95.24%**	90.71%	88.73%	87.50%
	19	97.02%	97.14%	94.37%	98.44%
	20	95.83%	94.29%	98.59%	96.88%
	21	94.64%	98.57%	***92.96%***	99.22%
Average		93.82%	92.55%	89.74%	93.82%

[11] Percentage in bold with italics indicates lower differences (5% or greater) in interrater agreement.
[12] Percentage in bold indicates higher differences (5% or greater) in interrater agreement

The large differences in interrater agreement were most likely due to differences in (a) content study teachers taught; (b) raters' interpretation of student work; and (c) proportion of student nonresponse. (In District 3, very few items were left blank; raters might also have had a more difficult time rating unexpected answers. In District 4, there were many more student nonresponses, which provided higher interrater agreement.)

Interrater Reliability by Program (Conventional or Mathematics in Context)

Conventional curricula. The interrater agreement on the Grade 8 Problem Solving Assessment from conventional curricula was high (94.24%; see Figure 49 and Table B9 in the Appendix). Interrater agreement was over 80% on all of the contexts and over 90% on six out of the seven of the contexts. The interrater agreement ranged from a low of 87.62% on the "Club Members" context (Item 1) to a high of 97.14% on the "Seesaw" context (Item 8–9).

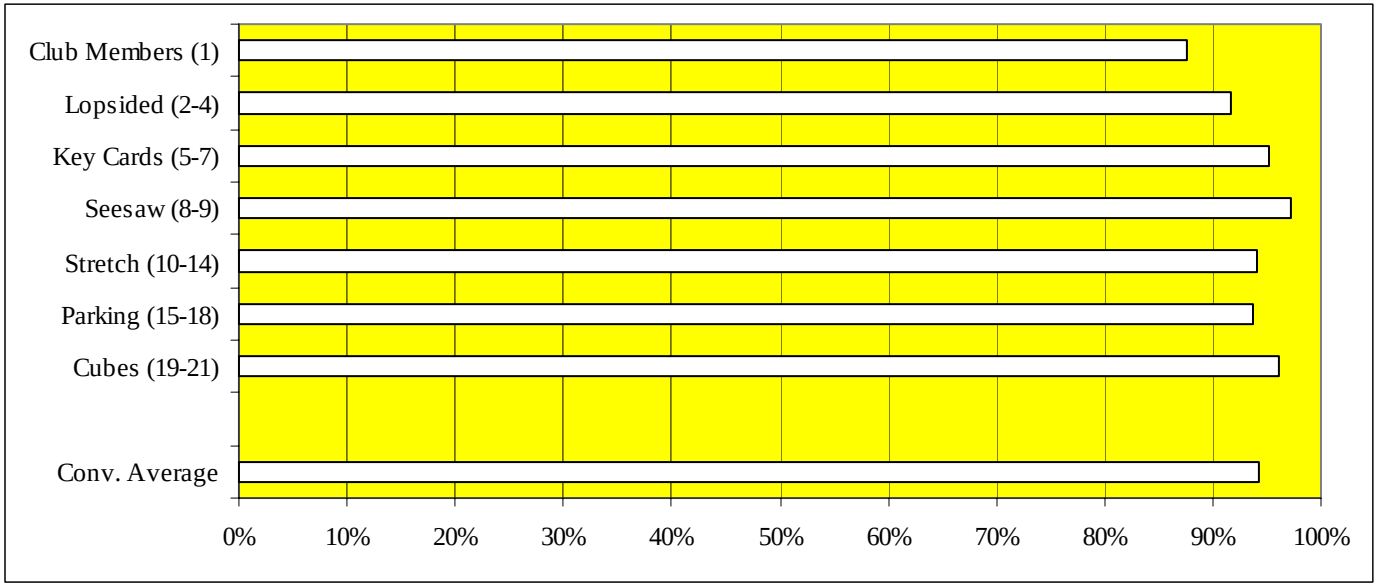

Figure 49. Interrater agreement on Grade 8 Problem Solving Assessment, by context: Conventional curricula.

All of the individual items had interrater agreement over 80%, and almost four-fifths of the items had agreement over 90% (see Figure 50 and Table B9 in the Appendix). The interrater agreement on individual items ranged from a low of 85.71% on Item 10 from "Seesaw" context to a high of 99.05% on Item 12 from the "Stretch" context.

Figure 50. Interrater agreement on Grade 8 Problem Solving Assessment, by item: Conventional curricula.

Mathematic in Context classes. The interrater agreement on the Grade 8 Problem Solving Assessment from *Mathematics in Context* classes was high (92.55%; see Figure 51 and Table B9 in the Appendix). Interrater agreement was over 80% on all contexts. The interrater agreement ranged from a low of 81.09% on the "Club Members" context (Item 1) to a high of 96.14% on the "Seesaw" context.

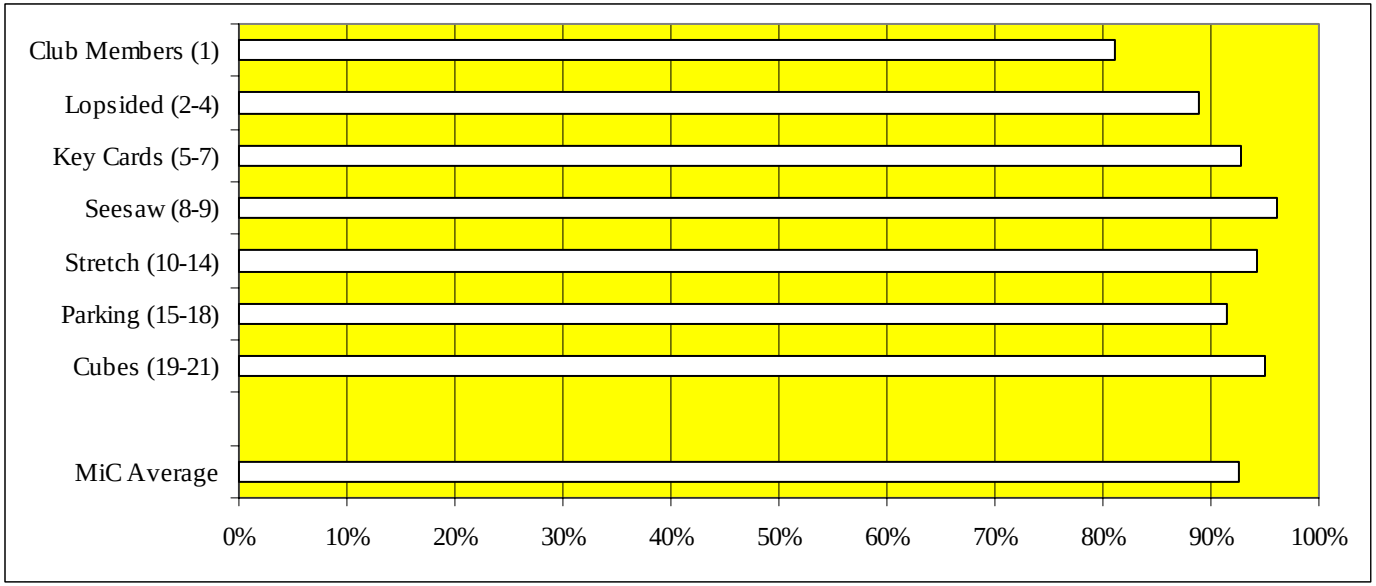

Figure 51. Interrater agreement on Grade 8 Problem Solving Assessment, by context: *Mathematics in Context* classes.

All of the individual items had interrater agreement over 80%, and four-fifths of the items had agreement over 90% (see Figure 52 and Table B9 in the Appendix). The interrater agreement on individual items ranged from a low of 81.09% on Item 1 from the "Club Members" context to a high of 99.00% on Item 12 from the "Stretch" context.

Figure 52. Interrater agreement on Grade 8 Problem Solving Assessment, by item: *Mathematics in Context* classes.

Across programs. Overall, the interrater agreement in conventional curricula and *Mathematics in Context* classes is similar (see Figure 53 and Table B9 in the Appendix). The average interrater agreement for conventional curricula was 94.24% and 92.55% for *Mathematics in Context* classrooms. The interrater agreement for the "Club Members" context (Item 1) was much higher (5% or greater) for the conventional classes.

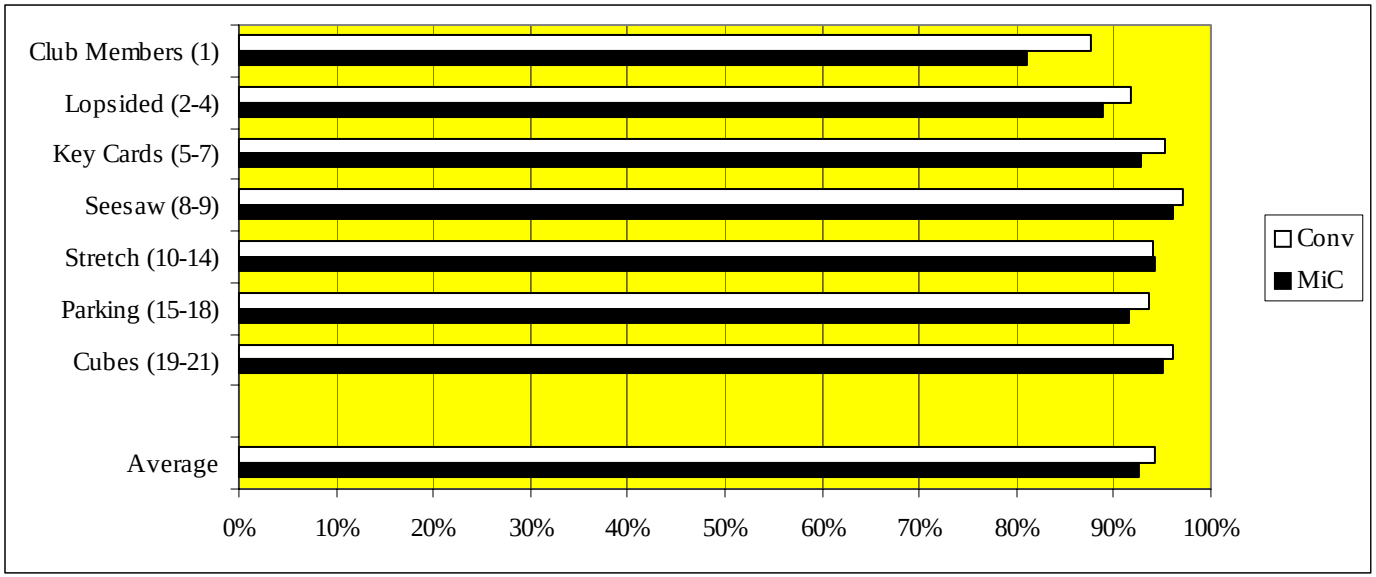

Figure 53. Interrater agreement on Grade 8 Problem Solving Assessment, by context: Conventional curricula and *Mathematics in Context* classes.

The interrater agreement on some individual items also revealed a difference between conventional curricula and *Mathematics in Context* classes (see Figure 54 and Table B9 in the Appendix). Assessments from conventional curricula had higher agreement (5% or greater) on Item 1 from the "Club Members" context. Assessments from the *Mathematics in Context* classes had higher agreement on Item 10 from the "Stretch" context.

The differences in interrater agreement were most likely due to differences in (a) content study teachers taught; (b) raters' interpretation of student work, and (c) proportion of student nonresponse.

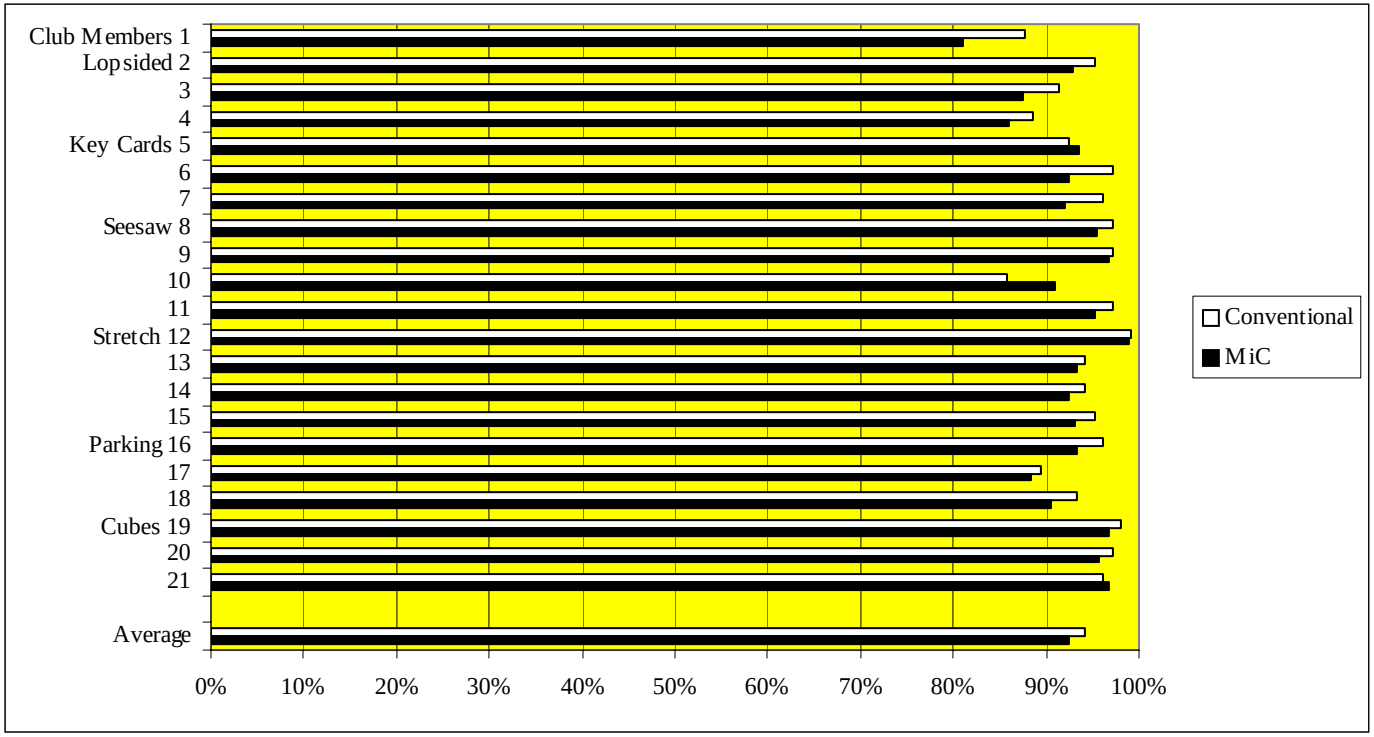

Figure 54. Interrater agreement on Grade 8 Problem Solving Assessment by item: Conventional and *Mathematics in Context* classes.

Conclusion

The interrater reliability was high for the Problem Solving Assessments. The factors that contributed to the high interrater agreement (and low adjudication) include (a) high quality training for raters; (b) rater experience; (c) well-defined and clarified rubrics; (d) effective scoring procedures; and (e) proportion of student nonresponse.

The factors contributing to the lower interrater agreement (and higher adjudication) include (a) multiple scoring criteria (some raters scored more leniently), (b) subtleties in graphs/figures, which students might not have marked clearly, (c) raters at the different sites may have been more perplexed with scoring certain items which were discussed more thoroughly at the later Madison scoring institutes, and (d) presenters at the different institutes where certain items were scored may have emphasized different criteria.

The differences in interrater agreement across districts were most likely due to differences in (a) presentation of rubrics during each scoring institute; (b) interpretation of rubrics during each scoring institute; (c) initial item scored at each scoring institute; (d) content study teachers taught; (e) time of day items were scored; (f) proportion of student nonresponse; and (g) number of items eliciting a higher level of reasoning that were left blank.

The differences in interrater agreement across programs (conventional curricula or *Mathematics in Context* classes) were most likely due to differences in (a) initial item scored at each institute; (b) content study teachers taught; (c) time of day items were scored; (d) raters' interpretation of student work; and (e) proportion of student nonresponse at the end of the assessment section completed on each day.

Interrater Reliability on External Assessments

All of the 1999 External Assessments were scored at one scoring institute in the summer, 1999 (see Table A1 in the Appendix). In contrast to the Problem Solving Assessment, six EA constructed-response items (anchor items) were repeated on the grade-specific assessment. For purposes of scoring, each set of anchor items was considered a context. On average, two contexts were scored each day. The rubrics used in scoring EA items were identical to rubrics used in the NAEP and TIMSS assessments. In general, EA rubrics were less complicated than PSA rubrics, but because they were anchor items recurring at each grade level, in most cases larger sets of assessments were scored for each EA context than for PSA contexts. In this section, interrater reliability is determined for each External Assessment by grade and context in three ways: (a) overall, (b) by districts and (c) by program (conventional curricula or *Mathematics in Context*).

Grade 6

Overall Interrater Reliability

The interrater agreement on the Grade 6 External Assessment was very high (94.67%; see Figure 55 and Appendix C1). Interrater agreement was over 80% on all items.[13] Three-quarters of the items had interrater agreement over 90%. The interrater agreement ranged from a low of 81.91% on Item 8 to a high of 99.44% on Item 24a.

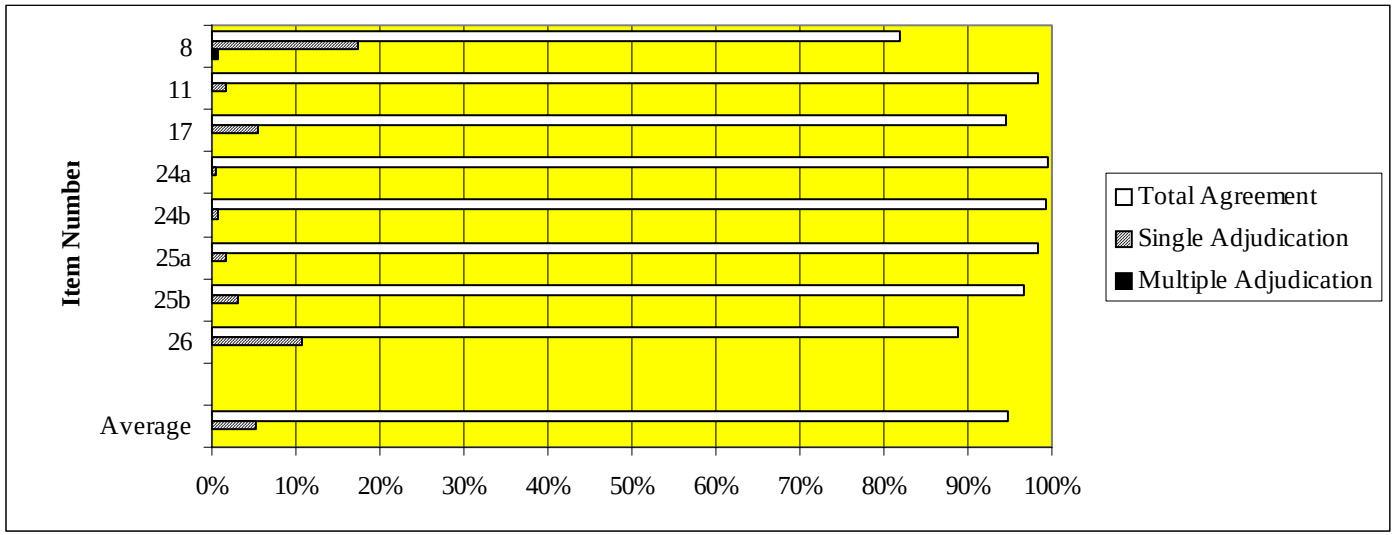

Figure 55. Interrater agreement on Grade 6 External Assessment, by item.

[13] External Assessment items are individually examined since there are few multiple-item contexts. The missing item numbers denote multiple-choice items requiring no interrater reliability analysis.

The incidence of single adjudication was inversely proportional to the incidence of total agreement (see Figure 55 and Table C1 in the Appendix). The percentage of single adjudication ranged from a low of 0.56% on Item 24a to a high of 17.39% on Item 8. The incidence of multiple adjudication was very low ranging from 0% on Items 11, 17, 24a, 24b and 25a to a high of 0.70% on Item 8.

Factors that contributed to the high interrater agreement and low adjudication include (a) high quality training for raters; (b) less complex rubrics which could not be changed; (c) effective scoring procedures; (d) the lower levels of reasoning elicitied (Item 24a), and (e) nonresponses or incorrect responses (Items 17, 24b, 25a, and 25b). Factors contributing to the lower interrater agreement (higher adjudication) on Item 8 were (a) difficulties with the open-ended format; (b) multiple scoring criteria; and (c) the higher levels of reasoning elicited.

Interrater Reliability by Districts

In District. In District 1, the interrater agreement on the Grade 6 External Assessment was very high (94.83%; see Figure 56 and Table C2 in the Appendix). Interrater agreement was over 80% on all items. Three-quarters of the items had interrater agreement over 90%. The interrater agreement ranged from a low of 81.93% on Item 8 to a high of 100% on Items 24a and 24b.

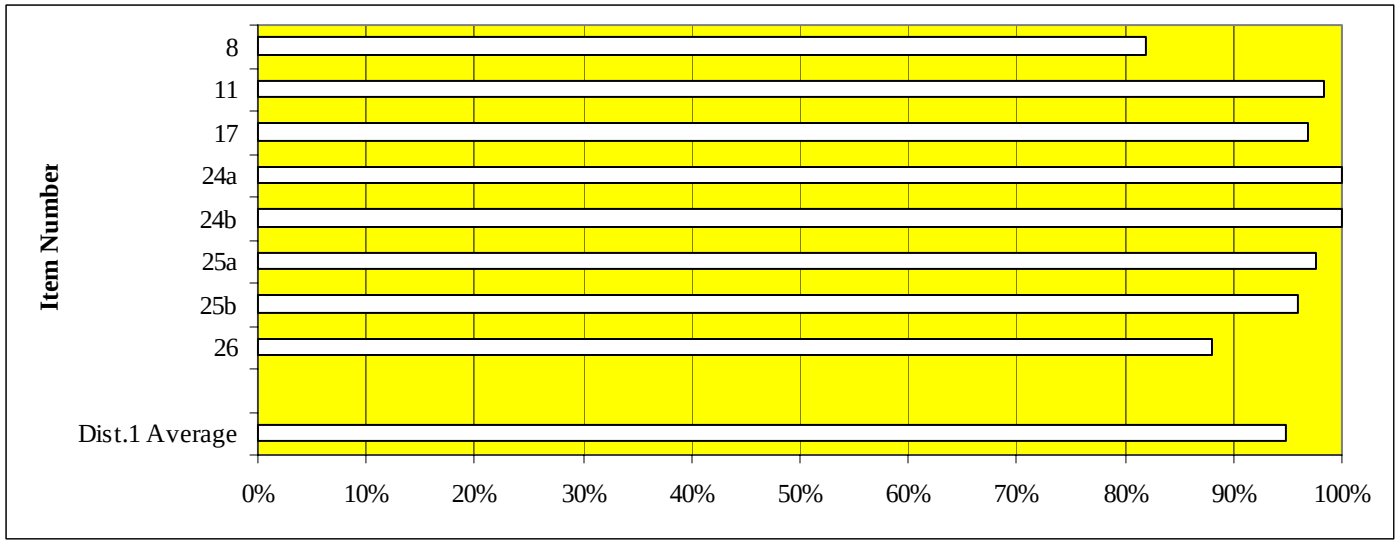

Figure 56. District 1 interrater agreement on Grade 6 External Assessment, by item.

District 2. In District 2, the interrater agreement on the Grade 6 External Assessment was very high (94.99%; see Figure 57 and Table C2 in the Appendix). Interrater agreement was over 80% on all items. Three-quarters of the items had interrater agreement over 90%. The interrater agreement ranged from a low of 82.64% on Item 8 to a high of 99.59% on Item 24a.

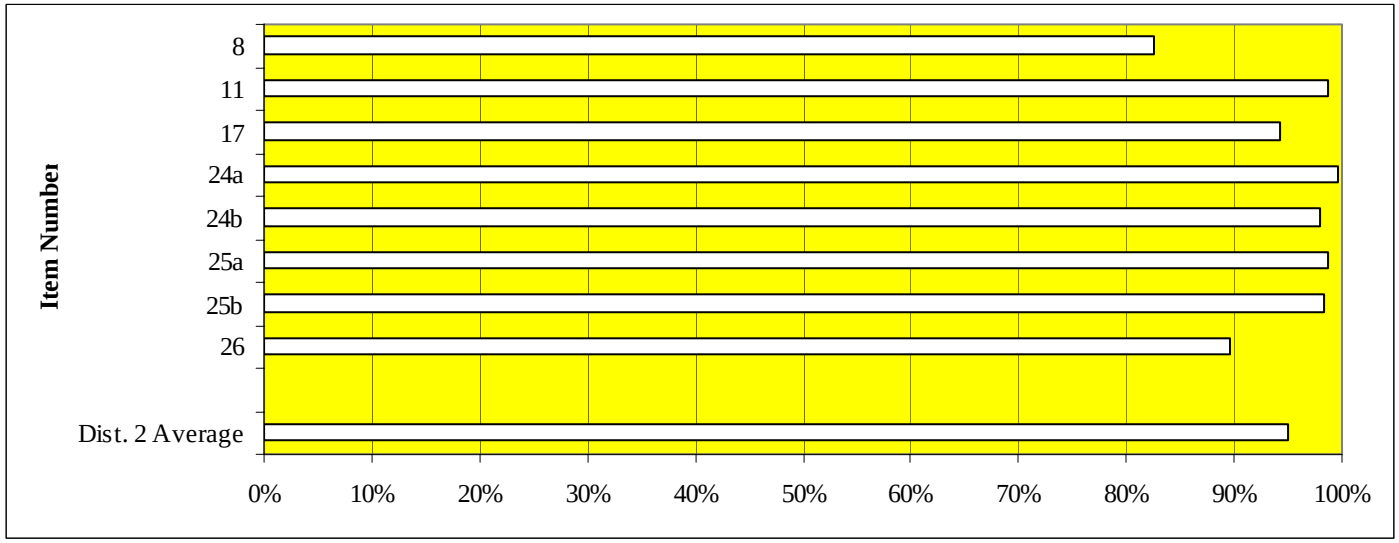

Figure 57. District 2 interrater agreement on Grade 6 External Assessment, by item.

District 3. In District 3, the interrater agreement on the Grade 6 External Assessment from District 3 was very high (93.01%; see Figure 58 and Table C2 in the Appendix). Interrater agreement was over 80% on seven out of the eight contexts. More than half of the items had interrater agreement over 90%. The interrater agreement ranged from a low of 76.27% on Item 8 to a high of 100% on Items 24a and 24b.

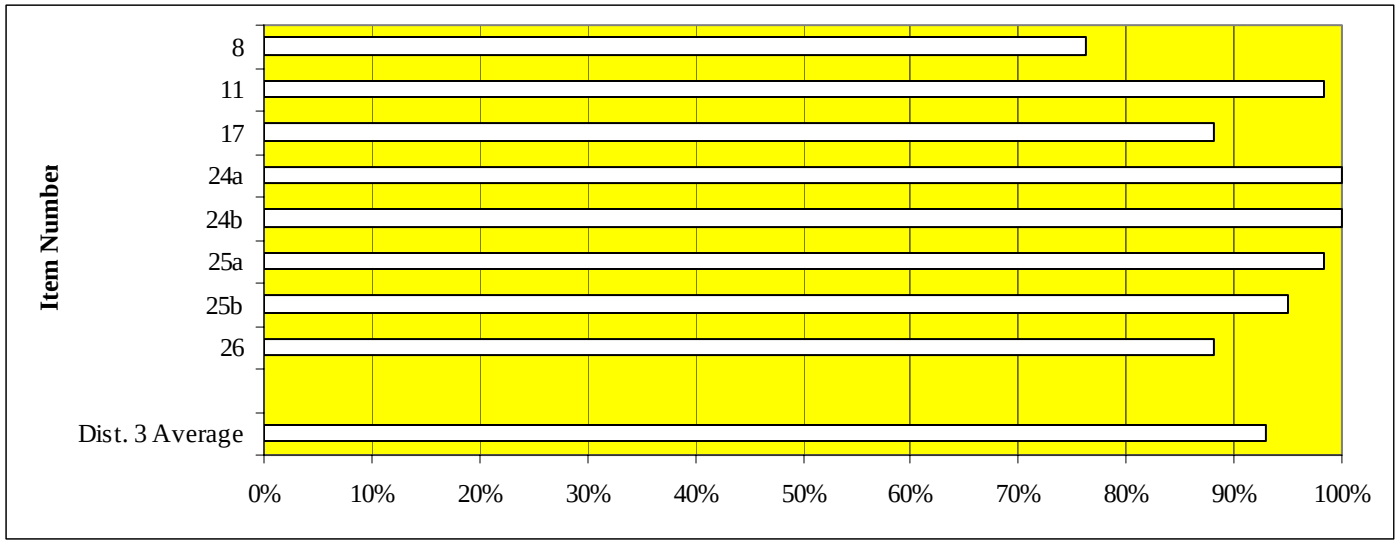

Figure 58. District 3 interrater agreement on Grade 6 External Assessment, by item.

District 4. In District 4, the interrater agreement on the Grade 6 External Assessment was very high (95.31%; see Figure 59 and Table C2 in the Appendix). Interrater agreement was over 80% on all items and over 90% on all but one item. The interrater agreement ranged from a low of 85.58% on Item 8 to a high of 100% on Item 24b.

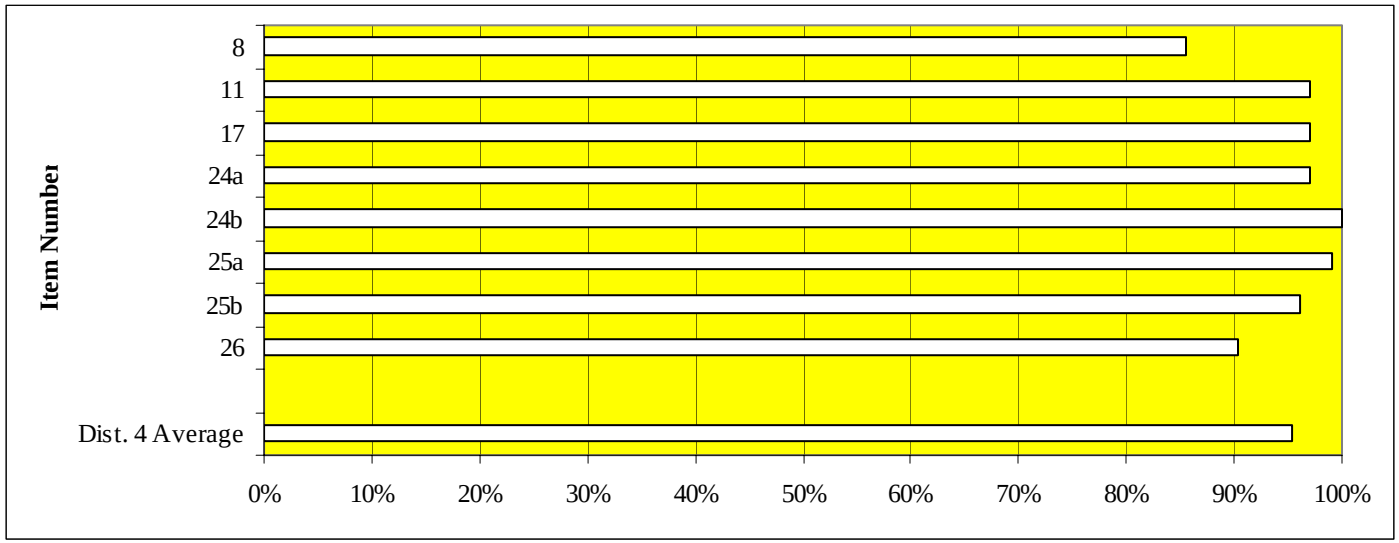

Figure 59. District 4 interrater agreement on Grade 6 External Assessment, by item.

Across districts. The interrater agreement across districts was very similar on most contexts (see Figure 60 and Table C2 in the Appendix). Some items had large (5% or greater) differences in interrater agreement. In Districts 1 and 2, there were no differences in interrater agreement greater than 5%. In District 3, interrater agreement was lower than other districts on Items 8 and 17. In District 4, interrater agreement was higher than other districts on Item 8.

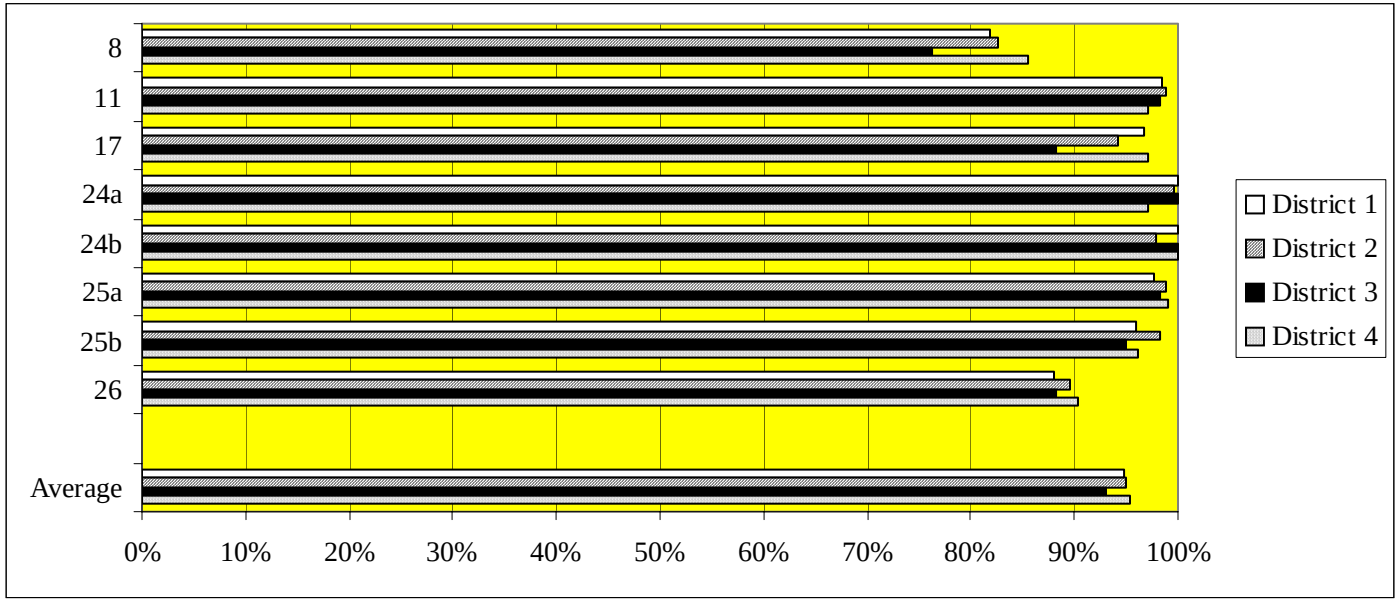

Figure 60. Across district interrater agreement on Grade 6 External Assessment, by item.

The differences in interrater agreement were most likely due to (a) content study teachers taught and (b) proportion of nonresponses and incorrect responses (e.g., in District 3, a much smaller proportion of students left answers blank than in the other districts).

Interrater Reliability by Program (Conventional Curricula or Mathematics in Context Classes)

Conventional curricula. The interrater agreement on the Grade 6 External Assessment from conventional curricula was very high (95.35%; see Figure 61 and Table C3 in the Appendix). Interrater agreement was over 80% all items. Three-quarters of the items had agreement over 90%. The interrater agreement ranged from a low of 84.88% on Item 8 to a high of 100% on Items 24a and 24b.

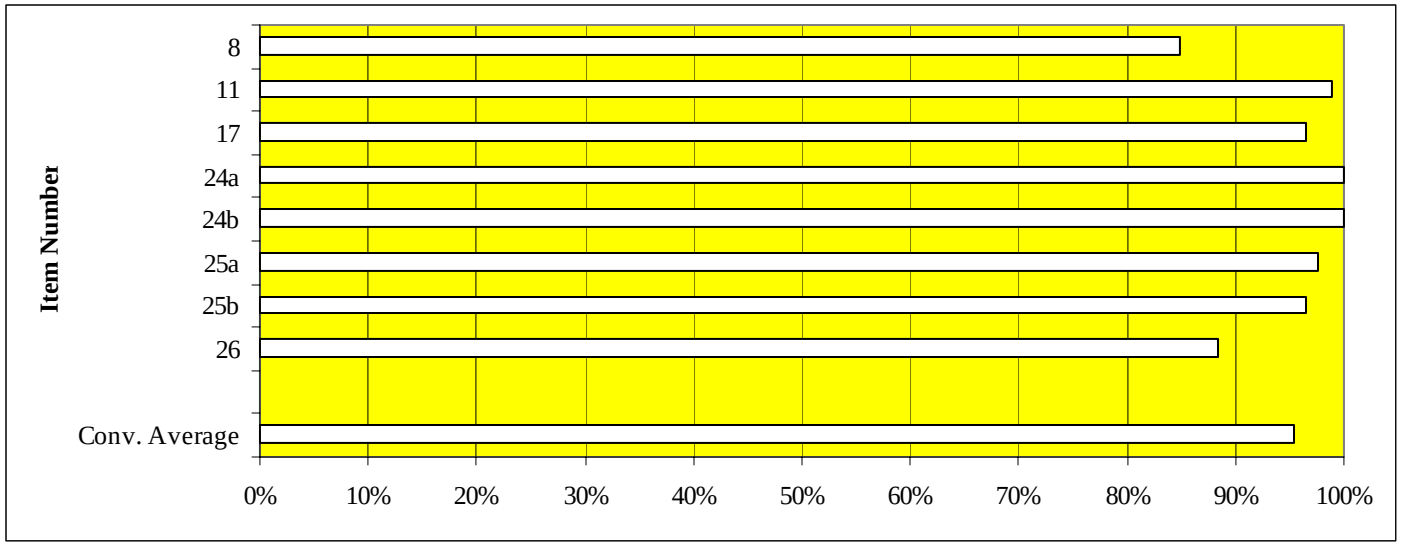

Figure 61. Interrater agreement on Grade 6 External Assessment, by item: Conventional curricula.

Mathematics in Context classes. The interrater agreement on the Grade 6 External Assessment from *Mathematics in Context* classes was very high (94.56%; see Figure 62 and Table C3 in the Appendix). Interrater agreement was over 80% on all items. Three-quarters of the items had interrater agreement over 90%. The interrater agreement ranged from a low of 81.34% on Item 8 to a high of 99.36% on Item 24a.

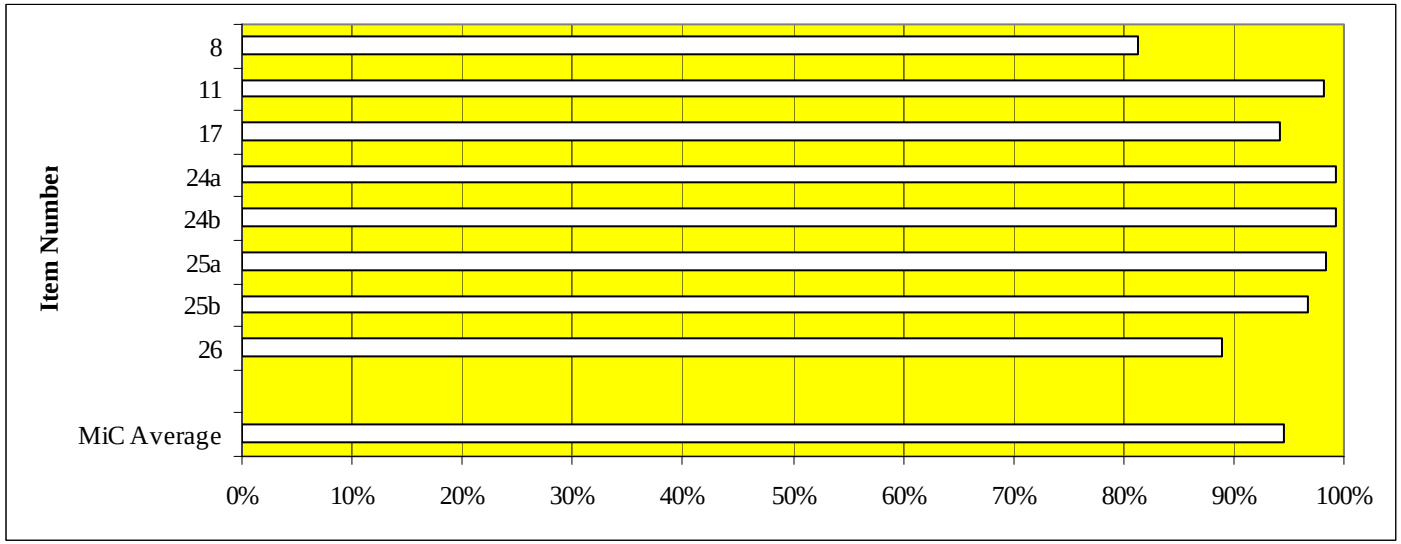

Figure 62. Interrater agreement on Grade 6 External Assessment by item *Mathematics in Context* classes.

Across program. Overall, the interrater agreement in conventional curricula and *Mathematics in Context* classes was similar (see Figure 63 and Table C3 in the Appendix). The average interrater agreement for conventional curricula was 95.35% and 94.56% for *Mathematics in Context* classes,. The difference in interrater agreement was never 5% or greater.

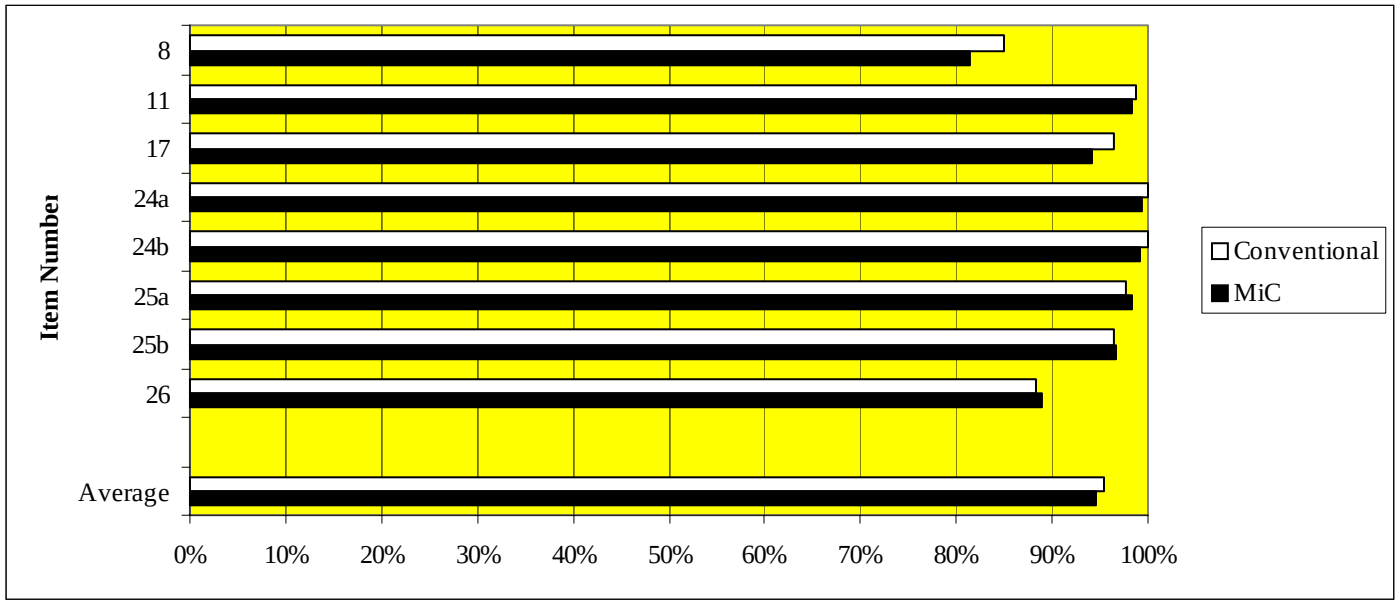

Figure 63. Interrater agreement on Grade 6 External Assessment, by item: Conventional curricula and *Mathematics in Context* classes.

Grade 7

Overall Interrater Reliability

The interrater agreement on the Grade 7 External Assessment was very high (91.75%; see Figure 64 and Appendix C4.) Interrater agreement was over 80% on seven out of the eight items.[14] All but two items had interrater agreement over 90%. The interrater agreement ranged from a low of 75.93% on Item 5 to a high of 99.26% on Item 22.

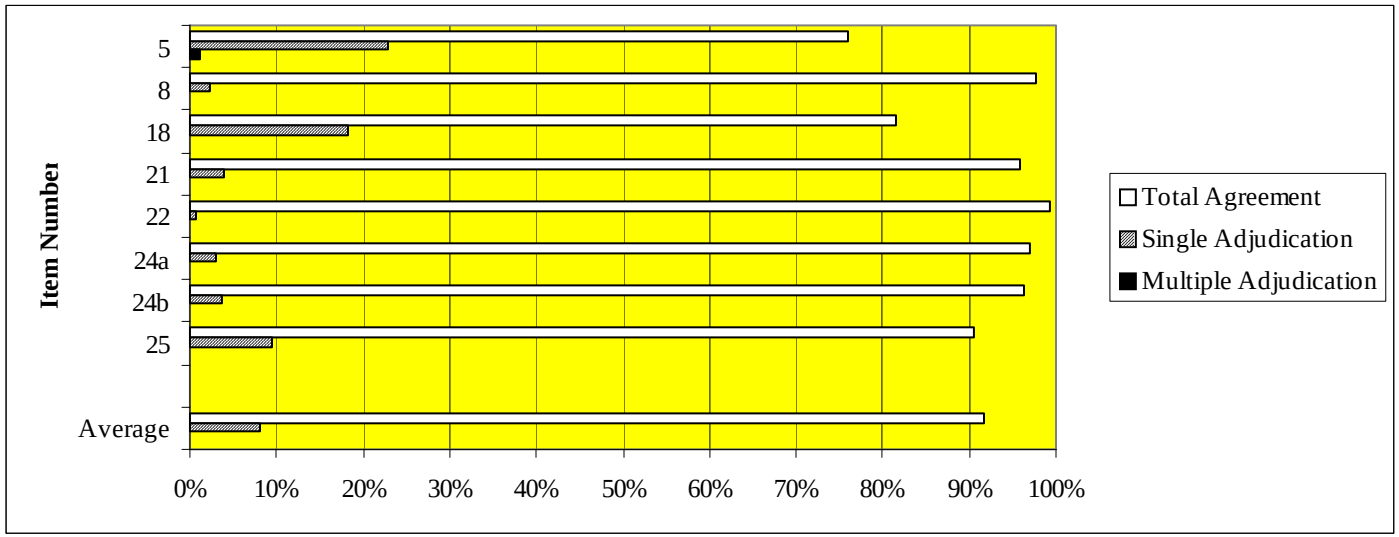

Figure 64. Interrater agreement on Grade 7 External Assessment, by item.

[14] External Assessment items are individually examined since there are few multiple-item contexts. The missing item numbers denote multiple choice items requiring no interrater reliability analysis.

The incidence of single adjudication was inversely proportional to the incidence of total agreement (see Figure 64 and Table C4 in the Appendix). The percentage of single adjudication ranged from a low of 0.74% on Item 22 to a high of 22.83% on Item 5. The incidence of multiple adjudication was very low ranging from 0% on Items 8, 22, and 24a to a high of 1.24% on Item 5.

Factors that contributed to the high interrater agreement and low adjudication include (a) high quality training for raters; (b) less complex rubrics, which could not be changed; (c) effective scoring procedures; and (d) the proportion of nonresponses or incorrect responses. Factors contributing to the lower interrater agreement (and higher adjudication) on Item 5 include (a) difficulties with the open-ended format and (b) multiple scoring criteria.

Interrater Reliability by Districts

District 1. In District 1, the interrater agreement on the Grade 7 External Assessment was very high (91.33%; see Figure 65 and Table C5 in the Appendix). Interrater agreement was over 80% on seven out of the eight items. Three-quarters of the items had interrater agreement over 90%. The interrater agreement ranged from a low of 71.77% on Item 5 to a high of 99.19% on the Item 22.

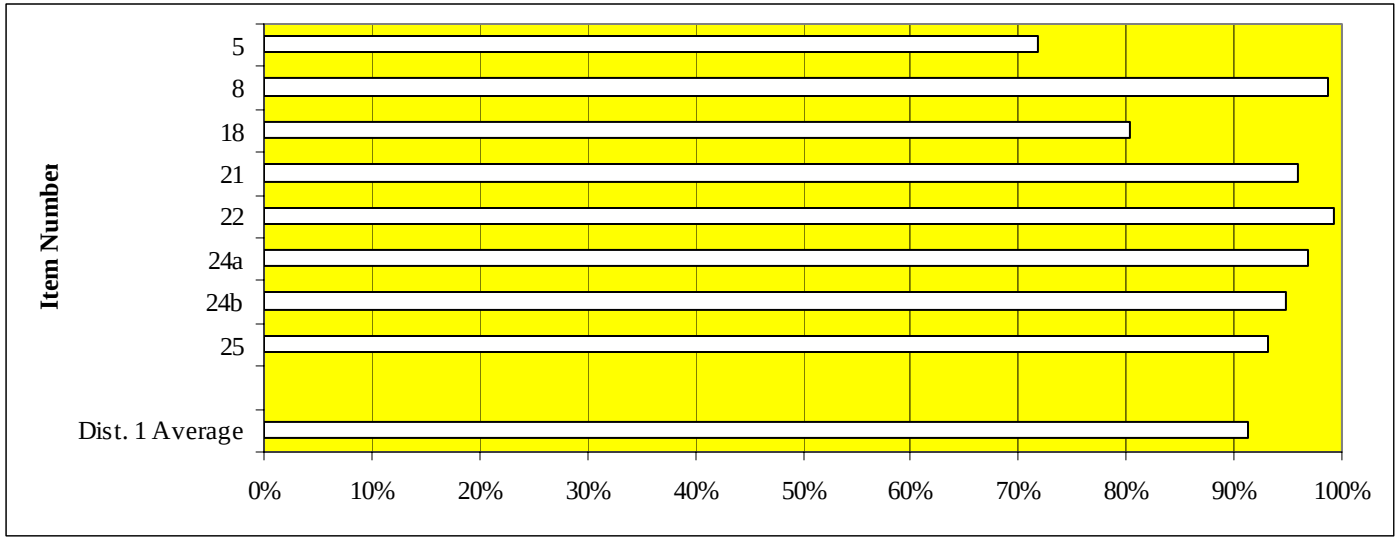

Figure 65. District 1 interrater agreement on Grade 7 External Assessment, by item.

District 2. In District 2, the interrater agreement on the Grade 7 External Assessment was very high (92.57%; see Figure 66 and Table C5 in the Appendix). Interrater agreement was over 80% on seven out of the eight items. Three-quarters of the items had interrater agreement over 90%. The interrater agreement ranged from a low of 79.06% on Item 5 to a high of 99.15% on Item 22.

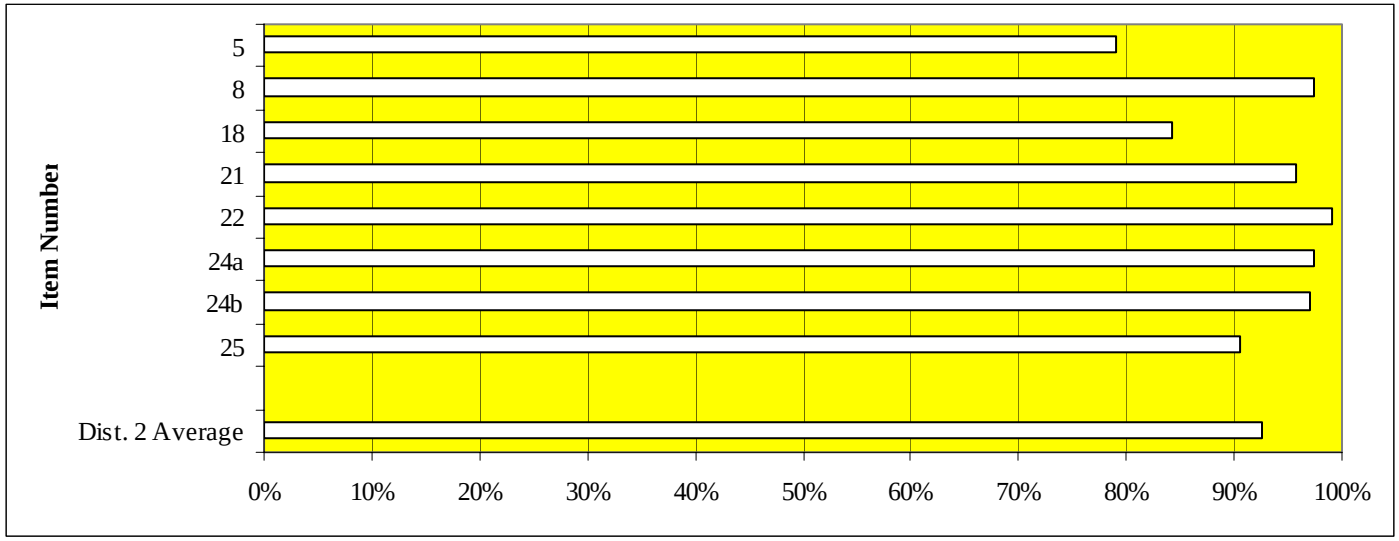

Figure 66. District 2 interrater agreement on Grade 7 External Assessment, by item.

District 3. In District 3, the interrater agreement on the Grade 7 External Assessment from District 3 was high (89.96%; see Figure 67 and Table C5 in the Appendix). Interrater agreement was over 80% on three-quarters of the items. Interrater agreement was over 90% on five out of the eight items. The interrater agreement ranged from a low of 71.53% on Item 18 to a high of 98.54% on Items 8 and 22. The other item with low interrater agreement was Item 5 at 75.00%.

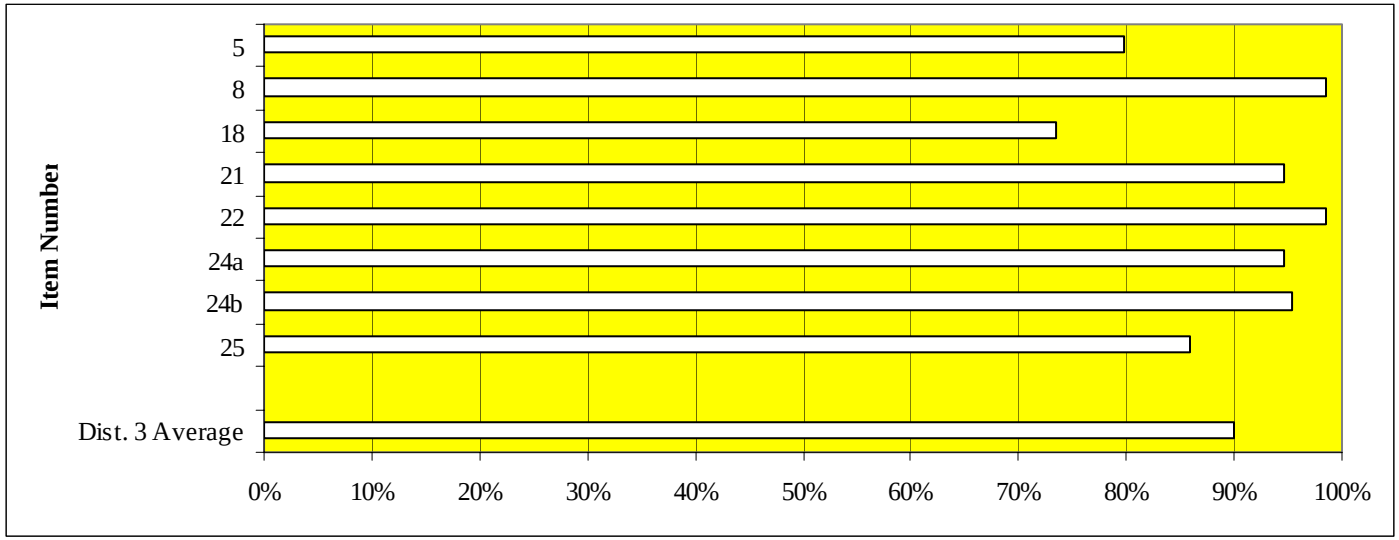

Figure 67. District 3 interrater agreement on Grade 7 External Assessment, by item.

District 4. In District 4, the interrater agreement on the Grade 7 External Assessment from District 4 was very high (92.41%; see Figure 68 and Table C5 in the Appendix). Interrater agreement was over 80% all but one item. Interrater agreement was over 90% on five out of the eight items. The interrater agreement ranged from a low of 75.00% on Item 5 to a high of 100% on Item 22.

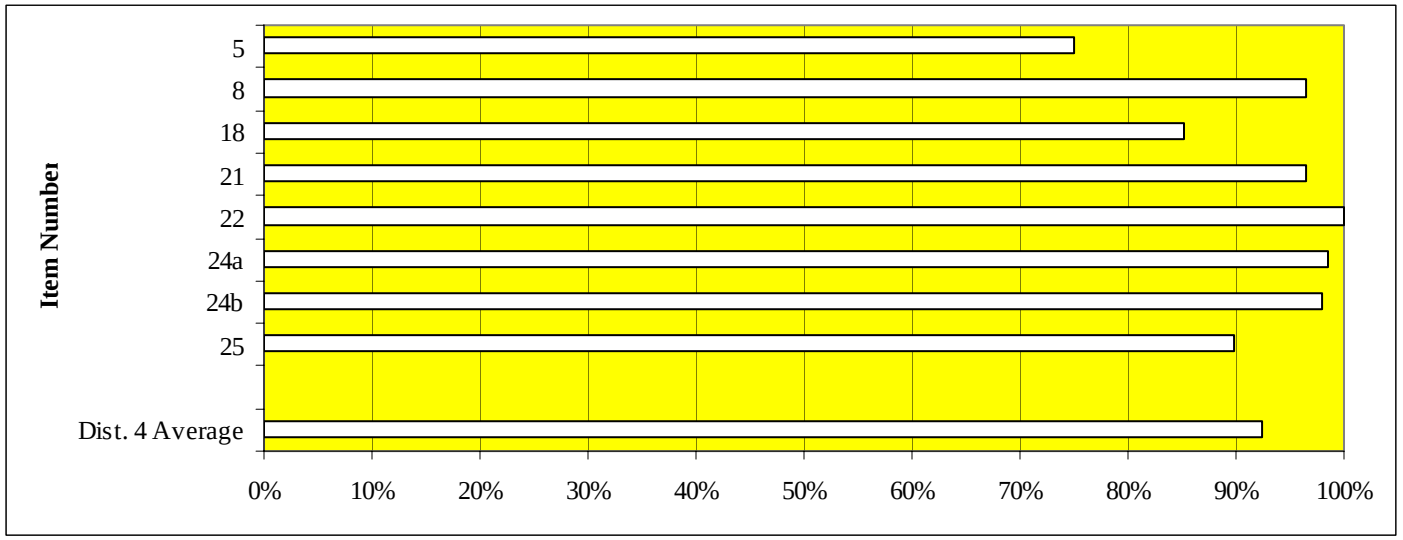

Figure 68. District 4 interrater agreement on Grade 7 External Assessment, by item.

Across districts. The interrater agreement across districts was very close on most items (see Figure 69 and Table C5 in the Appendix). Some items from each district had large (5% or greater) differences in interrater agreement. In District 1, interrater agreement was lower than the other districts on Item 5 and higher than other districts on Item 25. In District 2, interrater agreement was never larger that 5% greater difference than in the other districts. In District 3, interrater agreement was much lower than other districts on Items 18 and 25. In District 4, interrater agreement was higher than in the other districts on Item 18.

The differences in interrater agreement were most likely due to (a) content study teachers taught and (b) proportion of nonresponse and incorrect responses.

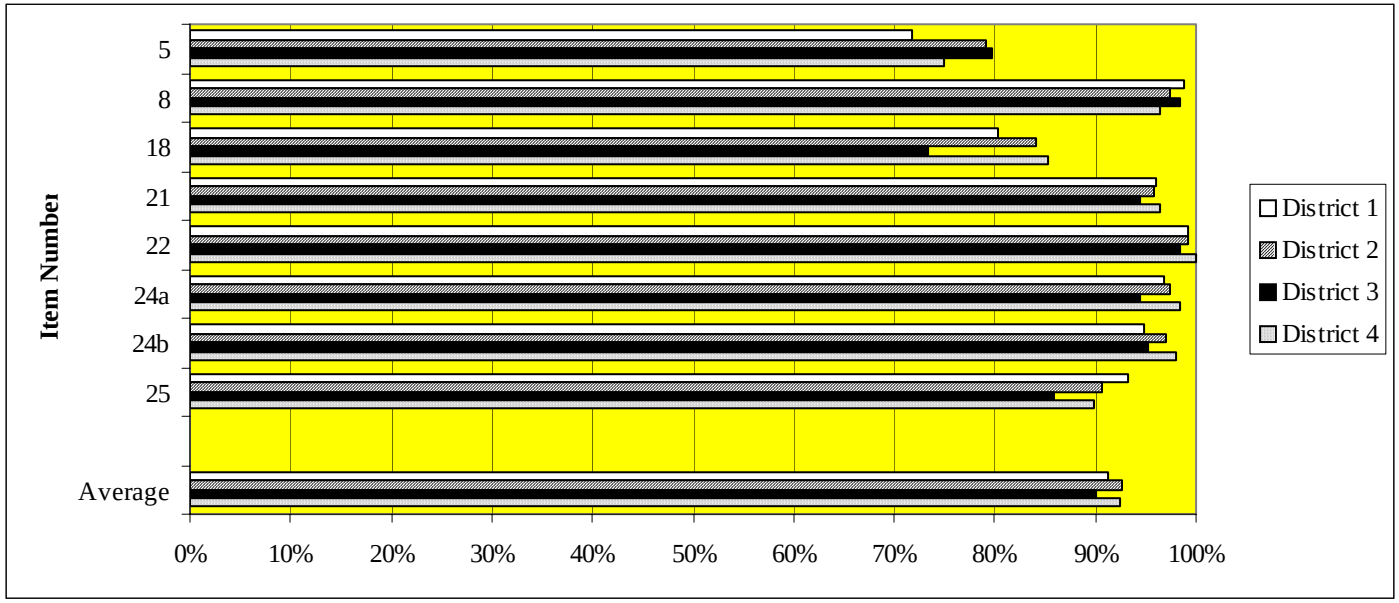

Figure 69. Across district interrater agreement on Grade 7 External Assessment, by item.

Interrater Reliability by Curricula (Conventional Curricula or Mathematics in Context Classes)

Conventional curricula. The interrater agreement on the Grade 7 External Assessment from conventional classes was very high (92.70%; see Figure 70 and Table C6 in the Appendix). Interrater agreement was over 80% on all but one item and over 90% on three-quarters of the contexts. The interrater agreement ranged from a low of 76.11% on Item 5 to a high of 98.23% on Item 22.

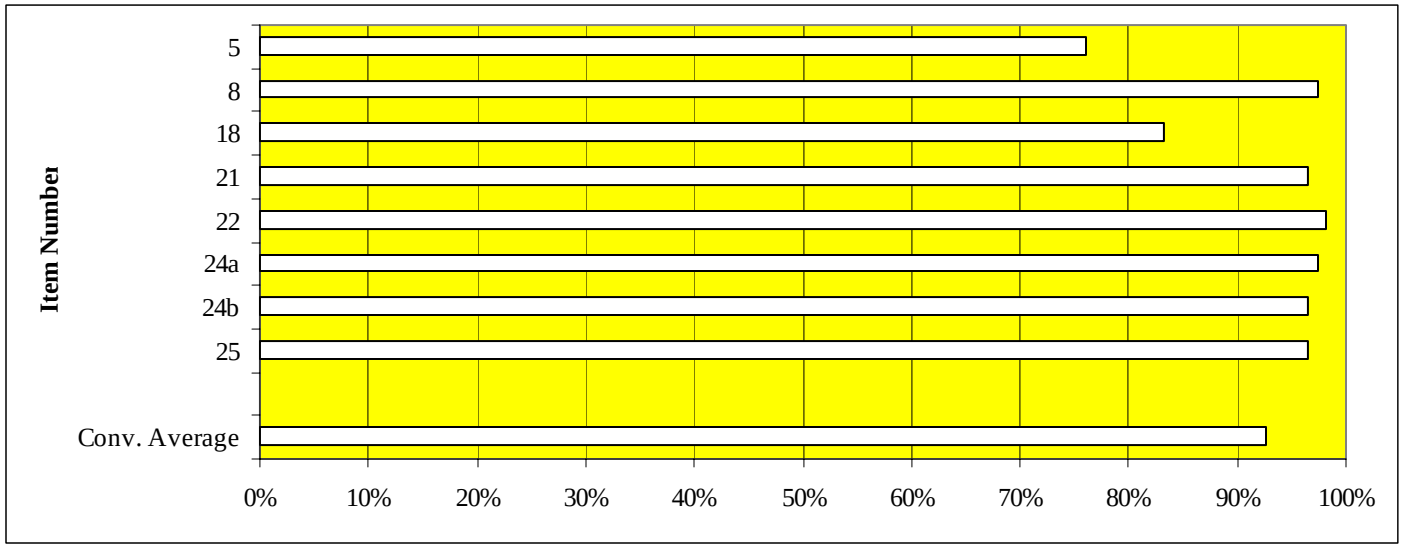

Figure 70. Interrater agreement on Grade 7 External Assessment, by item: Conventional curricula.

Mathematics in Context classes. The interrater agreement on the Grade 7 External Assessment from *Mathematics in Context* classes was very high (91.59%; see Figure 71 and Table C6 in the Appendix). Interrater agreement was over 80% on seven out of the eight items. Interrater agreement was over 90% on five out of eight items. The interrater agreement ranged from a low of 75.90% on Item 5 to a high of 99.42% on Item 22.

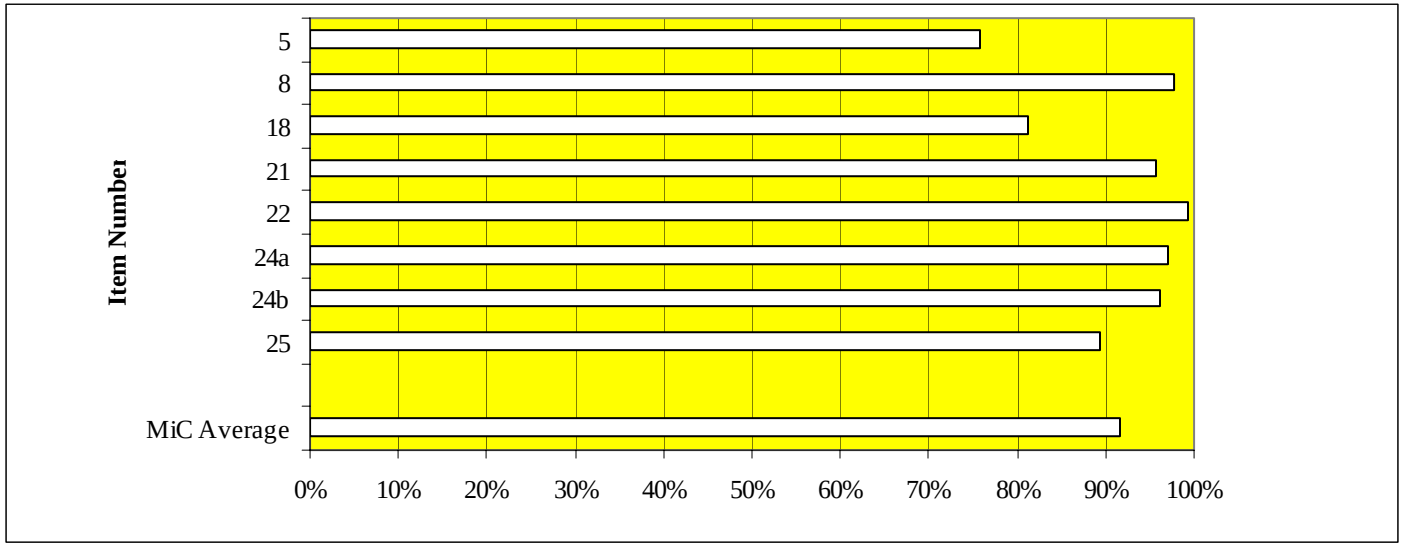

Figure 71. Interrater agreement on Grade 7 External Assessment, by item: *Mathematics in Context* classes.

Across program. Overall, the interrater agreement in conventional curricula and *Mathematics in Context* classes was similar (see Figure 72 and Table C6 in the Appendix). The average interrater agreement for conventional curricula was 92.70% and 91.59% for *Mathematics in Context* classes. The interrater agreement was higher (5% or greater) on assessments from the conventional curricula for Item 25.

This difference was most likely due to the content that the study teachers taught.

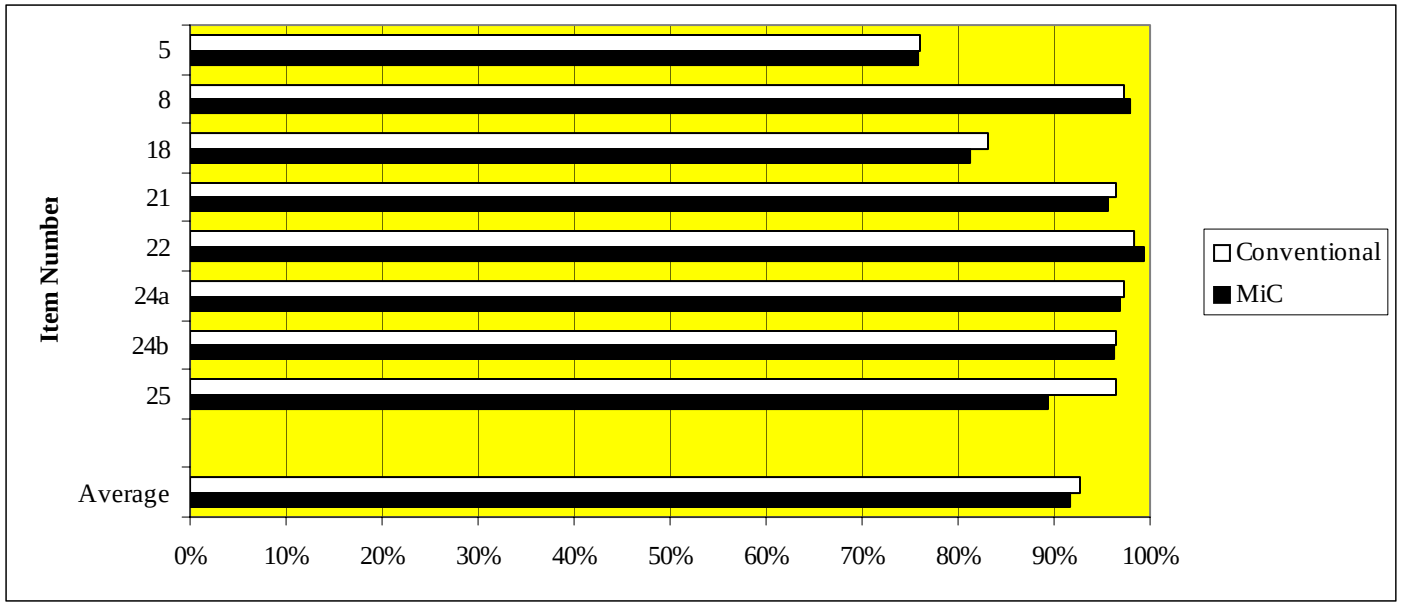

Figure 72. Interrater agreement on Grade 7 External Assessment, by item: Conventional curricula and *Mathematics in Context* classes.

Grade 8

Overall Interrater Reliability

The interrater agreement on the Grade 8 External Assessment was very high (92.96%; see Figure 73 and Appendix C7). Interrater agreement was over 80% on nine out of the ten items.[15] Interrater agreement was over 90% on four-fifths of the items. The interrater agreement ranged from a low of 76.91% on Item 1 to a high of 99.10% on Item 22b.

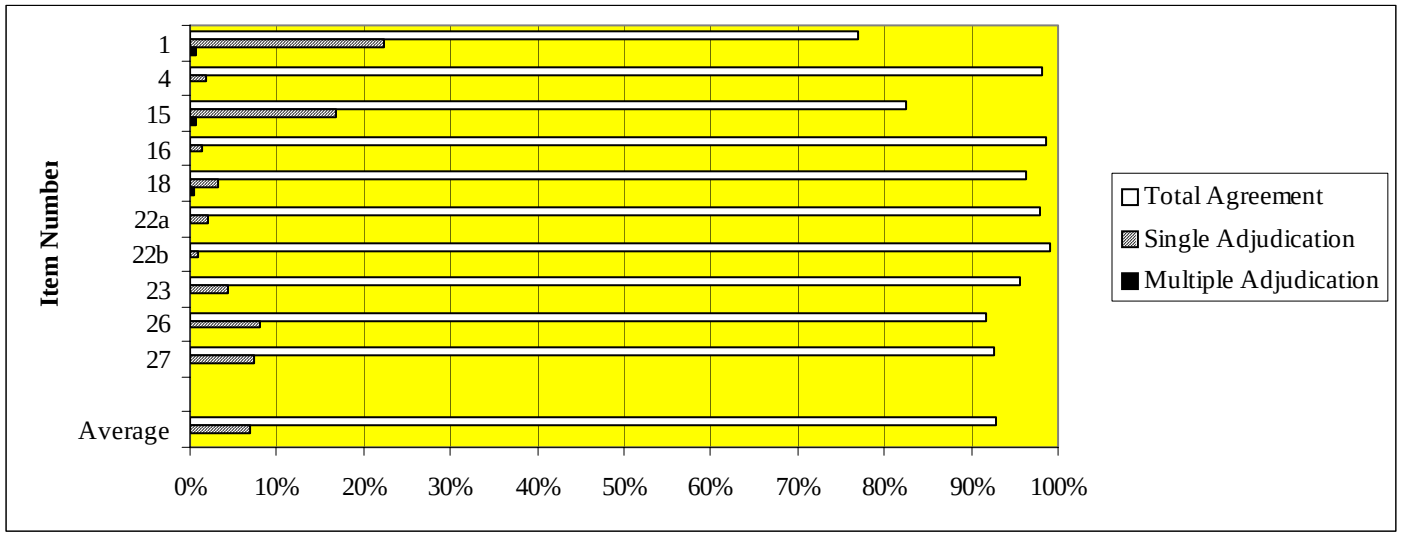

Figure 73. Interrater agreement on Grade 8 External Assessment, by item.

[15] External Assessment items are individually examined since there are few multiple-item contexts. The missing item numbers denote multiple choice items requiring no interrater reliability analysis.

The incidence of single adjudication was inversely proportional to the incidence of total agreement (see Figure 73 and Table C7 in the Appendix). The percentage of single adjudication ranged from a low of 0.90% on Item 22b to a high of 22.42% on Item 1. The incidence of multiple adjudication was very low ranging from 0% on Items 4, 16, 22a, 22b, 23, and 27 to a high of 0.67% on Items 1 and 15.

Factors that contributed to the high interrater agreement and low adjudication include (a) high quality training for raters; (b) less complex rubrics, which could not be changed; (c) effective scoring procedures; and (d) the proportion of nonresponses or incorrect responses. Factors contributing to the lower interrater agreement (and higher adjudication) on Item 1 include (a) difficulties with the open-ended format and (b) multiple scoring criteria.

Interrater Reliability by Districts

District 1. In District 1, the interrater agreement on the Grade 8 External Assessment was very high (93.84%; see Figure 74 and Table C8 in the Appendix). Interrater agreement was over 80% on nine out of the ten items. Four-fifths of the items had interrater agreement over 90%. The interrater agreement ranged from a low of 78.62% on Item 15 to a high of 99.37% on the Item 16.

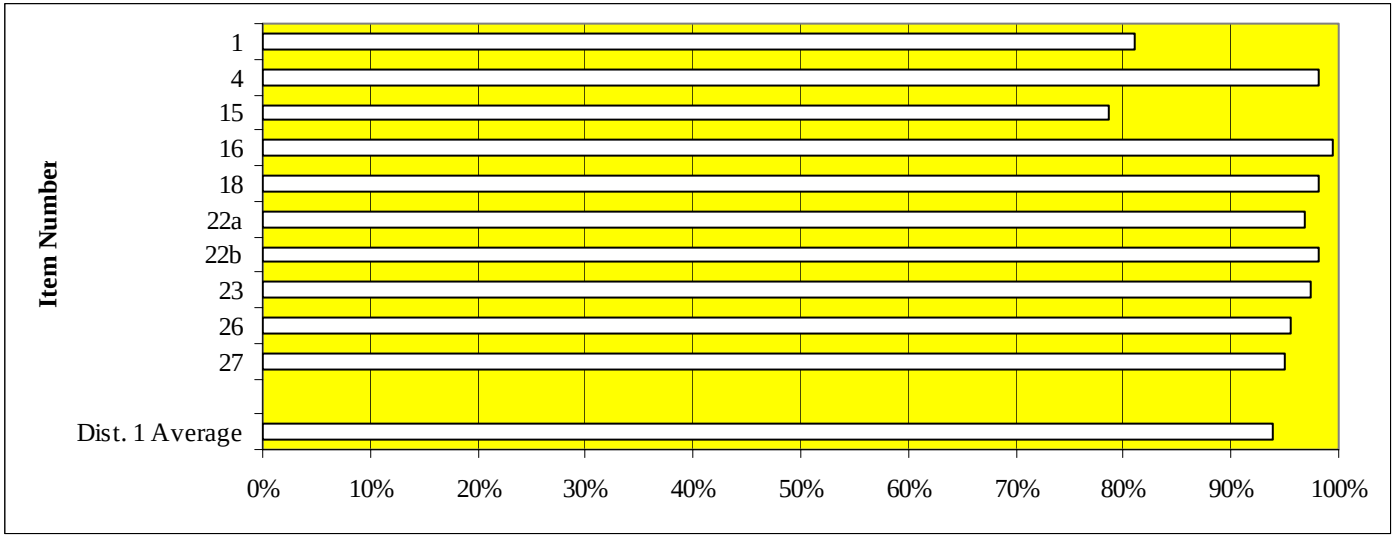

Figure 74. District 1 interrater agreement on Grade 8 External Assessment, by item.

District 2. In District 2, the interrater agreement on the Grade 8 External Assessment was very high (91.68%; see Figure 75 and Table C8 in the Appendix). Interrater agreement was over 80% on nine out of the ten items. Almost three-quarters of the items had interrater agreement over 90%. The interrater agreement ranged from a low of 73.28% on Item 1 to a high of 99.24% on Items 22a and 22b.

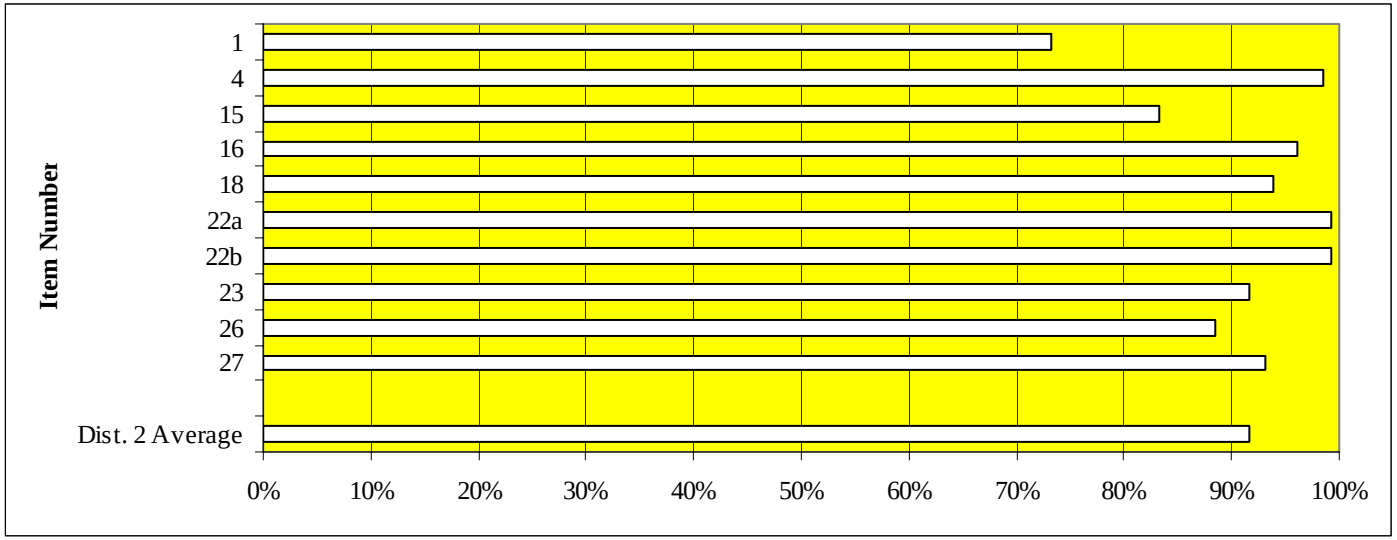

Figure 75. District 2 interrater agreement on Grade 8 External Assessment, by item.

District 3. In District 3, the interrater agreement on the Grade 8 External Assessment from District 3 was high (87.69%; see Figure 76 and Table C8 in the Appendix). Interrater agreement was over 80% on three-fifths of the items. Three-fifths of the items had interrater agreement over 90%. The interrater agreement ranged from a low of 70.77% on Item 15 to a high of 100% on Items 16 and 22b. The other items with low interrater agreement were Item 1 at 72.31%, Item 27 at 76.92%, and Item 26 at 78.46%.

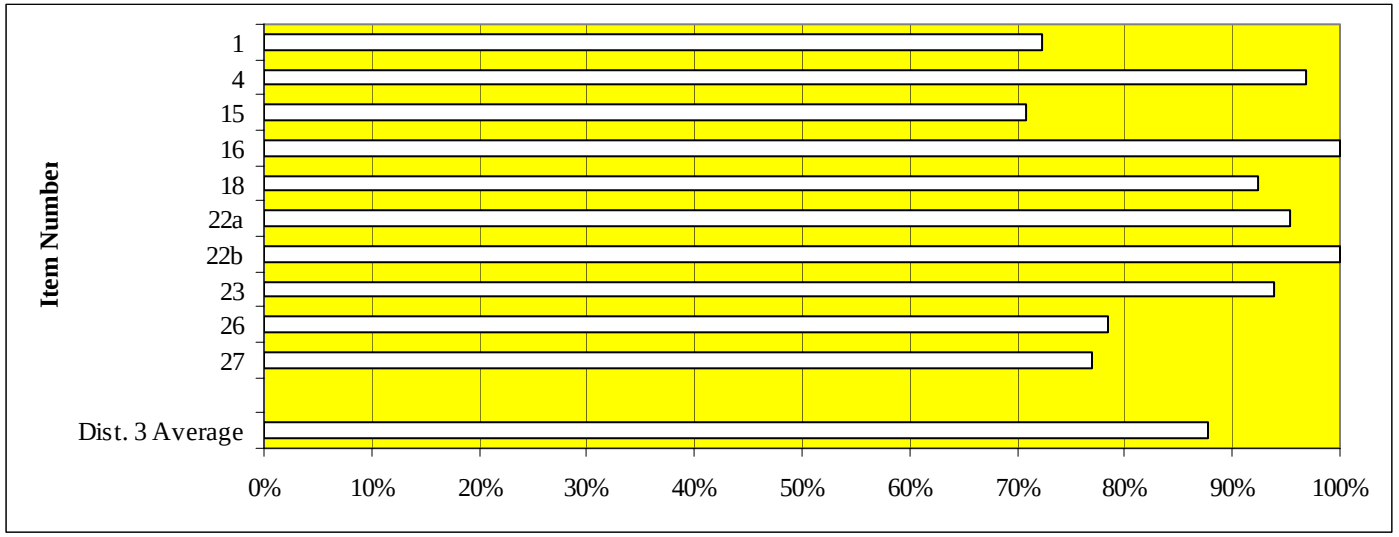

Figure 76. District 3 interrater agreement on Grade 8 External Assessment, by item.

District 4. In District 4, the interrater agreement on the Grade 8 External Assessment from District 4 was very high (97.03%; see Figure 77 and Table C8 in the Appendix). Interrater agreement was over 80% on nine out of the ten items. Interrater agreement was over 90% on nine out of the ten items. The interrater agreement ranged from a low of 78.02% on Item 1 to a high of 100% on Items 16, 18, 22a, and 22b.

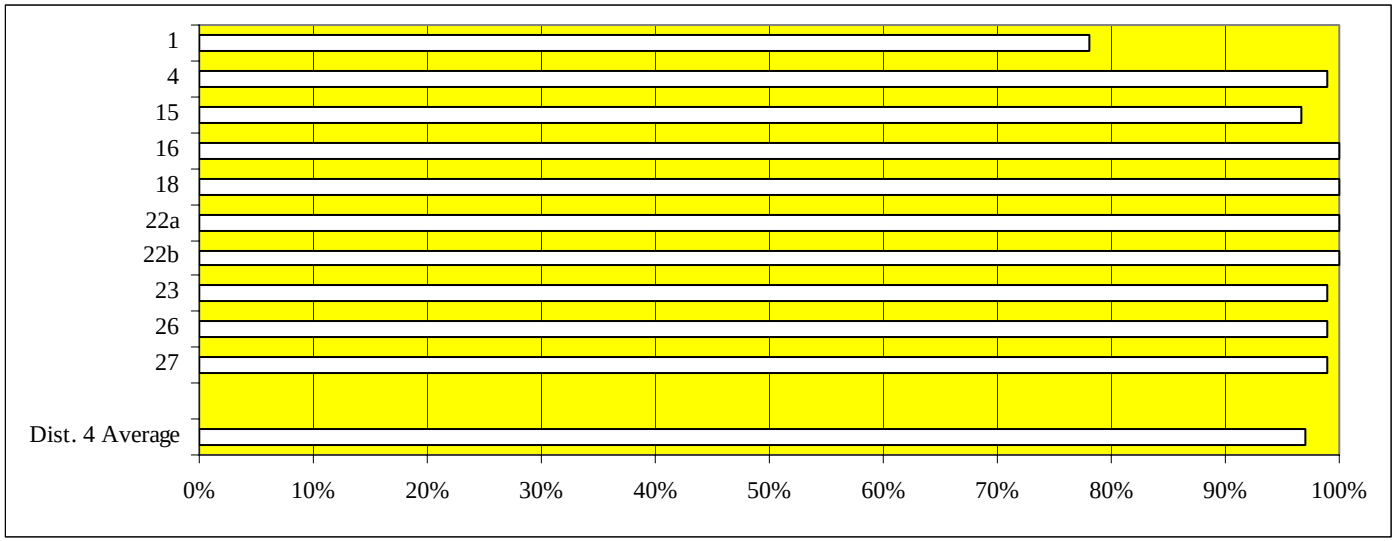

Figure 77. District 4 interrater agreement on Grade 8 External Assessment, by item.

Across districts. The interrater agreement across districts was very close on most items (see Figure 78 and Table C8 in the Appendix). Some items from each district had large (5% or greater) differences in interrater agreement. In District 1, interrater agreement was high on Items 1, 23, and 26. In District 2, interrater agreement was low on Items 18, 23, and 26. In District 3, interrater agreement was much lower on Items 15, 26, and 27; and low on Items 1, 18, 22a, and 23. In District 4, interrater agreement was much higher than in the other districts on Item 15 and high on Items 1, 18, 23, 26, and 27.

The large differences in interrater agreement were most likely due to (a) content study teachers taught and (b) proportion of nonresponse and incorrect responses.

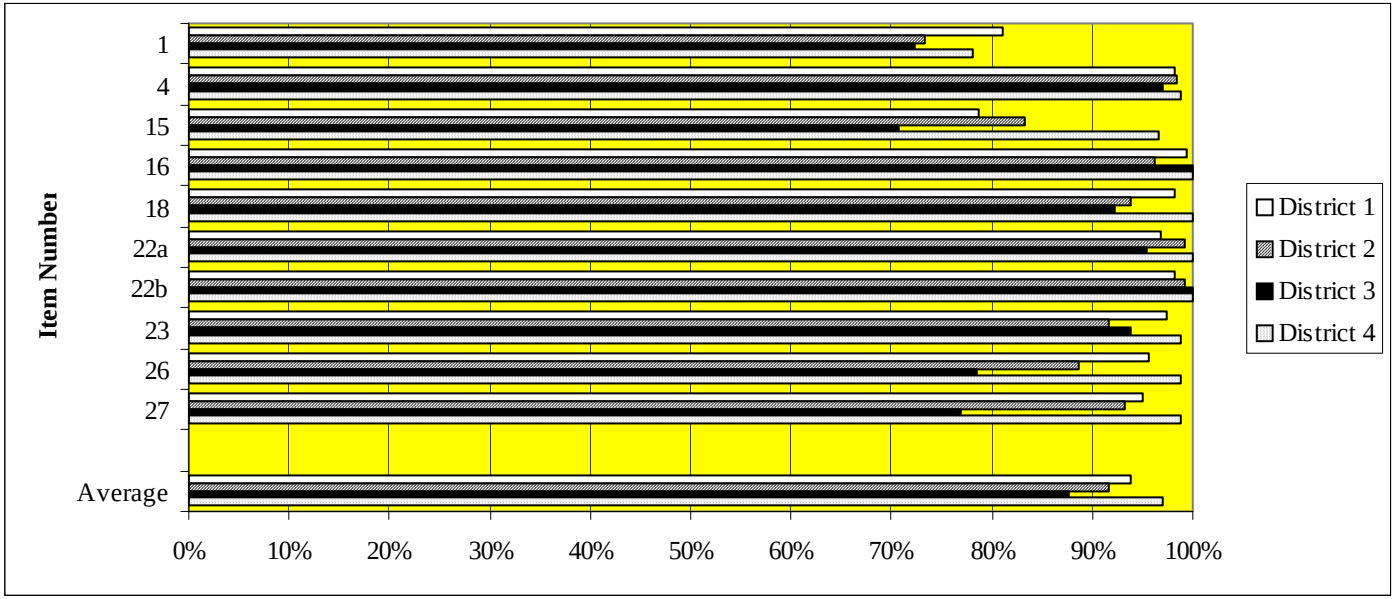

Figure 78. Across district interrater agreement on Grade 8 External Assessment, by item.

Interrater Reliability by Curricula (Conventional Curricula or Mathematics in Context Classes)

Conventional curricula. The interrater agreement on the Grade 8 External Assessment from conventional classes was very high (95.92%; see Figure 79 and Table C9 in the Appendix). Interrater agreement was over 80% on all items. Interrater agreement was over 90% on four-fifths of the items. The interrater agreement ranged from a low of 84.47% on Item 1 to a high of 99.03% on Items 4, 16, 22b, 23, and 26.

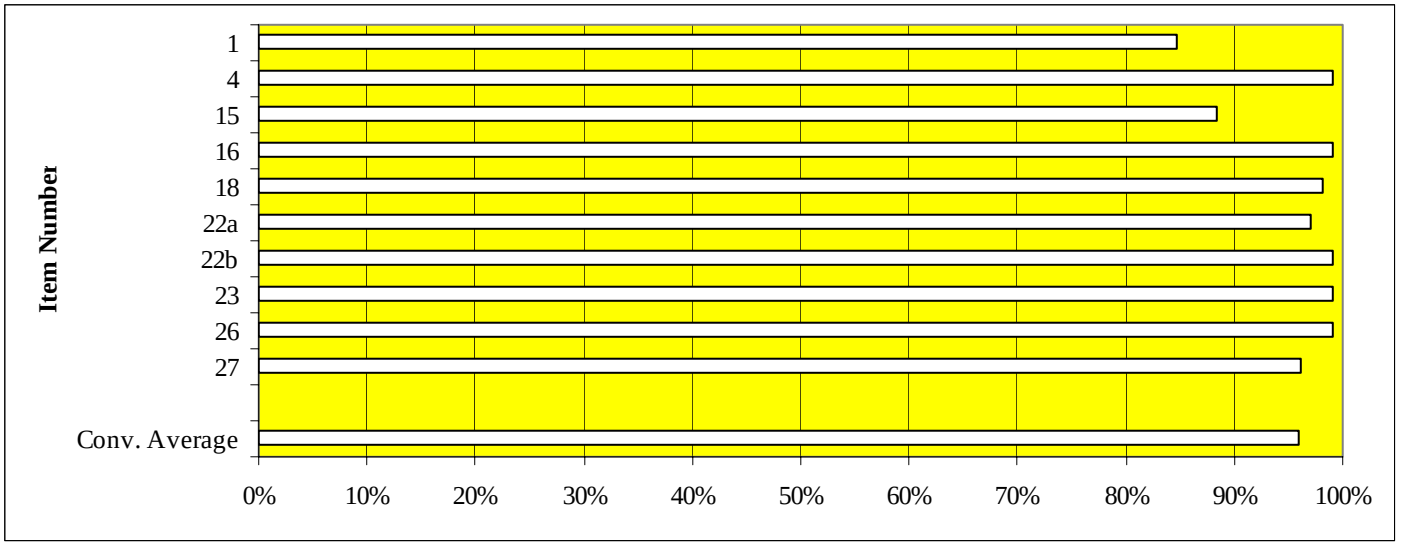

Figure 79. Interrater agreement on Grade 8 External Assessment, by item: Conventional curricula.

Mathematics in Context classes. The interrater agreement on the Grade 8 External Assessment from *Mathematics in Context* classes was very high (92.07%; see Figure 80 and Table C9 in the Appendix). Interrater agreement was over 80% on nine out of the ten items. More than two-thirds of the items had interrater agreement over 90%. The interrater agreement ranged from a low of 74.64% on Item 1 to a high of 99.13% on Item 22b.

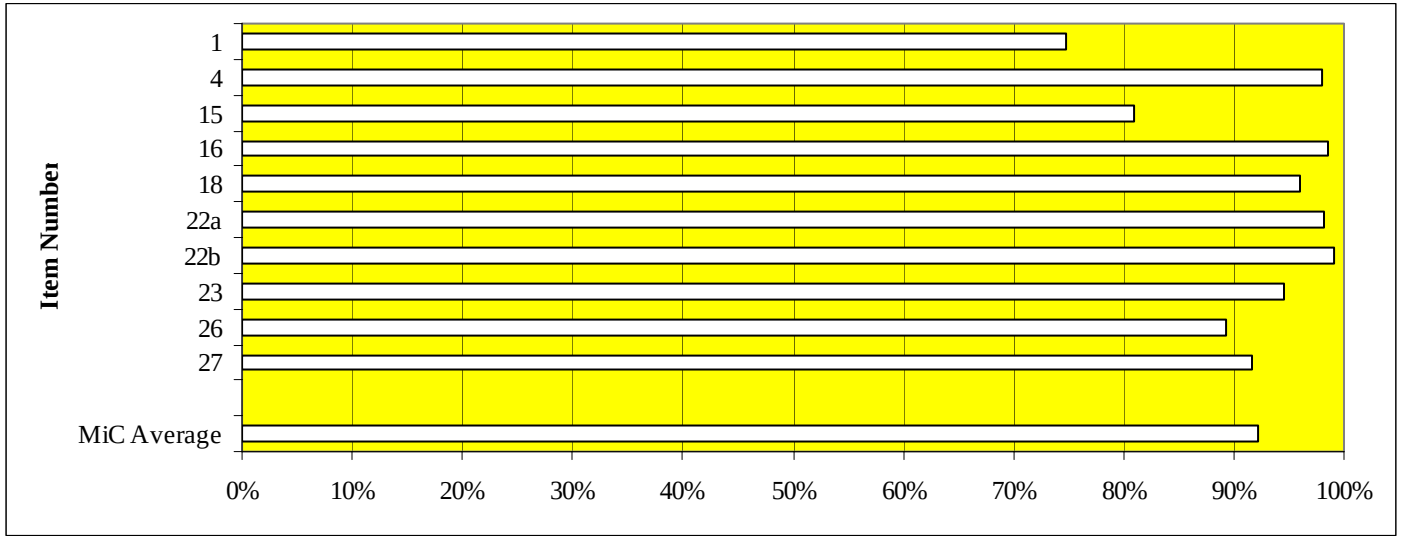

Figure 80. Interrater agreement on Grade 8 External Assessment, by item: *Mathematics in Context* classes.

Across program. Overall, the interrater agreement in conventional curricula and *Mathematics in Context* classes was similar (see Figure 81 and Table C9 in the Appendix). The average interrater agreement for conventional curricula was 95.92% and 92.07% for *Mathematics in Context* classes. Some items from each curricula had large (5% or greater) differences in interrater agreement. Interrater agreement was higher on assessments from the conventional curricula classrooms on Items 1, 15, and 26.

This difference was most likely due to the content that the study teachers taught.

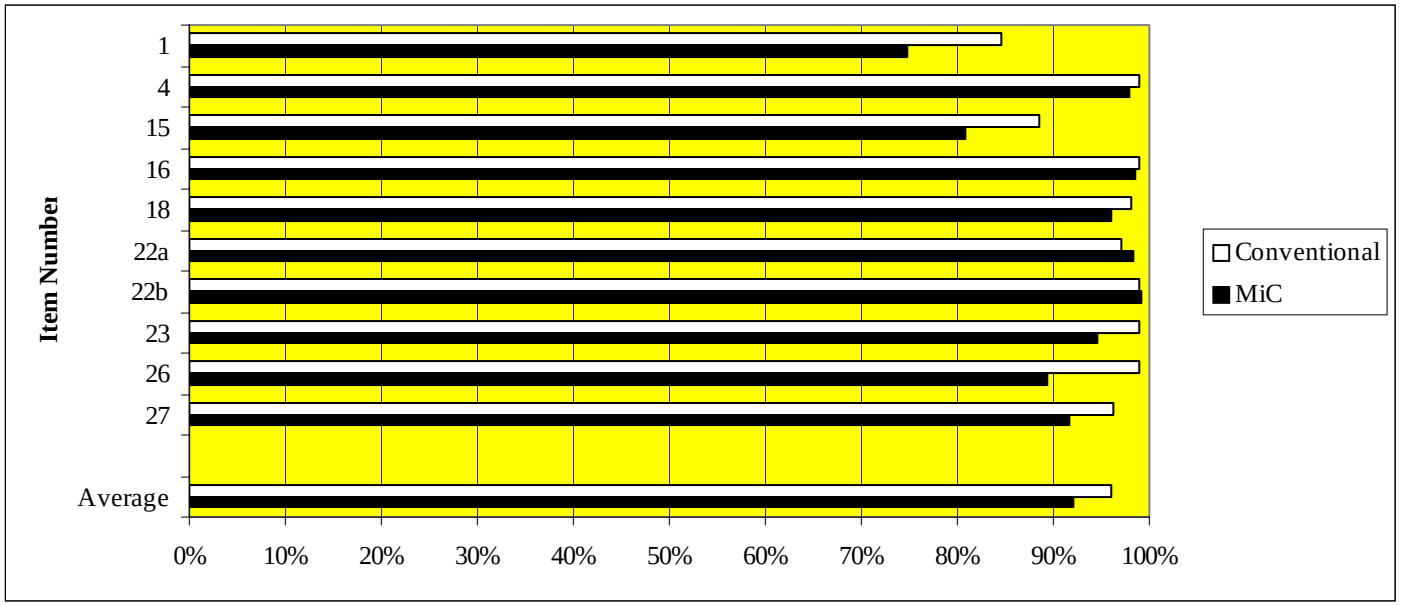

Figure 81. Interrater agreement on Grade 8 External Assessment, by item: Conventional curricula and *Mathematics in Context* classes.

Conclusion

By design, many of the items on the External Assessment were used at more than one grade level (see Figure 82 and Table C10 in the Appendix). The first context on each assessment tended to have lower interrater agreement: (Item 8 on EA6, Item 5 on EA7, and Item 1 on EA8). Also, one anchor item (Item 26 on EA6, Item 18 on EA7, and Item 26 on EA8) and the two more difficult contexts only on the Grade 8 assessment (Items 15 and 27 on EA8) had lower interrater agreement.

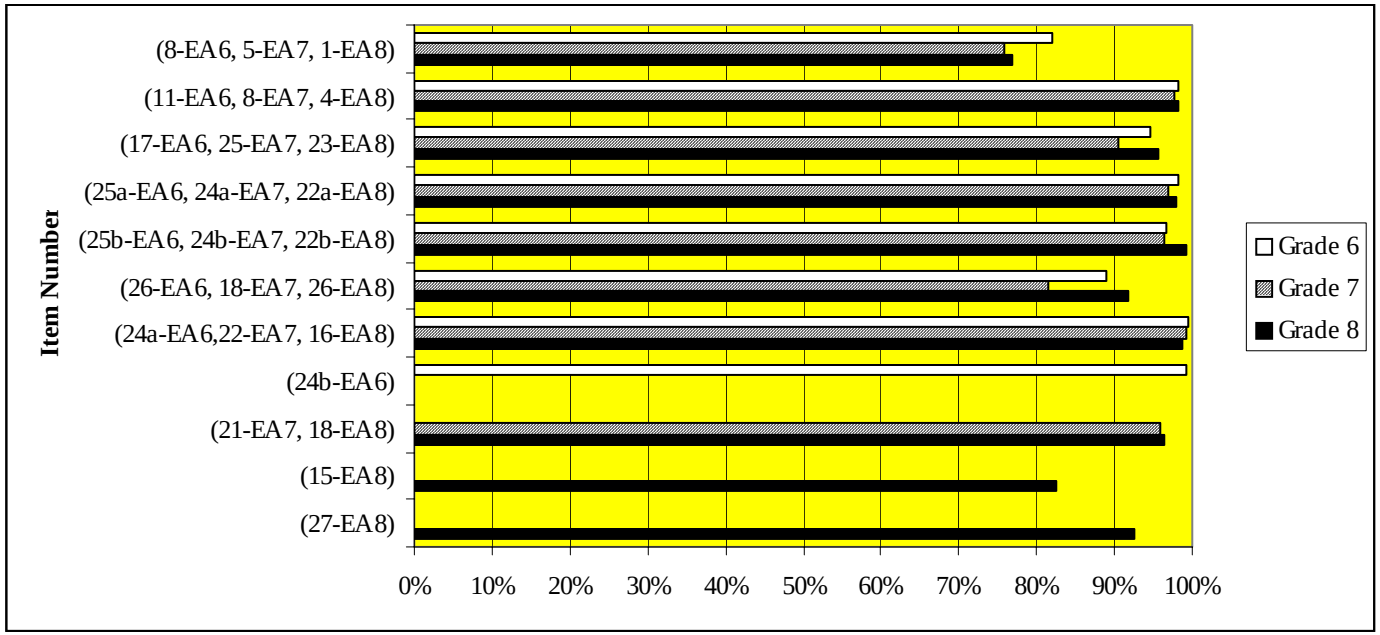

Figure 82. Interrater agreement of items from the Grades 6, 7, and 8 1998–1999 External Assessment.

Two factors led to higher interrater agreement. First, items eliciting lower level responses were less complex to score (Item 24a on the EA6, Item 22 on the EA7, and Item 16 on the EA8). Second, grade levels of each context were scored in succession. Interrater agreement tended to improve with each grade level on a particular context because of the cumulative experience and confidence of the raters. Factors leading to lower interrater agreement most likely were: (a) difficulties with the open-ended format; (b) multiple scoring criteria (Items 26 on EA6, Item 18 on EA7 and Items 15, 26 and 27 on EA8); and (c) the higher levels of reasoning elicited (Items 26 on EA6, Item 18 on EA7 and Items 15, 26 and 27 on EA8).

Conclusion

The interrater reliability was high on the External Assessments. The factors that contributed to the very high interrater agreement are (a) high quality training for raters; (b) rater experience; (c) effective scoring procedures; (d) less complex rubrics which could not be changed; and (e) proportion of nonresponses and incorrect responses. The factors that account for the lower interrater agreement were (a) difficulties with the open-ended format; (b) multiple scoring criteria; and (c) the higher levels of reasoning elicited.

The differences in interrater agreement among districts were most likely due to (a) content study teachers taught; (b) time of day items were scored; and (c) proportion of nonresponse and incorrect responses.

Differences in interrater agreement between conventional curricula and *Mathematics in Context* classes were most likely due to the content study teachers taught.

Appendix A

Interrater Reliability by Scoring Institute and by Rater

Table A1

Interrater Reliability by Scoring Institute and by Rater

Institute (Contexts Scored)*	Location (Date)	Assessments Rated *(N)*	Contexts Rated *(N)*	Items in Contexts *(N)*	Rater	Agreement *(N)*	Agreement *%*	Single Adjudication *(N)*	Single Adjudication *%*	Multiple Adjudication *(N)*	Multiple Adjudication *%*	Student Responses Rated *(N)*
1	District 3	PSA-6 (117),	5	2592	A	365	79.00%	90	19.48%	7	1.52%	462
PSA-6	(5/13/99-	PSA-7 (17), &			B	215	82.69%	44	16.92%	1	0.38%	260
RS #1-6,	5/14/99)	PSA-8 (20)			C	332	74.44%	105	23.54%	9	2.02%	446
PT #7-9,					D	317	78.47%	83	20.54%	4	0.99%	404
S #16,					E	168	78.50%	44	20.56%	2	0.93%	214
PSA-7					F	403	84.66%	72	15.13%	1	0.21%	476
PY #15-21,					G	269	81.52%	60	18.18%	1	0.30%	330
PSA-8				Total:	7	2069		498		25		2592
PK #15-17				**Average:**			**79.82%**		**19.21%**		**0.96%**	
2	District 1	PSA-6 (205)	1	2505	H	96	84.21%	16	14.04%	2	1.75%	114
PSA-6	(5/13/99-				I	132	88.00%	17	11.33%	1	0.67%	150
RS #1-6	(5/14/99)				J	158	81.03%	35	17.95%	2	1.03%	195
					K	254	89.12%	30	10.53%	1	0.35%	285
					L	295	91.05%	29	8.95%	0	0.00%	324
					M	169	89.42%	20	10.58%	0	0.00%	189
					N	306	89.47%	34	9.94%	2	0.58%	342
					O	111	80.43%	22	15.94%	5	3.62%	138
					P	265	87.46%	36	11.88%	2	0.66%	303
					Q	417	89.68%	45	9.68%	3	0.65%	465
				Total:	10	2203		284		18		2505
				Average:			**87.94%**		**11.34%**		**0.72%**	
3	District 2	PSA-6 (130)	4	2316	R	185	86.85%	25	11.74%	3	1.41%	213
PSA-6	(5/13/99-				S	113	89.68%	13	10.32%	0	0.00%	126
RS #1-6,	(5/14/99)				T	202	90.99%	20	9.01%	0	0.00%	222
PT #7-9,					U	141	85.45%	21	12.73%	3	1.82%	165
BW #14-15,					V	171	90.48%	15	7.94%	3	1.59%	189
S #16					W	204	85.00%	33	13.75%	3	1.25%	240
					X	186	89.86%	20	9.66%	1	0.48%	207
					Y	149	80.11%	36	19.35%	1	0.54%	186
					Z	142	78.89%	32	17.78%	6	3.33%	180
					AA	134	91.16%	11	7.48%	2	1.36%	147
					AB	114	86.36%	17	12.88%	1	0.76%	132
					AC	201	90.54%	21	9.46%	0	0.00%	222
					AD	38	74.51%	8	15.69%	5	9.80%	51
					AE	32	88.89%	4	11.11%	0	0.00%	36
				Total:	14	2012		276		28		2316
				Average:			**86.87%**		**11.92%**		**1.21%**	

Interrater Agreement, By Scoring Institute and Rater

Table A1 (*continued*)

Institute (Contexts Scored)	Location (Date)	Assessments Rated *(N)*	Contexts Rated *(N)*	Items in Contexts *(N)*	Rater	Agreement *(N)*	Agreement *%*	Single Adjudication *(N)*	Single Adjudication *%*	Multiple Adjudication *(N)*	Multiple Adjudication *%*	Student Responses Rated *(N)*
4	District 4	PSA-6 (55),	8	14	AF	70	86.42%	11	13.58%	0	0.00%	81
PSA-6	(5/25/99)	PSA-7 (37), &			AG	100	87.72%	14	12.28%	0	0.00%	114
PT #7-9,		PSA-8 (39)			AH	333	81.42%	74	18.09%	2	0.49%	409
S #16,					AI	183	82.81%	38	17.19%	0	0.00%	221
PSA-7				Total:	4	686		137		2		825
PY #15-21,				**Average:**			**83.15%**		**16.61%**		**0.24%**	
PSA-8												
PK #15-17												
5	Madison 1	PSA-6 (688)	6	25552	AJ	1981	92.44%	154	7.19%	8	0.37%	2143
PSA-6	(6/14/99-				AK	2106	89.73%	234	9.97%	7	0.30%	2347
RS #1-6,	6/18/99)				AL	1809	93.63%	121	6.26%	2	0.10%	1932
PT #7-9,					AM	2110	92.58%	162	7.11%	7	0.31%	2279
F #10-13,					AN	1859	90.64%	184	8.97%	8	0.39%	2051
BW #14-15,					AO	2131	90.84%	208	8.87%	7	0.30%	2346
S #16,					AP	2048	92.80%	154	6.98%	5	0.23%	2207
BS #17-24					AQ	1350	92.78%	100	6.87%	5	0.34%	1455
					AR	1854	92.75%	142	7.10%	3	0.15%	1999
					AS	1838	93.16%	132	6.69%	3	0.15%	1973
					AT	2442	90.58%	245	9.09%	9	0.33%	2696
					AU	1880	88.51%	236	11.11%	8	0.38%	2124
				Total:	12	23408		2072		72		25552
				Average:			**91.61%**		**8.11%**		**0.28%**	
6	Madison 2	PSA-7 (825)	4	19	AV	1829	90.28%	178	8.79%	19	0.94%	2026
PSA-7	(7/12/99-				AW	1516	89.33%	161	9.49%	20	1.18%	1697
BF #1-7,	7/16/99)				AX	1416	89.68%	150	9.50%	13	0.82%	1579
PN #8-10,					AY	1369	86.98%	189	12.01%	16	1.02%	1574
A #11-14,					AZ	2332	90.88%	221	8.61%	13	0.51%	2566
PYG #22-26					BA	2551	90.43%	263	9.32%	7	0.25%	2821
					BB	2420	92.09%	190	7.23%	18	0.68%	2628
					BC	726	86.74%	103	12.31%	8	0.96%	837
					BD	1587	92.05%	133	7.71%	4	0.23%	1724
					BE	1690	90.37%	163	8.72%	17	0.91%	1870
					BF	2636	90.62%	249	8.56%	24	0.83%	2909
					BG	2447	90.16%	253	9.32%	14	0.52%	2714
					BH	1529	88.90%	178	10.35%	13	0.76%	1720
					BI	2308	89.18%	263	10.16%	17	0.66%	2588
					BJ	1317	89.78%	135	9.20%	15	1.02%	1467
					BK	1189	88.14%	143	10.60%	17	1.26%	1349
					BL	3106	90.21%	313	9.09%	24	0.70%	3443
				Total:	17	31968		3285		259		35512
				Average:			**90.02%**		**9.25%**		**0.73%**	

Interrater Agreement, By Scoring Institute and Rater

Table A1 (*continued)*

Institute (Contexts Scored)	Location (Date)	Assessments Rated *(N)*	Contexts Rated *(N)*	Items in Contexts *(N)*	Rater	Agreement *(N)*	Agreement *%*	Single Adjudication *(N)*	Single Adjudication *%*	Multiple Adjudication *(N)*	Multiple Adjudication *%*	Student Responses Rated *(N)*
7	Madison 3	PSA-7 (825),	9	32	BM	1634	91.59%	146	8.18%	4	0.22%	1784
PSA-7	(7/26/99-	PSA-8 (503),			BN	1966	91.96%	171	8.00%	1	0.05%	2138
PY #15-21,	7/30/99)	EA-6 (713), &			BO	1380	93.31%	96	6.49%	3	0.20%	1479
PSA--8		EA-7 (810)			BP	1637	90.49%	165	9.12%	7	0.39%	1809
CM #1,					BQ	2036	92.42%	162	7.35%	5	0.23%	2203
LP #2-4,					BR	2526	94.11%	155	5.77%	3	0.11%	2684
KC #5-7,					BS	2164	95.12%	107	4.70%	4	0.18%	2275
SS #8-9,					BT	1504	93.77%	97	6.05%	3	0.19%	1604
ST #10-14,					BU	1787	92.98%	130	6.76%	5	0.26%	1922
PK #15-17,					BV	2354	93.30%	169	6.70%	0	0.00%	2523
CU #18-21,					BW	2410	92.87%	178	6.86%	7	0.27%	2595
EA-7					BX	1526	87.65%	210	12.06%	5	0.29%	1741
#24a+b					BY	1275	91.73%	108	7.77%	7	0.50%	1390
					BZ	2233	92.20%	183	7.56%	6	0.25%	2422
					CA	277	88.22%	37	11.78%	0	0.00%	314
					CB	1935	93.07%	139	6.69%	5	0.24%	2079
					CC	2421	92.79%	186	7.13%	2	0.08%	2609
					CD	1373	95.95%	54	3.77%	4	0.28%	1431
				Total:	18	32438		2493		71		35002
				Average:			**92.67%**		**7.12%**		**0.20%**	

Interrater Agreement, By Scoring Institute and Rater

Table A1 (*continued)*

Institute (Contexts Scored)	Location (Date)	Assessments Rated *(N)*	Contexts Rated *(N)*	Items in Contexts *(N)*	Rater	Agreement *(N)*	Agreement *%*	Single Adjudication *(N)*	Single Adjudication *%*	Multiple Adjudication *(N)*	Multiple Adjudication *%*	Student Responses Rated *(N)*
8	Madison 4	EA-6 (713),	8	14	CE	1655	90.83%	163	8.95%	4	0.22%	1822
EA-6 #8,	(8/2/99-	EA-7 (810), &			CF	1526	92.71%	115	6.99%	5	0.30%	1646
#11,	8/6/99)	EA-8 (446)			CG	1400	91.44%	126	8.23%	5	0.33%	1531
#17,					CH	2336	93.29%	162	6.47%	6	0.24%	2504
#24a+b,					CI	2110	92.87%	161	7.09%	1	0.04%	2272
#25a+b,					CJ	1723	93.54%	114	6.19%	5	0.27%	1842
#26,					CK	1136	93.81%	74	6.11%	1	0.08%	1211
EA-7 #5,					CL	1443	92.26%	119	7.61%	2	0.13%	1564
#8,					CM	1603	92.93%	119	6.90%	3	0.17%	1725
#18,					CN	1638	94.74%	88	5.09%	3	0.17%	1729
#21,					CO	1248	90.30%	130	9.41%	4	0.29%	1382
#22,					CP	1277	91.67%	110	7.90%	6	0.43%	1393
#25,					CQ	1670	93.14%	119	6.64%	4	0.22%	1793
EA-8 #1,					CR	1155	90.66%	113	8.87%	6	0.47%	1274
#3,					CS	2079	92.98%	154	6.89%	3	0.13%	2236
#15,					CT	1781	92.91%	135	7.04%	1	0.05%	1917
#16,					CW	186	86.92%	27	12.62%	1	0.47%	214
#18,					CX	1632	93.20%	111	6.34%	8	0.46%	1751
#22a+b,				Total:	18	27598		2140		68		29806
#23,				**Average:**			**92.59%**		**7.18%**		**0.23%**	
#26,												
#27												

* Context Key:

PSA-6 (6th Grade Problem Solving Assessment); Contexts (Question Numbers): R = Ranger Station (1-6), PT = A Patio (7-9), F = Fly One Day (10-13), B = Bird Watchers' Bulletin (14-15), S = Selling Tickets (16), and BS = Birds of All Sizes (17-24)

PSA-7 (7th Grade Problem Solving Assessment); Contexts (Item Numbers): B = Baby Feeding (1-7), PN = The Pentagon (8-10), A = Airships (11-14), PY = Pyramids (15-21), and PYG = Playground (22-26).

PSA-8 (8th Grade Problem Solving Assessment); Contexts (Item Numbers): CM = Club Members (1), LP = Lopsided (2-4), KC = Key Cards (5-7), SS = See Saw (8-9), ST = Stretch (10-14), PK = Parking (15-17), and CU = Cubes (18-21).

EA-6 (6th Grade External Assessment); Contexts = Item Numbers: 8, 11,17, 24a+24b, 25a+25b, and 26.

EA-7 (7th Grade External Assessment); Contexts = Item Numbers: 5, 8, 18, 21, 22, 24a+24b, and 25.

EA-8 (8th Grade External Assessment); Contexts = Item Numbers: 1, 3, 15, 16, 18, 22a+b, 23, 26 and 27.

Appendix B

Interrater Reliability–Problem Solving Assessment

Table B1
Interrater Agreement on 1998-1999 Grade 6 Problem-Solving Assessment

Context	Item Number	Assessments *(N)*	Total Agreement *(N)*	*%*	Single Adjudication *(N)*	*%*	Multiple Adjudications *(N)*	*%*
Ranger Station	1	700	673	96.14%	27	3.86%	0	0.00%
	2	700	581	83.00%	118	16.86%	1	0.14%
	3	700	592	84.57%	99	14.14%	9	1.29%
	4	700	651	93.00%	47	6.71%	2	0.29%
	5	700	630	90.00%	66	9.43%	4	0.57%
	6	700	576	82.29%	113	16.14%	11	1.57%
	Total	4200	3703		470		27	
	Average			**88.17%**		**11.19%**		**0.64%**
A Patio	7	700	689	98.43%	11	1.57%	0	0.00%
	8	700	661	94.43%	39	5.57%	0	0.00%
	9	700	556	79.43%	134	19.14%	10	1.43%
	Total	2100	1906		184		10	
	Average			**90.76%**		**8.76%**		**0.48%**
Fly One Day	10	700	684	97.71%	16	2.29%	0	0.00%
	11	700	665	95.00%	34	4.86%	1	0.14%
	12	700	646	92.29%	52	7.43%	2	0.29%
	13	700	645	92.14%	53	7.57%	2	0.29%
	Total	2800	2640		155		5	
	Average			**94.29%**		**5.54%**		**0.18%**
Bird Watchers' Bulletin	14	700	642	91.71%	57	8.14%	1	0.14%
	15	700	618	88.29%	79	11.29%	3	0.43%
	Total	1400	1260		136		4	
	Average			**90.00%**		**9.71%**		**0.29%**
Selling Tickets	16	700	591	84.43%	99	14.14%	10	1.43%
	Total	700	591		99		10	
	Average			**84.43%**		**14.14%**		**1.43%**
Birds of All Sizes	17	700	630	90.00%	67	9.57%	3	0.43%
	18	700	627	89.57%	71	10.14%	2	0.29%
	19	700	668	95.43%	31	4.43%	1	0.14%
	20	700	619	88.43%	81	11.57%	0	0.00%
	21	700	653	93.29%	46	6.57%	1	0.14%
	22	700	646	92.29%	54	7.71%	0	0.00%
	23	700	643	91.86%	54	7.71%	3	0.43%
	24	700	628	89.71%	69	9.86%	3	0.43%
	Total	5600	5114		473		13	
	Average			**91.32%**		**8.45%**		**0.23%**
PSA6	Total	16800	15214		1517		69	
	Average			**90.56%**		**9.03%**		**0.41%**

Interrater Agreement, By Scoring Institute and Rater

Table B2

Interrater Agreement by District for 1998-1999 Grade 6 Problem Solving Assessment

Context	Item Number	District 1 Assessment *(N)*	District 1 Agreement *(N)*	District 1 Agreement *%*	District 2 Assessment *(N)*	District 2 Agreement *(N)*	District 2 Agreement *%*	District 3 Assessment *(N)*	District 3 Agreement *(N)*	District 3 Agreement *%*	District 4 Assessment *(N)*	District 4 Agreement *(N)*	District 4 Agreement *%*
Ranger Station	1	196	187	95.41%	275	264	96.00%	122	121	99.18%	107	101	94.39%
	2	196	168	85.71%	275	228	82.91%	122	84	68.85%	107	101	94.39%
	3	196	158	80.61%	275	248	90.18%	122	90	73.77%	107	96	89.72%
	4	196	181	92.35%	275	259	94.18%	122	108	88.52%	107	103	96.26%
	5	196	170	86.73%	275	254	92.36%	122	104	85.25%	107	102	95.33%
	6	196	161	82.14%	275	234	85.09%	122	87	71.31%	107	94	87.85%
	Total	1176	1025		1650	1487		732	594		642	597	
	Average			**87.16%**			**90.12%**			**81.15%**			**92.99%**
A Patio	7	196	192	97.96%	275	273	99.27%	122	120	98.36%	107	104	97.20%
	8	196	185	94.39%	275	260	94.55%	122	113	92.62%	107	103	96.26%
	9	196	157	80.10%	275	221	80.36%	122	87	71.31%	107	91	85.05%
	Total	588	534		825	754		366	320		321	298	
	Average			**90.82%**			**91.39%**			**87.43%**			**92.83%**
Fly One Day	10	196	194	98.98%	275	267	97.09%	122	121	99.18%	107	102	95.33%
	11	196	189	96.43%	275	265	96.36%	122	110	90.16%	107	101	94.39%
	12	196	181	92.35%	275	254	92.36%	122	112	91.80%	107	99	92.52%
	13	196	176	89.80%	275	258	93.82%	122	107	87.70%	107	104	97.20%
	Total	784	740		1100	1044		488	450		428	406	
	Average			**94.39%**			**94.91%**			**92.21%**			**94.86%**
Bird Watchers' Bulletin	14	195	183	93.85%	275	246	89.45%	122	113	92.62%	107	99	92.52%
	15	196	175	89.29%	275	235	85.45%	122	105	86.07%	107	103	96.26%
	Total	391	358		550	481		244	218		214	202	
	Average			**91.56%**			**87.45%**			**89.34%**			**94.39%**
Selling Tickets	16	196	167	85.20%	275	247	89.82%	122	82	67.21%	107	95	88.79%
	Total	196	167		275	247		122	82		107	95	
	Average			**85.20%**			**89.82%**			**67.21%**			**88.79%**
Birds of All Sizes	17	196	175	89.29%	275	248	90.18%	122	111	90.98%	107	96	89.72%
	18	196	173	88.27%	275	255	92.73%	122	100	81.97%	107	99	92.52%
	19	196	186	94.90%	275	260	94.55%	122	119	97.54%	107	103	96.26%
	20	196	165	84.18%	275	250	90.91%	122	107	87.70%	107	97	90.65%
	21	196	182	92.86%	275	264	96.00%	122	110	90.16%	107	97	90.65%
	22	196	176	89.80%	275	263	95.64%	122	110	90.16%	107	97	90.65%
	23	196	169	86.22%	275	264	96.00%	122	108	88.52%	107	102	95.33%
	24	196	168	85.71%	275	250	90.91%	122	110	90.16%	107	100	93.46%
	Total	1568	1394		2200	2054		976	875		856	791	
	Average			**88.90%**			**93.36%**			**89.65%**			**92.41%**
PSA6	Total	4703	4218		6600	6067		2928	2539		2568	2389	
	Average			**89.69%**			**91.92%**			**86.71%**			**93.03%**

Interrater Agreement, By Scoring Institute and Rater

Table B3

Interrater Agreement by Program for 1998-1999 Grade 6 Problem Solving Assessments

Context	Item Number	Conventional Curricula			Mathematics in Context		
		Assessments	Agreement		Assessments	Agreement	
		(N)	*(N)*	*%*	*(N)*	*(N)*	*%*
Ranger Station	1	56	54	96.43%	644	619	96.12%
	2	56	50	89.29%	644	531	82.45%
	3	56	48	85.71%	644	544	84.47%
	4	56	50	89.29%	644	601	93.32%
	5	56	51	91.07%	644	579	89.91%
	6	56	49	87.50%	644	527	81.83%
	Total	336	302		3864	3401	
	Average			**89.88%**			**88.02%**
A Patio	7	56	56	100.00%	644	633	98.29%
	8	56	48	85.71%	644	613	95.19%
	9	56	44	78.57%	644	512	79.50%
	Total	168	148		1932	1758	
	Average			**88.10%**			**90.99%**
Fly One Day	10	56	55	98.21%	644	629	97.67%
	11	56	54	96.43%	644	611	94.88%
	12	56	53	94.64%	644	593	92.08%
	13	56	56	100.00%	644	589	91.46%
	Total	224	218		2576	2422	
	Average			**97.32%**			**94.02%**
Bird Watchers' Bulletin	14	56	45	80.36%	644	596	92.55%
	15	56	53	94.64%	644	565	87.73%
	Total	112	98		1288	1161	
	Average			**87.50%**			**90.14%**
Selling Tickets	16	56	48	85.71%	644	543	84.32%
	Total	56	48		644	543	
	Average			**85.71%**			**84.32%**
Birds of All Sizes	17	56	46	82.14%	644	584	90.68%
	18	56	47	83.93%	644	580	90.06%
	19	56	51	91.07%	644	617	95.81%
	20	56	44	78.57%	644	575	89.29%
	21	56	51	91.07%	644	602	93.48%
	22	56	53	94.64%	644	593	92.08%
	23	56	50	89.29%	644	593	92.08%
	24	56	48	85.71%	644	580	90.06%
	Total	448	390		5152	4724	
	Average			**87.05%**			**91.69%**
PSA6	Total	1344	1204		15456	14009	
	Average			**89.58%**			**90.64%**

Interrater Agreement, By Scoring Institute and Rater

Table B4

Interrater Agreement on 1998-1999 Grade 7 Problem-Solving Assessment

Context	Question Number	Assessments *(N)*	Total Agreement *(N)*	Total Agreement	Single Adjudication *(N)*	Single Adjudication	Multiple Adjudications *(N)*	Multiple Adjudication
BabyFeeding	1	826	806	97.58%	20	2.42%	0	0.00%
	2	826	716	86.68%	106	12.83%	4	0.48%
	3	826	760	92.01%	64	7.75%	2	0.24%
	4	826	702	84.99%	123	14.89%	1	0.12%
	5	826	778	94.19%	46	5.57%	2	0.24%
	6	826	783	94.79%	42	5.08%	1	0.12%
	7	826	733	88.74%	92	11.14%	1	0.12%
	Total	5782	5278		493		11	
	Average			**91.28%**		**8.53%**		**0.19%**
The Pentagon	8	826	698	84.50%	126	15.25%	2	0.24%
	9	826	584	70.70%	226	27.36%	16	1.94%
	10	826	770	93.22%	50	6.05%	6	0.73%
	Total	2478	2052		402		24	
	Average			**82.81%**		**16.22%**		**0.97%**
Airships	11	826	809	97.94%	16	1.94%	1	0.12%
	12	826	575	69.61%	190	23.00%	61	7.38%
	13	826	773	93.58%	50	6.05%	3	0.36%
	14	826	793	96.00%	33	4.00%	0	0.00%
	Total	3304	2950		289		65	
	Average			**89.29%**		**8.75%**		**1.97%**
Pyramids	15	826	678	82.08%	144	17.43%	4	0.48%
	16	826	768	92.98%	57	6.90%	1	0.12%
	17	826	792	95.88%	31	3.75%	3	0.36%
	18	826	693	83.90%	127	15.38%	6	0.73%
	19	826	753	91.16%	71	8.60%	2	0.24%
	20	826	770	93.22%	56	6.78%	0	0.00%
	21	826	761	92.13%	63	7.63%	2	0.24%
	Total	5782	5215		549		18	
	Average			**90.19%**		**9.49%**		**0.31%**
Playgrounds	22	826	803	97.22%	23	2.78%	0	0.00%
	23	826	795	96.25%	30	3.63%	1	0.12%
	24	826	734	88.86%	86	10.41%	6	0.73%
	25	826	802	97.09%	23	2.78%	1	0.12%
	26	826	785	95.04%	39	4.72%	2	0.24%
	Total	4130	3919		201		10	
	Average			**94.89%**		**4.87%**		**0.24%**
PSA7	Total	21476	19414		1934		128	
	Average			**90.40%**		**9.01%**		**0.60%**

Interrater Agreement, By Scoring Institute and Rater

Table B5

Interrater Agreement by District for 1998-1999 Grade 7 Problem Solving Assessment

Context	Item Number	District 1 Assessment (N)	District 1 Agreement (N)	District 1 Agreement %	District 2 Assessment (N)	District 2 Agreement (N)	District 2 Agreement %	District 3 Assessment (N)	District 3 Agreement (N)	District 3 Agreement %	District 4 Assessment (N)	District 4 Agreement (N)	District 4 Agreement %
Baby Feeding	1	249	247	99.20%	250	242	96.80%	138	135	97.83%	189	182	96.30%
	2	249	213	85.54%	250	216	86.40%	138	118	85.51%	189	169	89.42%
	3	249	231	92.77%	250	233	93.20%	138	128	92.75%	189	168	88.89%
	4	249	208	83.53%	250	214	85.60%	138	112	81.16%	189	168	88.89%
	5	249	237	95.18%	250	235	94.00%	138	125	90.58%	189	181	95.77%
	6	249	236	94.78%	250	238	95.20%	138	128	92.75%	189	181	95.77%
	7	249	227	91.16%	250	223	89.20%	138	122	88.41%	189	161	85.19%
	Total	1743	1599		1750	1601		966	868		1323	1210	
	Average			**91.74%**			**91.49%**			**89.86%**			**91.46%**
The Pentagon	8	249	202	81.12%	250	217	86.80%	138	114	82.61%	189	165	87.30%
	9	249	172	69.08%	250	177	70.80%	138	80	57.97%	189	155	82.01%
	10	249	231	92.77%	250	232	92.80%	138	129	93.48%	189	178	94.18%
	Total	747	605		750	626		414	323		567	498	
	Average			**80.99%**			**83.47%**			**78.02%**			**87.83%**
Airships	11	249	244	97.99%	250	245	98.00%	138	134	97.10%	189	186	98.41%
	12	249	171	68.67%	250	163	65.20%	138	96	69.57%	189	145	76.72%
	13	249	231	92.77%	250	234	93.60%	138	123	89.13%	189	185	97.88%
	14	249	236	94.78%	250	243	97.20%	138	131	94.93%	189	183	96.83%
	Total	996	882		1000	885		552	484		756	699	
	Average			**88.55%**			**88.50%**			**87.68%**			**92.46%**
Pyramids	15	249	213	85.54%	250	204	81.60%	138	109	78.99%	189	152	80.42%
	16	249	242	97.19%	250	235	94.00%	138	120	86.96%	189	171	90.48%
	17	249	240	96.39%	250	240	96.00%	138	127	92.03%	189	185	97.88%
	18	249	215	86.35%	250	210	84.00%	138	107	77.54%	189	161	85.19%
	19	249	231	92.77%	250	219	87.60%	138	126	91.30%	189	177	93.65%
	20	249	233	93.57%	250	230	92.00%	138	129	93.48%	189	178	94.18%
	21	249	231	92.77%	250	235	94.00%	138	112	81.16%	189	183	96.83%
	Total	1743	1605		1750	1573		966	830		1323	1207	
	Average			**92.08%**			**89.89%**			**85.92%**			**91.23%**
Playgrounds	22	249	245	98.39%	250	237	94.80%	138	134	97.10%	189	187	98.94%
	23	249	239	95.98%	250	241	96.40%	138	134	97.10%	189	181	95.77%
	24	249	215	86.35%	250	232	92.80%	138	115	83.33%	189	172	91.01%
	25	249	243	97.59%	250	244	97.60%	138	129	93.48%	189	186	98.41%
	26	249	234	93.98%	250	240	96.00%	138	131	94.93%	189	180	95.24%
	Total	1245	1176		1250	1194		690	643		945	906	
	Average			**94.46%**			**95.52%**			**93.19%**			**95.87%**
PSA7	Total	6474	5867		6500	5879		3588	3148		2568	2389	
	Average			**90.62%**			**90.45%**			**87.74%**			**93.03%**

Interrater Agreement, By Scoring Institute and Rater

Table B6

Interrater Agreement by Program for 1998-1999 Grade 7 Problem Solving Assessments

Context	Item Number	Conventional Curricula Assessments *(N)*	Conventional Curricula Agreement *(N)*	Conventional Curricula Agreement *%*	Mathematics in Context Assessments *(N)*	Mathematics in Context Agreement *(N)*	Mathematics in Context Agreement *%*
Baby Feeding	1	119	117	98.32%	707	689	97.45%
	2	119	103	86.55%	707	613	86.70%
	3	119	113	94.96%	707	647	91.51%
	4	119	98	82.35%	707	604	85.43%
	5	119	111	93.28%	707	667	94.34%
	6	119	114	95.80%	707	669	94.63%
	7	119	107	89.92%	707	626	88.54%
	Total	833	763		4949	4515	
	Average			**91.60%**			**91.23%**
Pentagon	8	119	101	84.87%	707	597	84.44%
	9	119	81	68.07%	707	503	71.15%
	10	119	112	94.12%	707	658	93.07%
	Total	357	294		2121	1758	
	Average			**82.35%**			**82.89%**
Airships	11	119	116	97.48%	707	693	98.02%
	12	119	91	76.47%	707	484	68.46%
	13	119	109	91.60%	707	664	93.92%
	14	119	111	93.28%	707	682	96.46%
	Total	476	427		2828	2523	
	Average			**89.71%**			**89.21%**
Pyramids	15	119	102	85.71%	707	576	81.47%
	16	119	115	96.64%	707	653	92.36%
	17	119	119	100.00%	707	673	95.19%
	18	119	105	88.24%	707	588	83.17%
	19	119	110	92.44%	707	643	90.95%
	20	119	112	94.12%	707	658	93.07%
	21	119	113	94.96%	707	648	91.65%
	Total	833	776		4949	4439	
	Average			**93.16%**			**89.69%**
Playgrounds	22	119	116	97.48%	707	687	97.17%
	23	119	113	94.96%	707	682	96.46%
	24	119	107	89.92%	707	627	88.68%
	25	119	119	100.00%	707	683	96.61%
	25	119	114	95.80%	707	671	94.91%
	Total	595	569		3535	3350	
	Average			**95.63%**			**94.77%**
PSA7	Total	3094	2829		18382	16585	
	Average			**91.44%**			**90.22%**

Interrater Agreement, By Scoring Institute and Rater

Table B7

Interrater Agreement on 1998-1999 Grade 8 Problem-Solving Assessment

Context	Question Number	Assessments *(N)*	Total Agreement *(N)*	Total Agreement	Single Adjudication (*N*)	Single Adjudication	Multiple Adjudications *(N)*	Multiple Adjudication
Club Members	1	507	418	82.45%	84	16.57%	5	0.99%
	Total	507	418		84		5	
	Average			**82.45%**		**16.57%**		**0.99%**
Lopsided	2	507	473	93.29%	32	6.31%	2	0.39%
	3	507	448	88.36%	59	11.64%	0	0.00%
	4	507	439	86.59%	64	12.62%	4	0.79%
	Total	1521	1360		155		6	
	Average			**89.41%**		**10.19%**		**0.39%**
Key Cards	5	507	473	93.29%	32	6.31%	2	0.39%
	6	507	474	93.49%	33	6.51%	0	0.00%
	7	507	471	92.90%	36	7.10%	0	0.00%
	Total	1521	1418		101		2	
	Average			**93.23%**		**6.64%**		**0.13%**
Seesaw	8	507	486	95.86%	21	4.14%	0	0.00%
	9	507	491	96.84%	15	2.96%	1	0.20%
	Total	1014	977		36		1	
	Average			**96.35%**		**3.55%**		**0.10%**
Stretch	10	507	456	89.94%	51	10.06%	0	0.00%
	11	507	485	95.66%	22	4.34%	0	0.00%
	12	507	502	99.01%	5	0.99%	0	0.00%
	13	507	474	93.49%	33	6.51%	0	0.00%
	14	507	471	92.90%	35	6.90%	1	0.20%
	Total	2535	2388		146		1	
	Average			**94.20%**		**5.76%**		**0.04%**
Parking	15	507	474	93.49%	33	6.51%	0	0.00%
	16	507	476	93.89%	31	6.11%	0	0.00%
	17	507	449	88.56%	55	10.85%	3	0.59%
	Total	1521	1399		119		3	
	Average			**91.98%**		**7.82%**		**0.20%**
Cubes	18	507	462	91.12%	42	8.28%	3	0.59%
	19	507	492	97.04%	14	2.76%	1	0.20%
	20	507	487	96.06%	20	3.94%	0	0.00%
	21	507	490	96.65%	17	3.35%	0	0.00%
	Total	2028	1931		93		4	
	Average			**95.22%**		**4.59%**		**0.20%**
PSA8	Total	10647	9891		734		22	
	Average			**92.90%**		**6.89%**		**0.21%**

Interrater Agreement, By Scoring Institute and Rater

Table B8

Interrater Agreement by District for 1998-1999 Grade 8 Problem Solving Assessment

Context	Item Number	District 1 Assessment (N)	District 1 Agreement (N)	District 1 Agreement %	District 2 Assessment (N)	District 2 Agreement (N)	District 2 Agreement %	District 3 Assessment (N)	District 3 Agreement (N)	District 3 Agreement %	District 4 Assessment (N)	District 4 Agreement (N)	District 4 Agreement %
Club Members	1	168	144	85.71%	140	116	82.86%	71	60	84.51%	128	98	76.56%
	Total	168	144		140	116		71	60		128	98	
	Average			**85.71%**			**82.86%**			**84.51%**			**76.56%**
Lopsided	2	168	164	97.62%	140	125	89.29%	71	63	88.73%	128	121	94.53%
	3	168	154	91.67%	140	122	87.14%	71	57	80.28%	128	115	89.84%
	4	168	156	92.86%	140	112	80.00%	71	54	76.06%	128	117	91.41%
	Total	504	474		420	359		213	174		384	353	
	Average			**94.05%**			**85.48%**			**81.69%**			**91.93%**
Key Cards	5	168	152	90.48%	140	130	92.86%	71	70	98.59%	128	121	94.53%
	6	168	156	92.86%	140	134	95.71%	71	67	94.37%	128	117	91.41%
	7	168	159	94.64%	140	129	92.14%	71	61	85.92%	128	122	95.31%
	Total	504	467		420	393		213	198		384	360	
	Average			**92.66%**			**93.57%**			**92.96%**			**93.75%**
Seesaw	8	168	159	94.64%	140	137	97.86%	71	64	90.14%	128	126	98.44%
	9	168	159	94.64%	140	139	99.29%	71	69	97.18%	128	124	96.88%
	Total	336	318		280	276		142	133		256	250	
	Average			**94.64%**			**98.57%**			**93.66%**			**97.66%**
Stretch	10	168	153	91.07%	140	125	89.29%	71	62	87.32%	128	116	90.63%
	11	168	161	95.83%	140	135	96.43%	71	65	91.55%	128	124	96.88%
	12	168	167	99.40%	140	140	100.00%	71	68	95.77%	128	127	99.22%
	13	168	158	94.05%	140	127	90.71%	71	65	91.55%	128	124	96.88%
	14	168	157	93.45%	140	127	90.71%	71	63	88.73%	128	124	96.88%
	Total	840	796		700	654		355	323		640	615	
	Average			**94.76%**			**93.43%**			**90.99%**			**96.09%**
Parking	15	168	158	94.05%	140	134	95.71%	71	67	94.37%	128	115	89.84%
	16	168	163	97.02%	140	134	95.71%	71	60	84.51%	128	119	92.97%
	17	168	147	87.50%	140	122	87.14%	71	57	80.28%	128	123	96.09%
	Total	504	468		420	390		213	184		384	357	
	Average			**92.86%**			**92.86%**			**86.38%**			**92.97%**
Cubes	18	168	160	95.24%	140	127	90.71%	71	63	88.73%	128	112	87.50%
	19	168	163	97.02%	140	136	97.14%	71	67	94.37%	128	126	98.44%
	20	168	161	95.83%	140	132	94.29%	71	70	98.59%	128	124	96.88%
	21	168	159	94.64%	140	138	98.57%	71	66	92.96%	128	127	99.22%
	Total	672	643		560	533		284	266		512	489	
	Average			**95.68%**			**95.18%**			**93.66%**			**95.51%**
PSA8	Total	3528	3310		2940	2721		1491	1338		2688	2522	
	Average			**93.82%**			**92.55%**			**89.74%**			**93.82%**

Interrater Agreement, By Scoring Institute and Rater

Table B9

Interrater Agreement by Program for 1998-1999 Grade 8 Problem Solving Assessments

Context	Item Number	Conventional Curricula Assessments *(N)*	Conventional Curricula Agreement *(N)*	Conventional Curricula Agreement *%*	Mathematics in Context Assessments *(N)*	Mathematics in Context Agreement *(N)*	Mathematics in Context Agreement *%*
Club Members	1	105	92	87.62%	402	326	81.09%
	Total	105	92		402	326	
	Average			**87.62%**			**81.09%**
Lopsided	2	105	100	95.24%	402	373	92.79%
	3	105	96	91.43%	402	352	87.56%
	4	105	93	88.57%	402	346	86.07%
	Total	315	289		1206	1071	
	Average			**91.75%**			**88.81%**
Key Cards	5	105	97	92.38%	402	376	93.53%
	6	105	102	97.14%	402	372	92.54%
	7	105	101	96.19%	402	370	92.04%
	Total	315	300		1206	1118	
	Average			**95.24%**			**92.70%**
Seesaw	8	105	102	97.14%	402	384	95.52%
	9	105	102	97.14%	402	389	96.77%
	Total	210	204		804	773	
	Average			**97.14%**			**96.14%**
Stretch	10	105	90	85.71%	402	366	91.04%
	11	105	102	97.14%	402	383	95.27%
	12	105	104	99.05%	402	398	99.00%
	13	105	99	94.29%	402	375	93.28%
	14	105	99	94.29%	402	372	92.54%
	Total	525	494		2010	1894	
	Average			**94.10%**			**94.23%**
Parking	15	105	100	95.24%	402	374	93.03%
	16	105	101	96.19%	402	375	93.28%
	17	105	94	89.52%	402	355	88.31%
	Total	315	295		1206	1104	
	Average			**93.65%**			**91.54%**
Cubes	18	105	98	93.33%	402	364	90.55%
	19	105	103	98.10%	402	389	96.77%
	20	105	102	97.14%	402	385	95.77%
	21	105	101	96.19%	402	389	96.77%
	Total	420	404		1608	1527	
	Average			**96.19%**			**94.96%**
PSA8	Total	2205	2078		8442	7813	
	Average			**94.24%**			**92.55%**

Interrater Agreement, By Scoring Institute and Rater

Appendix C

Interrater Reliability–External Assessment

Table C1
Interrater Agreement on 1998-1999 Grade 6 External Assessments

Costructed Response Item	Assessments *(N)*	Total Agreement *(N)*	Total Agreement *%*	Single Adjudication *(N)*	Single Adjudication *%*	Multiple Adjudication *(N)*	Multiple Adjudication *%*
8	713	584	81.91%	124	17.39%	5	0.70%
11	713	701	98.32%	12	1.68%	0	0.00%
17	713	674	94.53%	39	5.47%	0	0.00%
24a	713	709	99.44%	4	0.56%	0	0.00%
24b	713	708	99.30%	5	0.70%	0	0.00%
25a	713	701	98.32%	12	1.68%	0	0.00%
25b	713	689	96.63%	22	3.09%	2	0.28%
26	713	634	88.92%	77	10.80%	2	0.28%
Total	5704	5400		295		9	
Average			**94.67%**		**5.17%**		**0.16%**

Interrater Agreement, By Scoring Institute and Rater

Table C2

Interrater Agreement by District for 1998-1999 Grade 6 External Assessment

Costructed Response Item	District 1			District 2			District 3			District 4		
	Assessments	Total Agreement		Assessments	Total Agreement		Assessments	Total Agreement		Assessments	Total Agreement	
	(N)	*(N)*	*%*	*(N)*	*(N)*	*%*	*(N)*	*(N)*	*%*	*(N)*	*(N)*	*%*
8	249	204	81.93%	242	200	82.64%	118	90	76.27%	104	89	85.58%
11	249	245	98.39%	242	239	98.76%	118	116	98.31%	104	101	97.12%
17	249	241	96.79%	242	228	94.21%	118	104	88.14%	104	101	97.12%
24a	249	249	100.00%	242	241	99.59%	118	118	100.00%	104	101	97.12%
24b	249	249	100.00%	242	237	97.93%	118	118	100.00%	103	103	100.00%
25a	249	243	97.59%	242	239	98.76%	118	116	98.31%	104	103	99.04%
25b	249	239	95.98%	242	238	98.35%	118	112	94.92%	104	100	96.15%
26	249	219	87.95%	242	217	89.67%	118	104	88.14%	104	94	90.38%
Total	1992	1889		1936	1839		944	878		831	792	
Average			**94.83%**			**94.99%**			**93.01%**			**95.31%**

Table C3

Interrater Agreement by Program for 1998-1999 Grade 6 External Assessments

Conventional				***Mathematics in Context***			
Costructed-Response Item	Assessments *(N)*	Total Agreement *(N)*	*%*	Costructed-Response Item	Assessments *(N)*	Total Agreement *(N)*	*%*
8	86	73	84.88%	8	627	510	81.34%
11	86	85	98.84%	11	627	616	98.25%
17	86	83	96.51%	17	627	591	94.26%
24a	86	86	100.00%	24a	627	623	99.36%
24b	86	86	100.00%	24b	626	621	99.20%
25a	86	84	97.67%	25a	627	617	98.41%
25b	86	83	96.51%	25b	627	606	96.65%
26	86	76	88.37%	26	627	558	89.00%
Total	688	656		Total	5015	4742	
Average			**95.35%**	**Average**			**94.56%**

Interrater Agreement, By Scoring Institute and Rater

Table C4
Interrater Agreement on 1998-1999 Grade 7 External Assessments

Costructed-Response Item	Assessments *(N)*	Total Agreement *(N)*	Total Agreement *%*	Single Adjudication *(N)*	Single Adjudication *%*	Multiple Adjudication *(N)*	Multiple Adjudication *%*
5	806	612	75.93%	184	22.83%	10	1.24%
8	806	788	97.77%	18	2.23%	0	0.00%
18	806	657	81.51%	147	18.24%	2	0.25%
21	806	772	95.78%	32	3.97%	2	0.25%
22	806	800	99.26%	6	0.74%	0	0.00%
24a	806	782	97.02%	24	2.98%	0	0.00%
24b	806	776	96.28%	29	3.60%	1	0.12%
25	806	729	90.45%	76	9.43%	1	0.12%
Total	6448	5916		516		16	
Average			**91.75%**		**8.00%**		**0.25%**

Interrater Agreement, By Scoring Institute and Rater

Table C5

Interrater Agreement by District for 1998-1999 Grade 7 External Assessment

Costructed-Response Item	District 1			District 2			District 3			District 4		
	Assessments *(N)*	Total Agreement *(N)*	*%*	Assessments *(N)*	Total Agreement *(N)*	*%*	Assessments *(N)*	Total Agreement *(N)*	*%*	Assessments *(N)*	Total Agreement *(N)*	*%*
5	248	178	71.77%	234	185	79.06%	128	102	79.56%	196	147	75.00%
8	248	245	98.79%	234	228	97.44%	128	126	98.54%	196	189	96.43%
18	248	199	80.24%	234	197	84.19%	128	94	71.53%	196	167	85.20%
21	248	238	95.97%	234	224	95.73%	128	121	94.89%	196	189	96.43%
22	248	246	99.19%	234	232	99.15%	128	126	98.54%	196	196	100.00%
24a	248	240	96.77%	234	228	97.44%	128	121	94.89%	196	193	98.47%
24b	248	235	94.76%	234	227	97.01%	128	122	95.62%	196	192	97.96%
25	248	231	93.15%	234	212	90.60%	128	110	86.13%	196	176	89.80%
Total	1984	1812		1872	1733		1096	986		1568	1449	
Average			**91.33%**			**92.57%**			**89.96%**			**92.41%**

Interrater Agreement, By Scoring Institute and Rater

Table C6
Interrater Agreement by Program for 1998-1999 Grade 7 External Assessments

	Conventional			***Mathematics in Context***			
Costructed-Response Item	Assessments (N)	Total Agreement *(N)*	*%*	Costructed-Response Item	Assessments (N)	Total Agreement *(N)*	*%*
5	113	86	76.11%	5	693	526	75.90%
8	113	110	97.35%	8	693	678	97.84%
18	113	94	83.19%	18	693	563	81.24%
21	113	109	96.46%	21	693	663	95.67%
22	113	111	98.23%	22	693	689	99.42%
24a	113	110	97.35%	24a	693	672	96.97%
24b	113	109	96.46%	24b	693	667	96.25%
25	113	109	96.46%	25	693	620	89.47%
Total	904	838		Total	5544	5078	
Average			**92.70%**	**Average**			**91.59%**

Table C7
Interrater Agreement on 1998-1999 Grade 8 External Assessments

Costructed-Response Item	Assessments *(N)*	Total Agreement *(N)*	Total Agreement *%*	Single Adjudication *(N)*	Single Adjudication *%*	Multiple Adjudication *(N)*	Multiple Adjudication *%*
1	446	343	76.91%	100	22.42%	3	0.67%
4	446	438	98.21%	8	1.79%	0	0.00%
15	446	368	82.51%	75	16.82%	3	0.67%
16	446	440	98.65%	6	1.35%	0	0.00%
18	446	430	96.41%	14	3.14%	2	0.45%
22a	446	437	97.98%	9	2.02%	0	0.00%
22b	446	442	99.10%	4	0.90%	0	0.00%
23	446	426	95.52%	20	4.48%	0	0.00%
26	446	409	91.70%	36	8.07%	1	0.22%
27	446	413	92.60%	33	7.40%	0	0.00%
Total	4460	4146		305		9	
Average			**92.96%**		**6.84%**		**0.20%**

Interrater Agreement, By Scoring Institute and Rater

Table C8

Interrater Agreement by District for 1998-1999 Grade 8 External Assessment

Costructed-Response Item	District 1			District 2			District 3			District 4		
	Assessments	Total Agreement		Assessments	Total Agreement		Assessments	Total Agreement		Assessments	Total Agreement	
	(N)	*(N)*	*%*	*(N)*	*(N)*	*%*	*(N)*	*(N)*	*%*	*(N)*	*(N)*	*%*
1	159	129	81.13%	131	96	73.28%	65	47	72.31%	91	71	78.02%
4	159	156	98.11%	131	129	98.47%	65	63	96.92%	91	90	98.90%
15	159	125	78.62%	131	109	83.21%	65	46	70.77%	91	88	96.70%
16	159	158	99.37%	131	126	96.18%	65	65	100.00%	91	91	100.00%
18	159	156	98.11%	131	123	93.89%	65	60	92.31%	91	91	100.00%
22a	159	154	96.86%	131	130	99.24%	65	62	95.38%	91	91	100.00%
22b	159	156	98.11%	131	130	99.24%	65	65	100.00%	91	91	100.00%
23	159	155	97.48%	131	120	91.60%	65	61	93.85%	91	90	98.90%
26	159	152	95.60%	131	116	88.55%	65	51	78.46%	91	90	98.90%
27	159	151	94.97%	131	122	93.13%	65	50	76.92%	91	90	98.90%
Total	1590	1492		1310	1201		650	570		910	883	
Average			**93.84%**			**91.68%**			**87.69%**			**97.03%**

Table C9
Interrater Agreement by Program for 1998-1999 Grade 8 External Assessments

	Conventional				***Mathematics in Context***		
Costructed-Response Item	Assessments *(N)*	Total Agreement *(N)*	*%*	Costructed-Response Item	Assessments *(N)*	Total Agreement *(N)*	*%*
1	103	87	84.47%	1	343	256	74.64%
4	103	102	99.03%	4	343	336	97.96%
15	103	91	88.35%	15	343	277	80.76%
16	103	102	99.03%	16	343	338	98.54%
18	103	101	98.06%	18	343	329	95.92%
22a	103	100	97.09%	22a	343	337	98.25%
22b	103	102	99.03%	22b	343	340	99.13%
23	103	102	99.03%	23	343	324	94.46%
26	103	102	99.03%	26	343	307	89.50%
27	103	99	96.12%	27	343	314	91.55%
Total	1030	988		Total	3430	3158	
Average			**95.92%**	**Average**			**92.07%**

Interrater Agreement, By Scoring Institute and Rater

Table C10

Interrater Agreement for 1998-1999 External Assessment by Question Across Grades 6, 7, and 8

Context	6th Grade		7th Grade		8th Grade		Average Grades 6, 7, 8 Agreement
	Item	Agreement	Item	Agreement	Item	Agreement	
1	*8*	81.91%	*5*	75.93%	*1*	76.91%	78.32%
2	*11*	98.32%	*8*	97.77%	*4*	98.21%	98.07%
3	*17*	94.53%	*25*	90.45%	*23*	95.52%	93.08%
4a	*25a*	98.32%	*24a*	97.02%	*22a*	97.98%	97.71%
4b	*25b*	96.63%	*24b*	96.28%	*22b*	99.10%	97.05%
5	*26*	88.92%	*18*	81.51%	*26*	91.70%	86.51%
6a	*24a*	99.44%	*22*	99.26%	*16*	98.65%	99.19%
6b	*24b*	99.30%	--	--	--	--	99.30%
7	--	--	*21*	95.78%	*18*	96.41%	96.01%
8	--	--	--	--	*15*	82.51%	82.51%
9	--	--	--	--	*27*	92.60%	92.60%