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

Interrater Reliability at the 1999-2000 Scoring Institutes

(Working Paper #23)

Lorene Folgert and Mary Shafer

University of Wisconsin-Madison

September 2001

Folgert, L., & Shafer, M. C. (2001). *Interrater Reliability at the 1999-2000 Scoring Institutes* (*Mathematics in Context* Longitudinal/Cross-Sectional Study Working Paper No. 23). 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 7, and 8 in 1999-2000). The variables have been organized in five categories (prior, independent, intervening, outcome, and consequent). (See Figure 1 for variables and hypothesized relationships.)

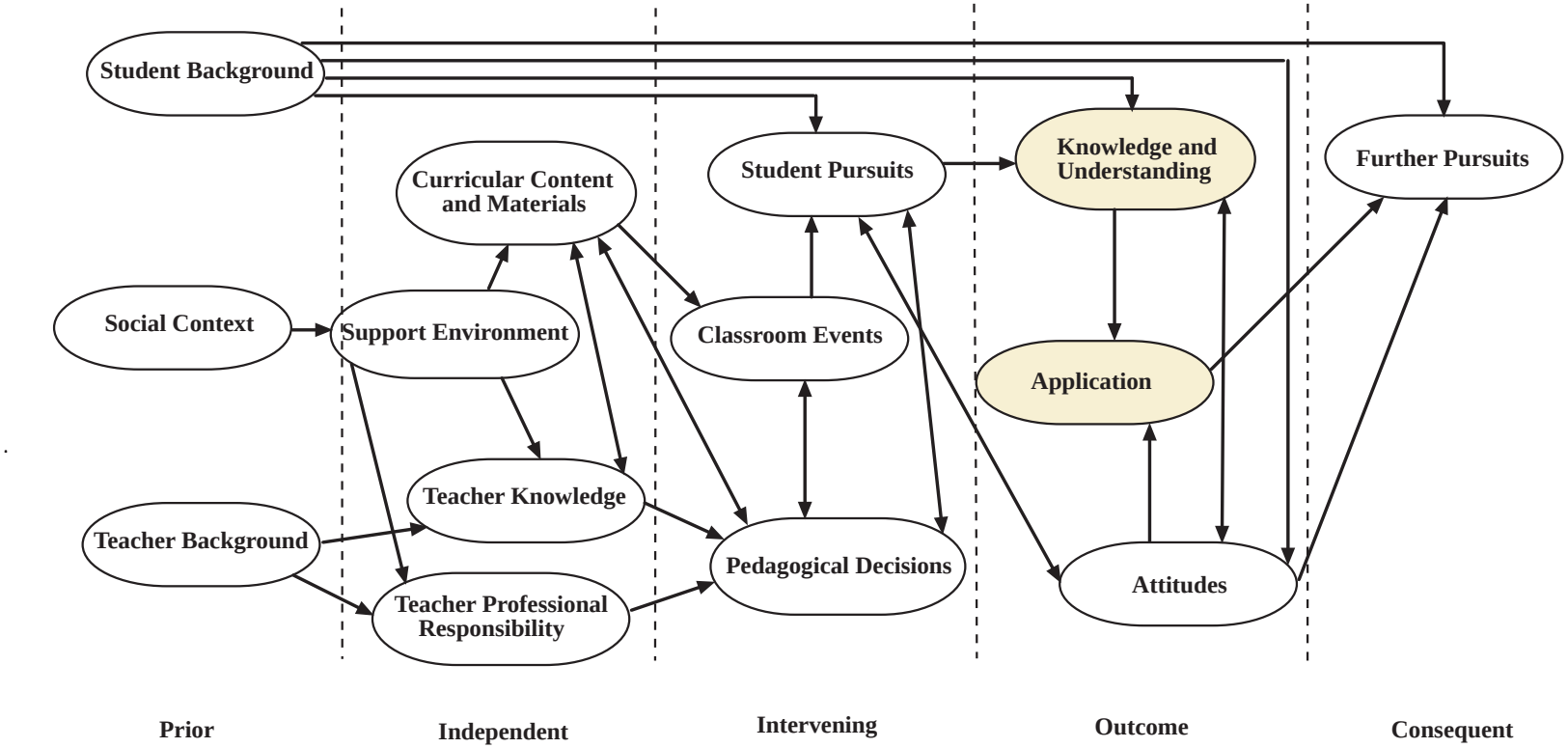

Figure 1. Revised model for the monitoring of school mathematics.

Interrater Reliability at the 1999-2000 Scoring Institutes

Four scoring institutes were held in 2000 to score the Problem Solving Assessments and External Assessments administered to seventh, and eighth-grade students in spring 2000 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 7–8) in each of the four districts in the study. The first two scoring institutes, held in spring 2000, were conducted in two of 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. Two more scoring institutes were held at the University of Wisconsin-Madison; teachers (Grades 3–12) from schools in the Madison area were the raters. One of the Madison institutes lasted one week and the other one lasted two days. These institutes were held in summer 2000.

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)*
7	318	315
8	374	308

[1] If there was a discrepancy between the scores, a third rater adjudicated. Occasionally, more adjudications were necessary. When two raters agreed upon a score, 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 12 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.

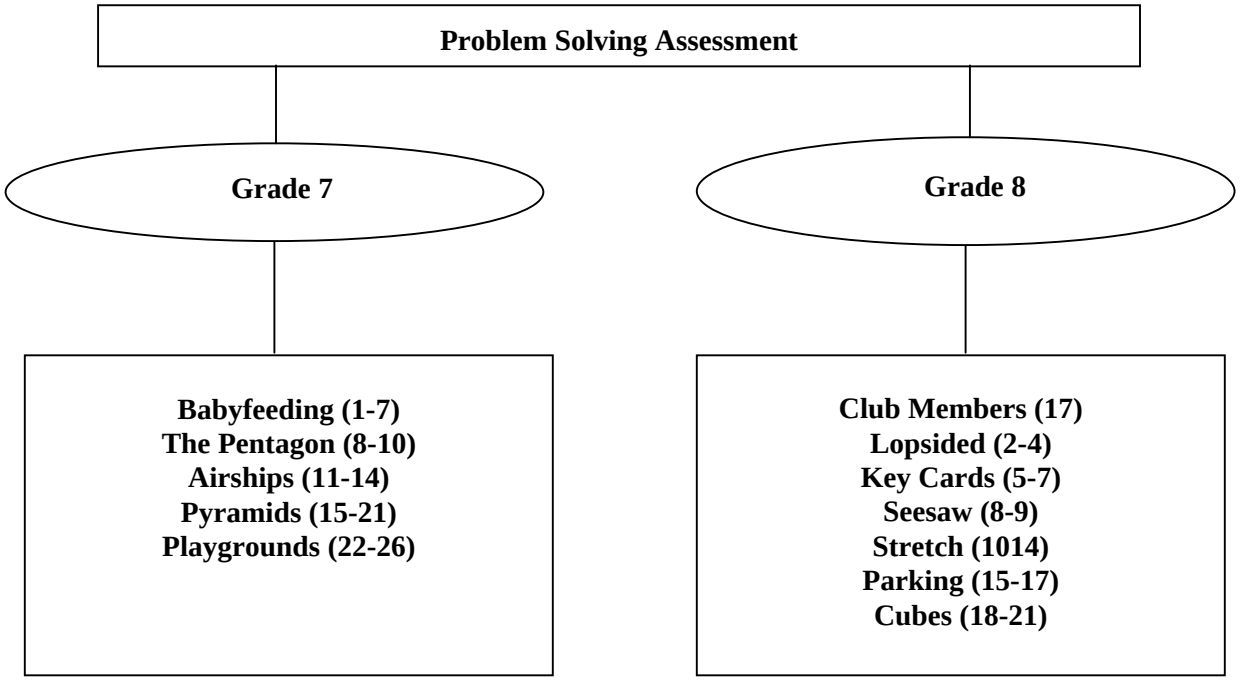

Figure 2. Contexts of the Grade 7 and Grade 8 Problem Solving Assessments.

External Assessment System

The External Assessment System is a set of grade-specific assessments 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, seven EA anchor items (constructed-response items) were repeated on each grade-specific assessment. In addition, two other constructed-response items were scored (Contexts 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.)

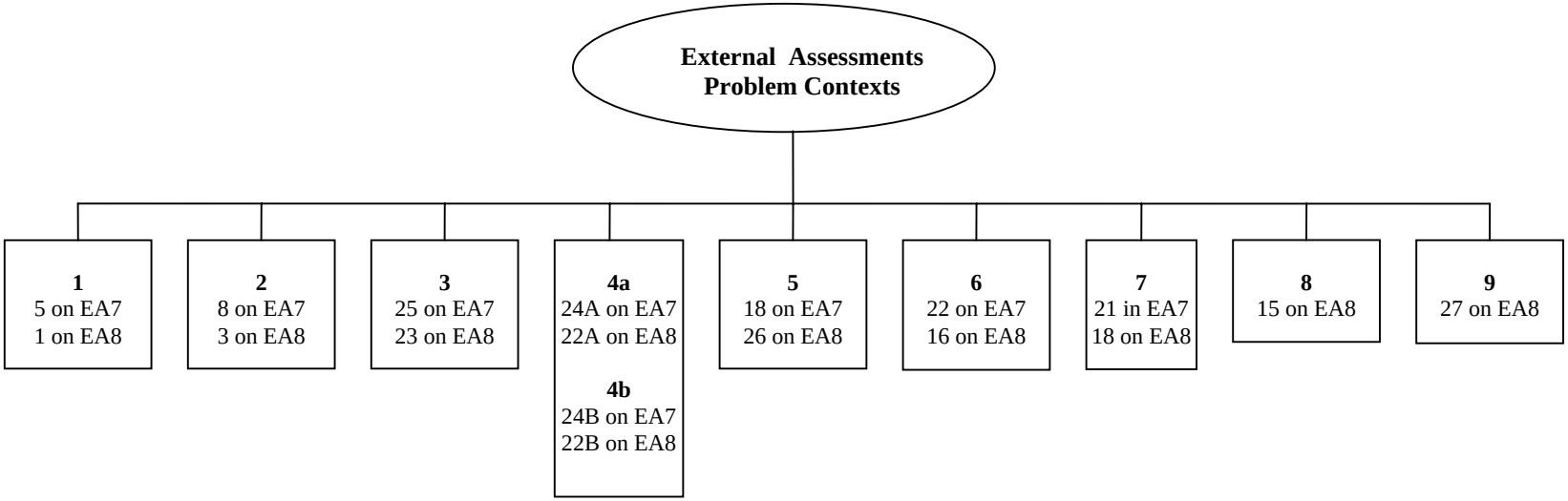

Figure 3. Constructed-Response Contexts of the Grade 7 and Grade 8 External Assessments

[2] The context numbers refer only to the context groupings in this paper.

Assessments: Scoring Procedures

A total of 21 contexts on both assessments was scored. The number of contexts scored at each institute varied from 1 to 13 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 institute (July 24, 2000), all raters were trained in the scoring of "Baby Feeding" (the first context, items 1–7, on the Grade 7 PSA). Test packets were randomly distributed. After raters finished rating this context, the coordinator randomly distributed the packet to different raters. The each set of scores was compared and the assessments with discrepancies were routed to third raters (called adjudicators). When all of the Grade 7 PSAs were scored and adjudicated for this context, the raters were trained in the scoring of "Airships" (the third context, Items 11–14, on the Grade 7 PSA). Raters then started scoring "Airships" in the same manner. (The next day after a quick review of the rubric, raters completed the scoring of "Airships".)

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 who 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 ↓ More	X	0	1			
	X	0	1	2		
	X	0	1	2	3	
	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 four scoring institutes were held on May 24–25, June 8, July 24–28, and August 7–8 (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

2000 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	11	184	4	3221	2799	86.90%	392	12.17%	30	0.93%
2	3	42	1	84	60	71.43%	22	26.19%	2	2.00%
3	18	1315	13	31686	29260	92.34%	2360	7.45%	66	0.21%
4	15	623	8	9341	8546	91.49%	775	8.30%	20	0.21%

The number of student responses given the same score points by two raters was determined and percentages were calculated. For example, of the 3221 student responses rated for the first scoring institute, 2799 student responses were assigned the same point scores by two raters. Therefore, the raters agreed on the point scores 86.90% of the time during the first scoring institute.

Interrater agreement was high for all four scoring institutes, ranging from a low of 71.43% at the second institute to a high of 92.34% at the third institute. The last 2 scoring institutes had much higher agreement than the first 2 scoring institutes. The higher agreement from the last 2 institutes was probably due to rater experience. These institutes were held in Madison by raters returning with 2 years of experience scoring the majority of the assessments. At the first two institutes a much smaller number of assessments were scored by raters with less experience (since they rated far fewer assessments the previous 2 years than their Madison counterparts). The high overall interrater agreement could be attributed to a high quality scoring procedure, the increasing experience of presenters and raters over time, the opportunity for manipulation of scores by the second rater when s/he compared scores, and the general movement from rating harder-to-score PSAs to easier-to-score EAs.

Interrater Reliability by Rater

Interrater reliability was calculated for all raters at each institute (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 306 of the 370 student responses or 82.70% of the time. (This includes when Rater A was the second rater.)

Interrater agreement was high for all raters, ranging from a low of 72.09% (Rater L) at the second institute to a high of 95.28% (Rater AO) at the fourth institute. Five out of the 47 raters had less than 80% agreement with the second rater. Almost two-thirds 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.

There was considerable variation in the number of assessments each rater scored due to the length of the scoring session in which the rater participated, the number of assessments prepared for scoring, and, especially, the speed at which the rater scored assessments.

Table 5

Interrater Reliabiltiy by Rater

Institute (Student Responses Rated)	Rater	Student Responses Rated (N)	Agreement %	Single Adjudication %	Multiple Adjudication %
1 (3221)	A	370	82.70%	15.95%	1.35%
	B	333	88.89%	9.61%	1.50%
	C	98	78.57%	20.41%	1.02%
	D	91	78.02%	20.88%	1.10%
	E	335	86.57%	11.34%	2.09%
	F	321	82.55%	17.13%	0.31%
	G	328	88.11%	11.59%	0.30%
	H	421	89.07%	9.74%	1.19%
	I	284	90.85%	8.45%	0.70%
	J	382	90.58%	8.90%	0.52%
	K	258	87.60%	12.40%	0.00%
2 (84)	L	42	72.09%	25.58%	2.33%
	M	27	77.36%	22.64%	0.00%
	N	15	73.33%	20.00%	6.67%
3 (31686)	O	2159	94.03%	5.84%	0.14%
	P	1713	92.29%	7.53%	0.18%
	Q	1638	93.53%	6.41%	0.06%
	R	3014	92.14%	7.76%	0.10%
	S	1568	90.75%	8.86%	0.38%
	T	1603	91.52%	8.30%	0.19%
	U	1636	92.97%	7.03%	0.00%
	V	1823	92.92%	6.80%	0.27%
	W	2699	91.81%	7.93%	0.26%
	X	1696	91.04%	8.61%	0.35%
	Y	1516	92.68%	7.06%	0.26%
	Z	1744	91.34%	8.20%	0.46%
	AA	2204	93.74%	6.03%	0.23%
	AB	1609	92.98%	6.84%	0.19%
	AC	1150	91.30%	8.61%	0.09%
	AD	1493	90.62%	9.11%	0.27%
	AE	2421	92.94%	6.90%	0.17%
	AF	178	83.15%	15.73%	1.12%
4 (9341)	AG	729	93.42%	6.31%	0.27%
	AH	690	89.13%	10.87%	0.00%
	AI	873	92.67%	7.22%	0.11%
	AJ	635	91.18%	8.35%	0.47%
	AK	632	93.20%	6.65%	0.16%
	AL	608	87.17%	12.50%	0.33%
	AM	618	90.94%	8.90%	0.16%
	AN	677	93.35%	6.35%	0.30%
	AO	509	95.28%	4.72%	0.00%
	AP	579	94.13%	5.70%	0.17%
	AQ	618	90.45%	9.55%	0.00%
	AR	510	91.57%	8.24%	0.20%
	AS	431	91.42%	8.35%	0.23%
	AT	523	88.15%	11.66%	0.19%
	AU	709	89.99%	9.45%	0.56%

Conclusion

The high interrater agreement at all the 2000 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 3 scoring institutes in 2000 (see Table A1 in the Appendix). The number of assessments varied at each institute depending on number of contexts covered 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 12 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 7

Overall Interrater Reliability

The interrater agreement on the Grade 7 Problem Solving Assessment was high (89.61% see Figure 4 and Appendix B1). 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 87.10% on "The Pentagon" context (Items 8–10) to a high of 91.98% on the "Playground" context (Items 22–26).

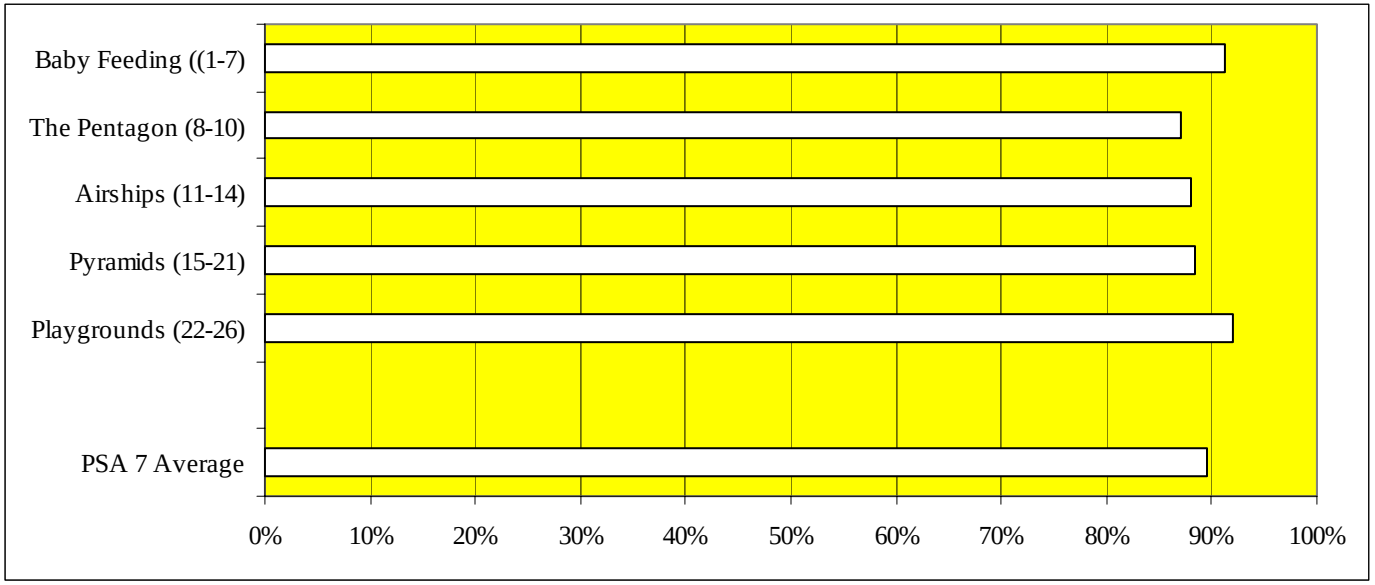

Figure 4. Interrater agreement on Grade 7 Problem Solving Assessment, by context.

The interrater agreement on individual items was over 80% on 23 out of 26 items, and more than half 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 74.29% on Item 12 from the "Airships" context to a high of 98.02% on Item 22 from the "Playground" context. The other items with low interrater agreement were Item 15 from the "Pyramids" context at 77.97% and Item 18 from the "Pyramids" context at 79.94%.

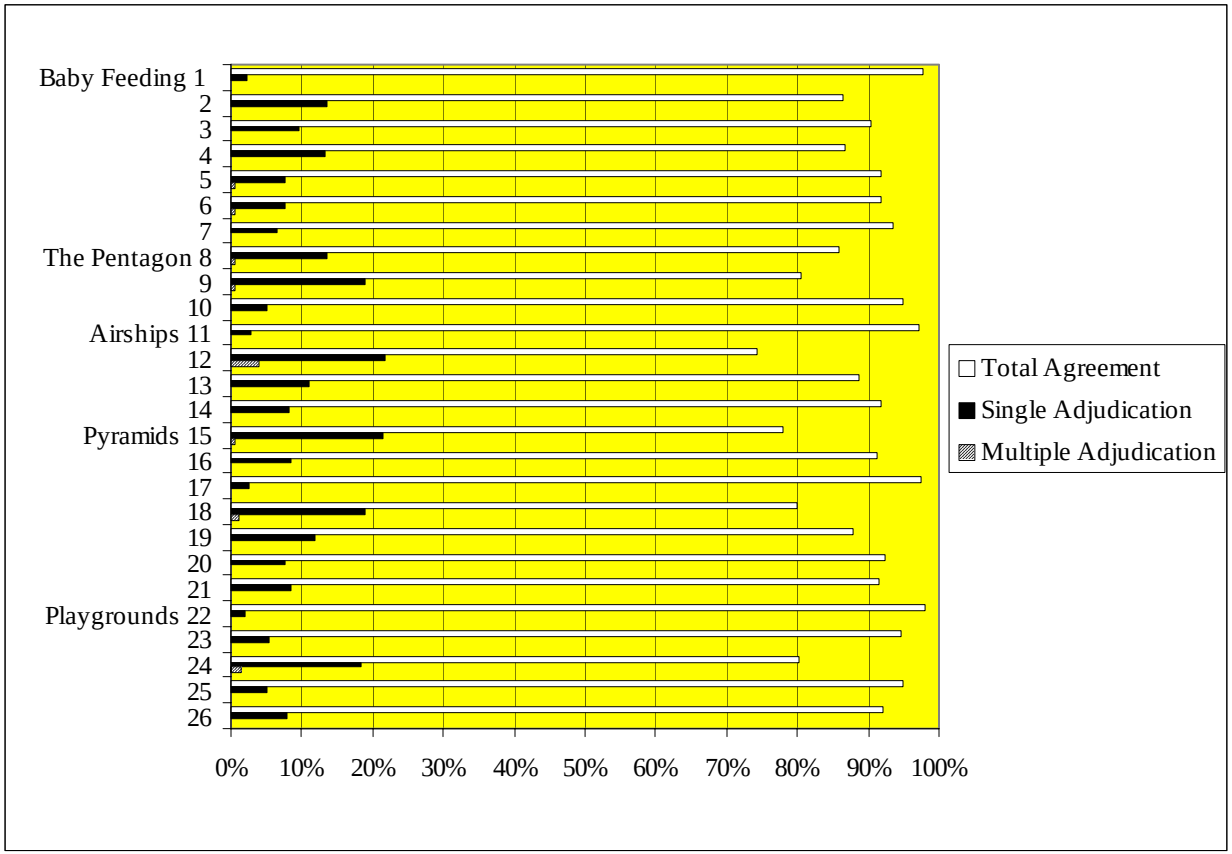

Figure 5. Interrater agreement on Grade 7 Problem Solving Assessment, by item.

The incidence of single adjudication was inversely proportional to the incidence of total agreement (see Figure 5 and Table B1 in the Appendix). The percentage of single adjudication ranged from a low of 1.98% on Item 22 from the "Playground" context to a high of 21.75% on Item 12 from the "Airships" context. The incidence of multiple adjudication was very low. It ranged from 0% on 15 items (Items 1, 2, 3, 4 and 7 from the "Baby Feeding" context, Item 10 from "The Pentagon" context, Items 11 and 14 from the "Airships" context, Item 17, 20, and 21 from the "Pyramids" context, and Items 22, 23, 25, and 26 context) to a high of 3.95% 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) well-defined and clarified rubrics; (c) effective scoring procedures; (d) easy-to-score answers that require no interpretation, especially Item 17 from the "Pyramids" context; (e) lowest level of reasoning required in student response (e.g., Item 1 from the "Baby Feeding" and Item 22 from the "Playgrounds" context); and (f) many items with nonresponses or incorrect responses (Item 11 from the "Airships" context). Factors contributing to the lower interrater agreement (and higher adjudication) include (a) difficulties with the open-ended format; (b) multiple scoring criteria (some raters scored more leniently; Item 9 from "The Pentagon" context, Item 12 from the "Airships" context, Item 15 from the "Pyramids" context, and Item 24 from the "Playground" context); (c) higher levels of reasoning elicited (Item 9 from "The Pentagon" context); (d) answers dependent upon answers to other items (Item 12 from the "Airships" context[4], Item 18 from the "Pyramids" context, and Item 24 from the "Playground" context); (e) subtleties in graphs/figures which students may not have marked clearly (Item 12 from the "Airships" context and Item 15 from the "Pyramids" context); and (f) the need for raters to interpret students' graphs/drawings, which were occasionally constructed in unexpected ways (Item 12 from the "Airships" context and Item 15 from the "Pyramids" context).

[4] Determination of point scores for Item 12 involved the use of three scoring guidelines. Along with using an item-specific rubric in scoring Item 12, raters needed to consider the student's response to Item 11 and were to give full credit to Item 12 if a specific correct student response to Item 13 was given.

Interrater Reliability by Districts

District 1. In District 1, the interrater agreement on the Grade 7 Problem Solving Assessment was high (92.42%; see Figure 6 and Table B2 in the Appendix). Interrater agreement was over 80% on all of the contexts and over 90% on four out of the five contexts. The interrater agreement ranged from a low of 89.00% on the "Airships" context (Items 11-14) to a high of 96.00% on the "Playground" context (Items 22–26).

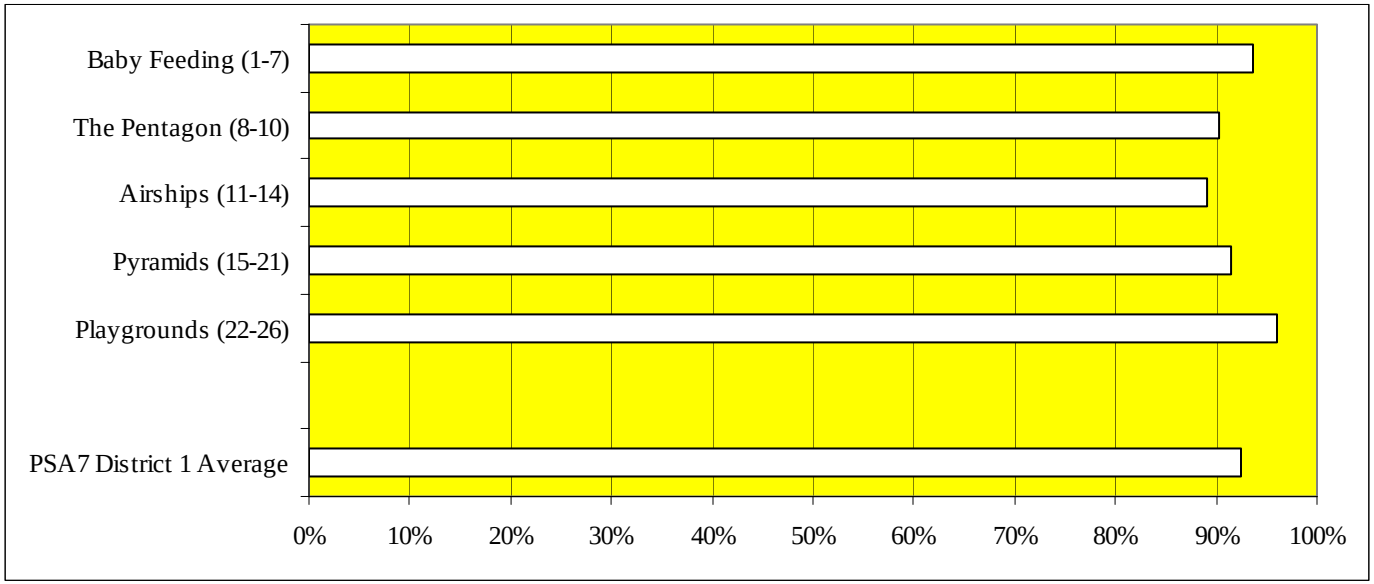

Figure 6. District 1 interrater agreement on Grade 7 Problem Solving Assessment, by context.

On 24 of the 26 individual items interrater agreement was over 80% and 21 of the 26 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 73.00% on Item 12 from the "Airships" to a high of 99.00% on 2 items (Item 11 from the "Airships" context and Item 17 from the "Pyramids" context). The other item with low interrater agreement was Item 15 from the "Pyramids" context at 76.00%.

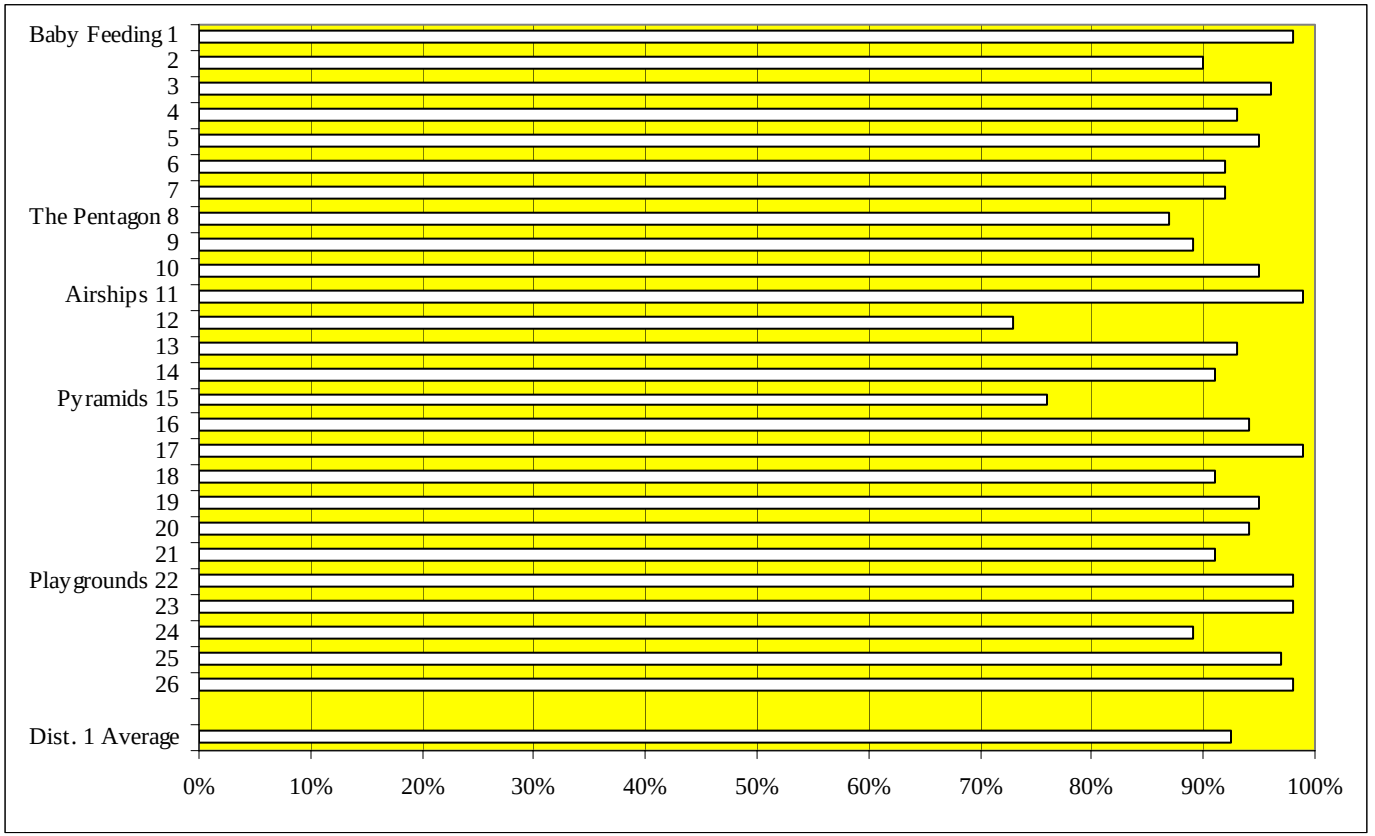

Figure 7. District 1 interrater agreement on Grade 7 Problem Solving Assessment, by item.

District 2. In District 2, the interrater agreement on the Grade 7 Problem Solving Assessment was high (86.05%; see Figure 8 and Table B2 in the Appendix). Interrater agreement was over 80% on all but one context, and it was over 90% on a one context. The interrater agreement ranged from a low of 78.74% on the "Pyramids" context (Item 15–21) to a high of 91.20% on the "Baby Feeding" context (Items 1–7).

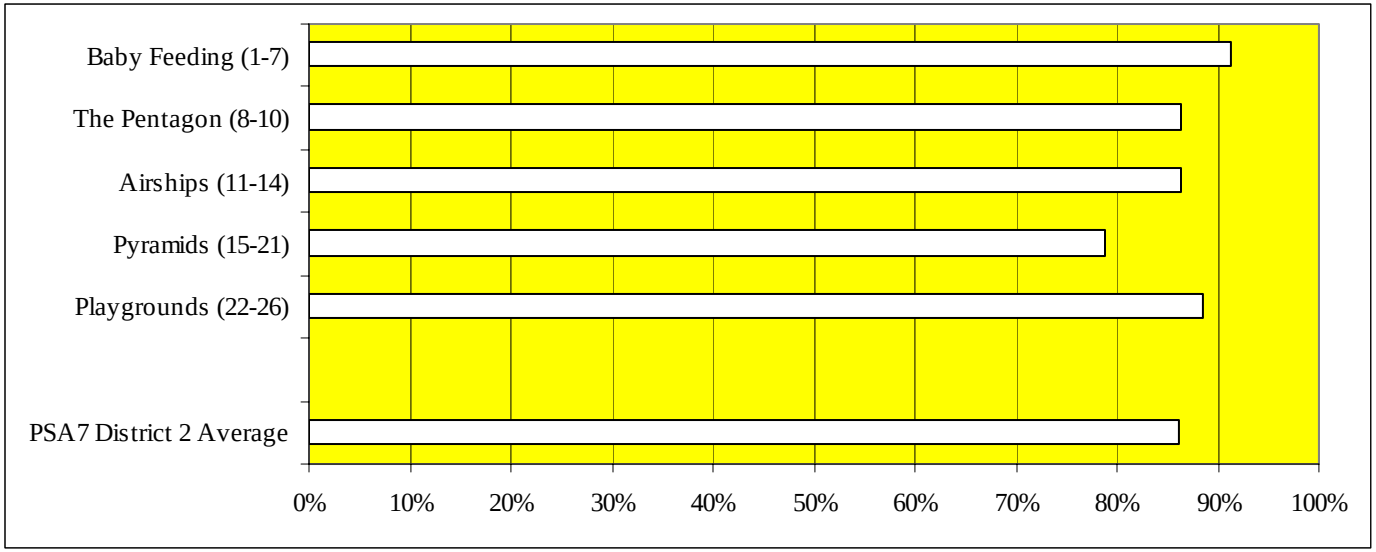

Figure 8. District 2 interrater agreement on Grade 7 Problem Solving Assessment, by context.

All but four of the individual items had interrater agreement over 80%, and more than a third 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 63.95% on Item 18 from the "Pyramids" to a high of 97.67% on 2 items (Item 1 from the "Baby Feeding" context and Item Item 11 from the "Airships" context). The other items with low agreement were Items 19 and 15 from the "Pyramids" context at 65.12% and 68.60% respectively, and Item 12 from the "Airships" context at 67.44%.

Figure 9. 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 (90.01%; see Figure 10 and Table B2 in the Appendix). Interrater agreement was over 80% on all contexts, and over 90% on three out of the five contexts. The interrater agreement ranged from a low of 84.88% on "The Pentagon" context (Items 8–10) to a high of 91.96% on the "Playground" context (Items 22–26).

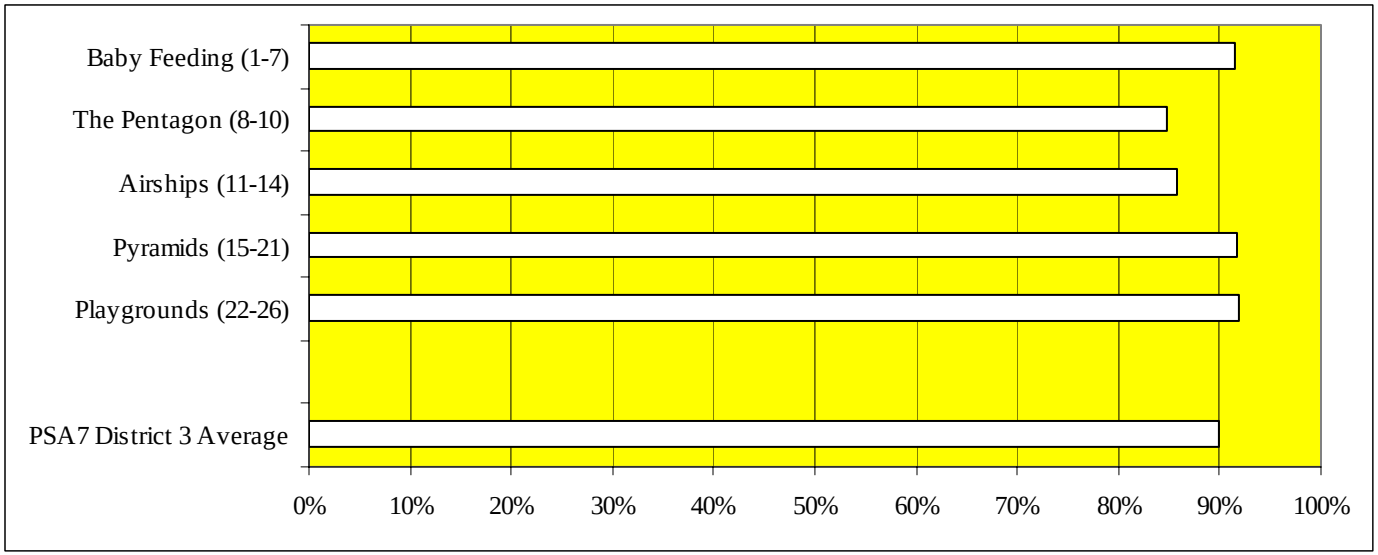

Figure 10. District 3 interrater agreement on Grade 7 Problem Solving Assessment, by context.

Interrater agreement was over 80% on 23 out of 26 individual items. More than half 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 74.23% on Item 12 from the "Airships" to a high of 100% on Item 22 from "The Pentagon" context. Other individual items with low interrater agreement were Item 9 from "The Pentagon" context at 75.26% and Item 24 from the "Playgrounds" context at 77.32%.

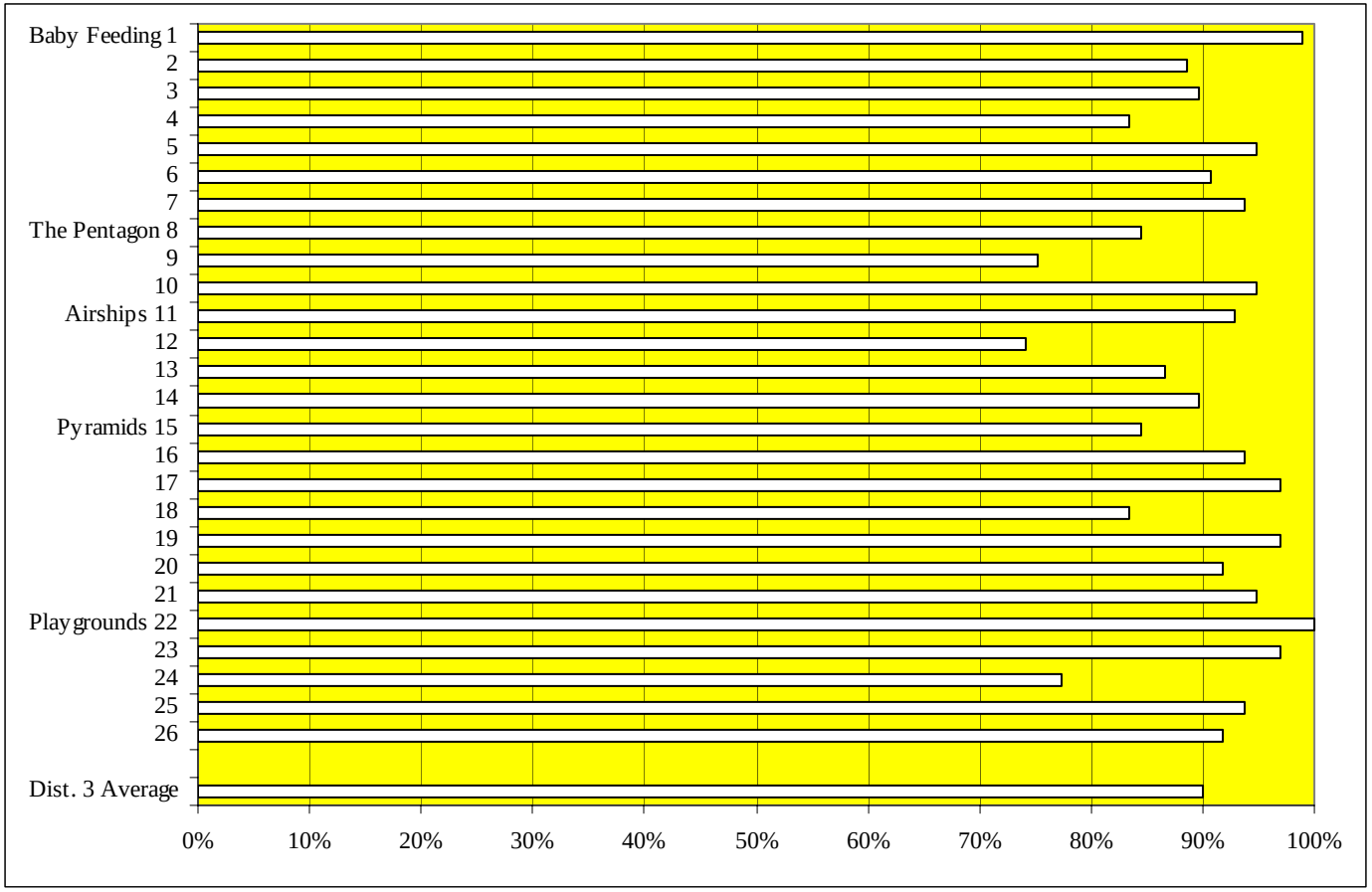

Figure 11. District 3 interrater agreement on Grade 7 Problem Solving Assessment, by item.

District 4. In District 4, the interrater agreement on the Grade 7 Problem Solving Assessment was high (93.03%; see Figure 12 and Table B2 in the Appendix). Interrater agreement was over 80% on all contexts, and it was over 90% on three out of the five contexts. The interrater agreement ranged from a low of 86.38% on "The Pentagon" context (Item 8–10) to a high of 91.55% on the "Airships" context (Items 11–14).

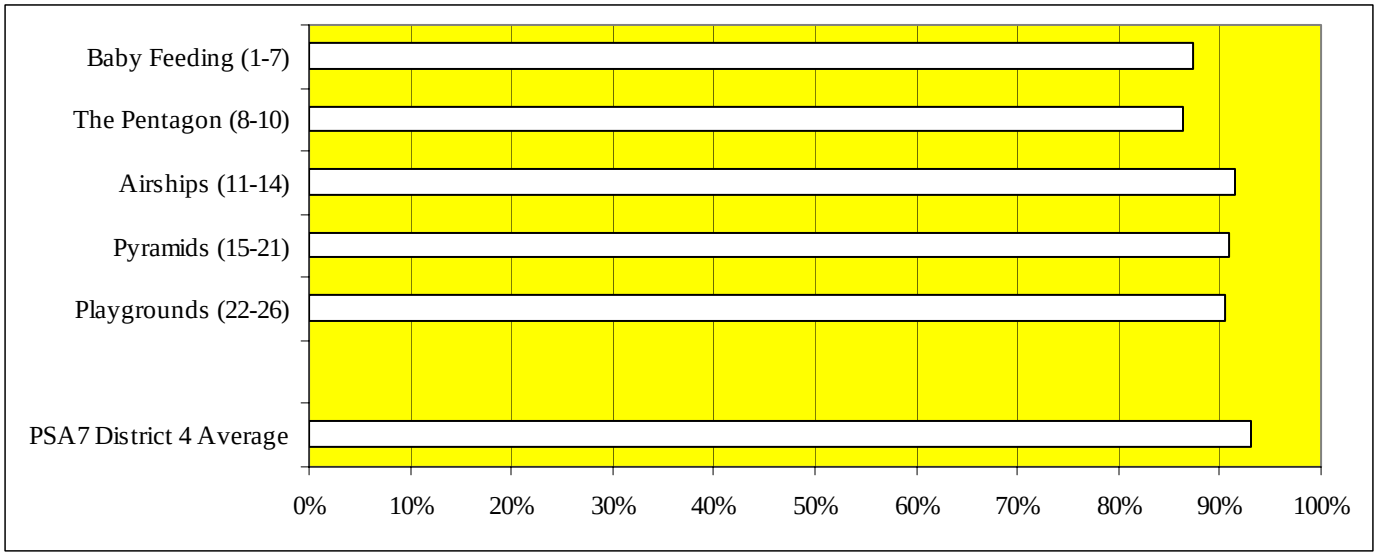

Figure 12. District 4 interrater agreement on Grade 7 Problem Solving Assessment, by context.

All but four of the individual items had interrater agreement over 80%, and more than half 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 71.83% on Item 24 from the "Playgrounds" context to a high of 100% on Item 11 from the "Airships" context. The other individual items with low interrater agreement were Item 9 from "The Pentagon" context at 73.24%, Item 2 from the "Baby Feeding" context at 77.46%, and Item 18 from the "Pyramids" context at 78.87%.

Figure 13. 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 14 and Table B2 in the Appendix). In District 1, interrater agreement was higher than other districts on the "Baby Feeding," "The Pentagon," and "Playground" contexts. In District 2, interrater agreement was lower than other districts overall, especially on the "Pyramids," and "Playground" contexts. In District 3, interrater agreement was lower than other districts on "The Pentagon" context. In District 4, interrater agreement was lower than the other districts on the "Baby Feeding" context, but higher on the "Airships" context.

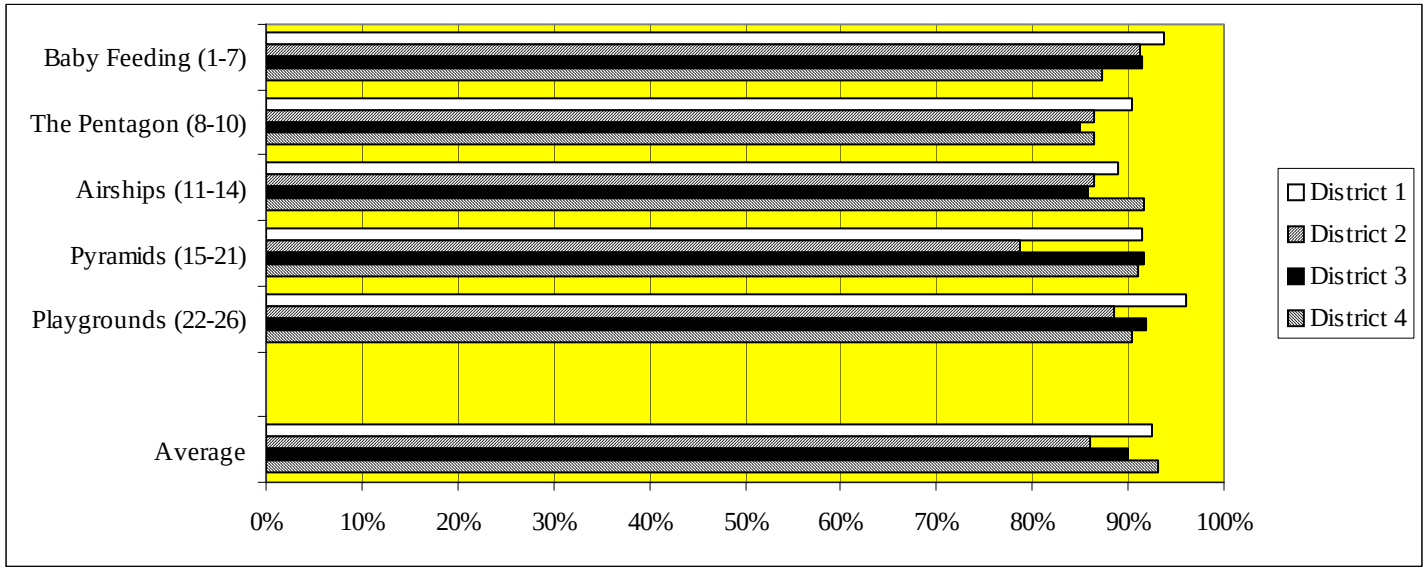

Figure 14. Across district interrater agreement on Grade 7 Problem Solving Assessment, by context.

Some individual items from each district have large (5% or greater) differences in interrater agreement (see Table 6 and Table B2 in the Appendix). In District 1, interrater agreement was very high on Item 18 and high on Items 3, 4, 9, 13, 24, and 26. In District 2, interrater agreement was low overall and on Items 8, 13, 20, 21, and 25; very low on Items 12, 15, 16, 18, 19, and 23; but high on Item 14. In District 3, interrater agreement was lower than other districts on Items 9 and 11. In District 4, interrater agreement was much lower than other districts on Item 2 and 24; low on Items 4, 5, and 9; very high on Item 12; and high on Item 8 and 20.

Table 6
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	98.00%	97.67%	98.97%	95.77%
	2	90.00%	87.21%	88.66%	***77.46%*** [5]
	3	**96.00%** [6]	86.05%	89.69%	88.73%
	4	**93.00%**	88.37%	83.51%	***80.28%***
	5	95.00%	90.70%	94.85%	***84.51%***
	6	92.00%	93.02%	90.72%	91.55%
	7	92.00%	95.35%	93.81%	92.96%
The Pentagon	8	87.00%	***82.56%***	84.54%	**90.14%**
	9	**89.00%**	82.56%	***75.26%***	***73.24%***
	10	95.00%	94.19%	94.85%	95.77%
Airships	11	99.00%	97.67%	***92.78%***	100.00%
	12	73.00%	***67.44%***	74.23%	**84.51%**
	13	**93.00%**	***84.88%***	86.60%	90.14%
	14	91.00%	**95.35%**	89.69%	91.55%
Pyramids	15	76.00%	***68.60%***	84.54%	83.10%
	16	94.00%	***83.72%***	93.81%	92.96%
	17	99.00%	95.35%	96.91%	98.59%
	18	**91.00%**	***63.95%***	83.51%	78.87%
	19	95.00%	***65.12%***	96.91%	92.96%
	20	94.00%	***88.37%***	91.75%	**95.77%**
	21	91.00%	***86.05%***	94.85%	94.37%
Playgrounds	22	98.00%	95.35%	100.00%	98.59%
	23	98.00%	***87.21%***	96.91%	95.77%
	24	**89.00%**	80.23%	77.32%	***71.83%***
	25	97.00%	***90.70%***	93.81%	98.59%
	26	**98.00%**	89.53%	91.75%	87.32%
Average		92.42%	***86.05%***	90.01%	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 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 2, most of the District 2 assessments were rated by study teachers. Most of the assessments from the other districts were scored by teachers in Madison who had more experience due to the greater number of assessments scored at scoring institutes they attended in the previous two years.)

Interrater Reliability by Program (Conventional or Mathematics in Context)

Conventional curricula. The interrater agreement on the Grade 7 Problem Solving Assessment from conventional curricula was high (90.21%; see Figure 15 and Table B3 in the Appendix). Interrater agreement was over 80% on all of the contexts and over 90% on three-fifths of the contexts. The interrater agreement ranged from a low of 85.28% on the "Pyramids" context (Items 15–21) to a high of 94.37% on the "Playground" context (Items 1–7).

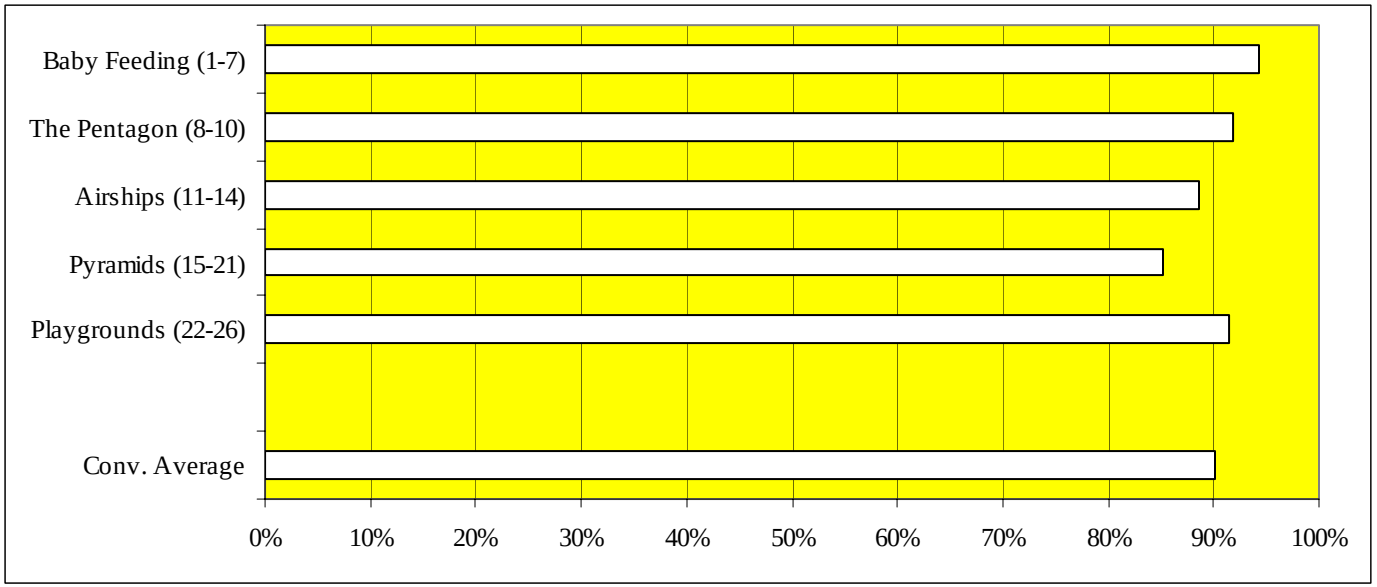

Figure 15. Interrater agreement on Grade 7 Problem Solving Assessment, by context: Conventional curricula.

All but 3 of the individual items had interrater agreement over 80%, and about two-thirds 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 66.67% on Item 18 from the "Pyramids" context to a high of 100% on Item 17 from the "Pyramids" context. The other items with low interrater agreement were Item 12 from the "Airships" context at 72.73% and Item 15 from the "Pyramids" context at 75.76%.

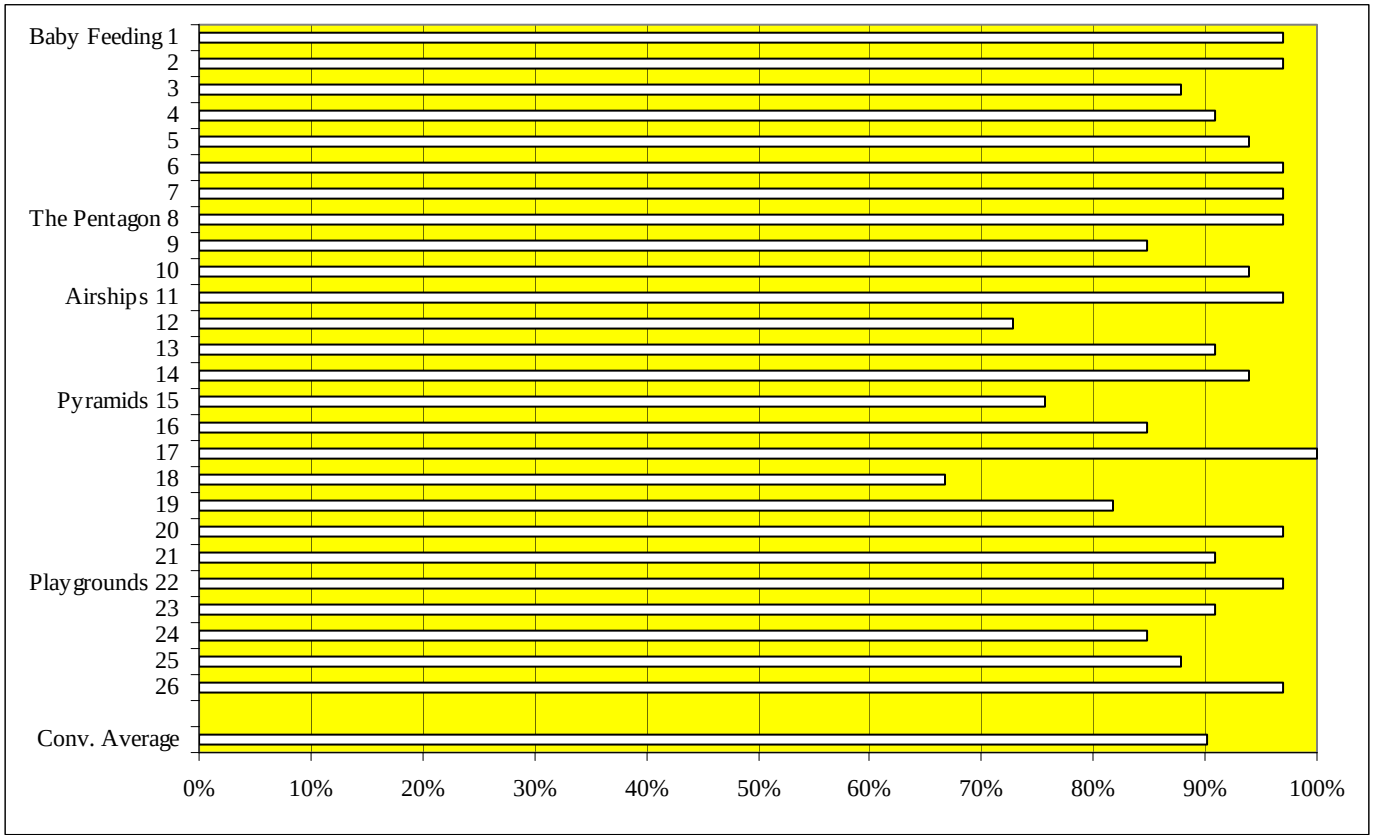

Figure 16. Interrater agreement on Grade 7 Problem Solving Assessment, by item: Conventional curricula.

Mathematic in Context classes. The interrater agreement on the Grade 7 Problem Solving Assessment from *Mathematics in Context* classes was high (89.55%; see Figure 17 and Table B3 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 86.60% on "The Pentagon" context (Items 8–10) to a high of 92.02% on the "Playground" context (Items 22–26).

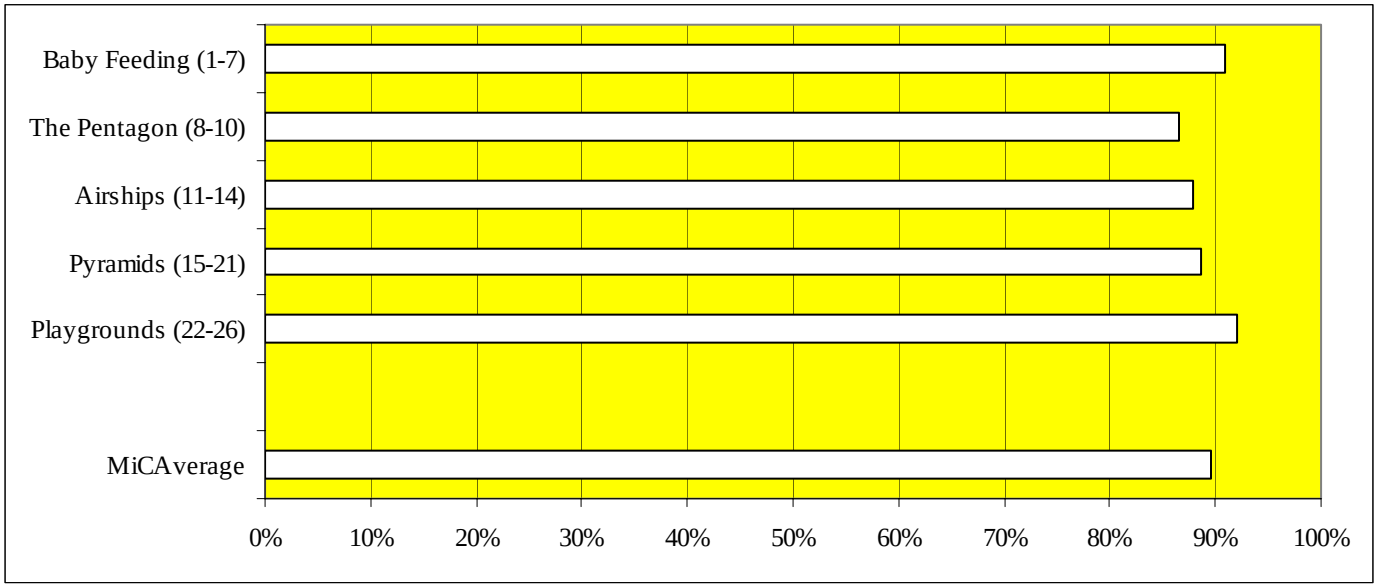

Figure 17. Interrater agreement on Grade 7 Problem Solving Assessment, by context: *Mathematics in Context* classes.

All but 3 of the individual items had interrater agreement over 80%, and about 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 74.45% on Item 12 from the "Airships" context to a high of 98.13% on Item 22 from the "Playgrounds" context. The other items with low interrater agreement were Item 15 from the "Pyramids" context at 78.19%, and Item 24 from the "Playgrounds" context at 79.75%.

Figure 18. 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 is similar (see Figure 19 and Table B3 in the Appendix). The average interrater agreement for conventional curricula was 90.21% and for *Mathematics in Context* classrooms 89.55%. The interrater agreement for "The Pentagon" context (Items 8–10) was higher for the conventional classes.

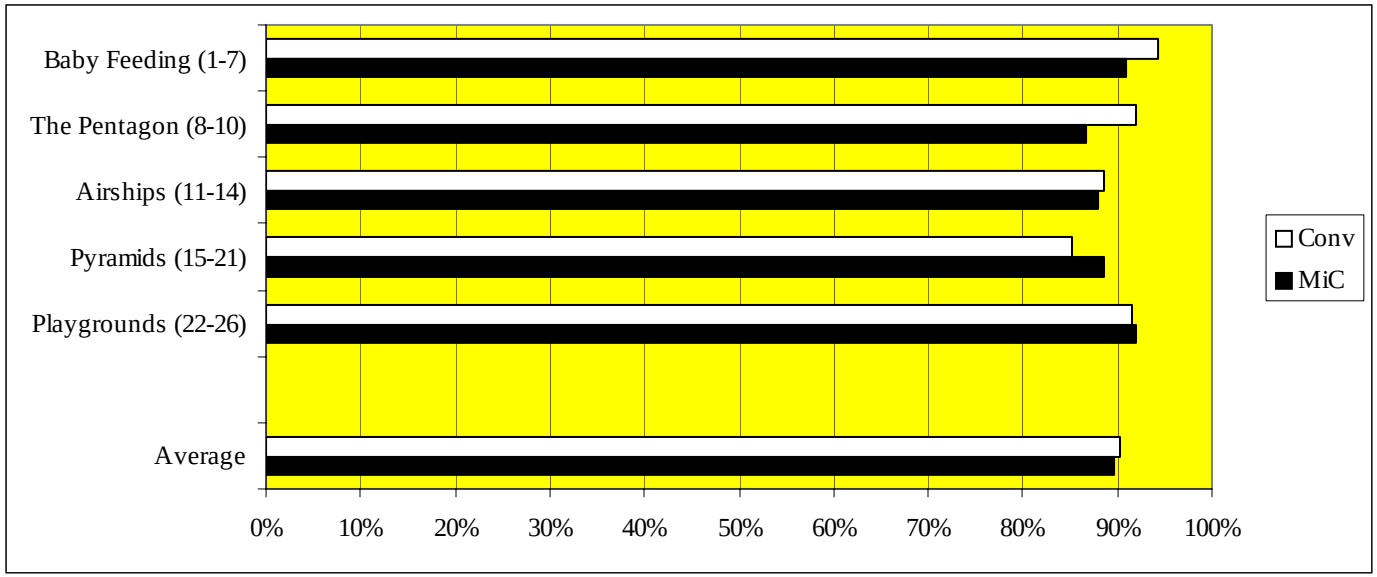

Figure 19. Interrater agreement on Grade 7 Problem Solving Assessment, by context: Conventional curricula and *Mathematics in Context* classes.

The interrater agreement on some individual items also revealed a difference 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 Item 2 from the "Baby Feeding" context, Item 8 from "The Pentagon" context, Item 20 from the "Pyramids" context, and Items 24 and 26 from the "Playgrounds" context.

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.

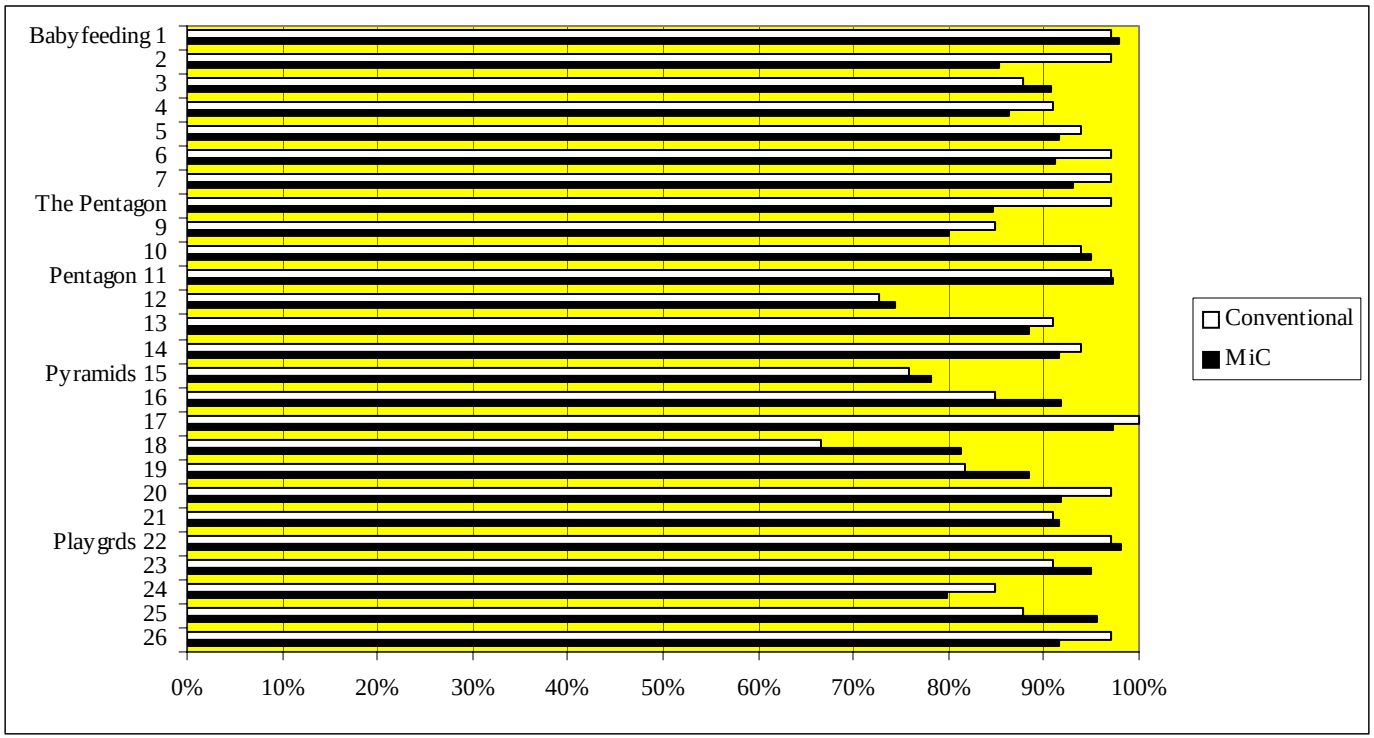

Figure 20. Interrater agreement on Grade 7 Problem Solving Assessment by item: Conventional and *Mathematics in Context* classes.

Grade 8

Overall Interrater Reliability

The interrater agreement on the Grade 8 Problem Solving Assessment was high (93.55% see Figure 21 and Appendix B4). Interrater agreement was over 80% on all of the contexts and over 90% on three-quarters of the contexts. The interrater agreement ranged from a low of 82.80% on the "Club Members" context (Item 1) to a high of 96.30% on the "Key Cards" context (Items 5–7).

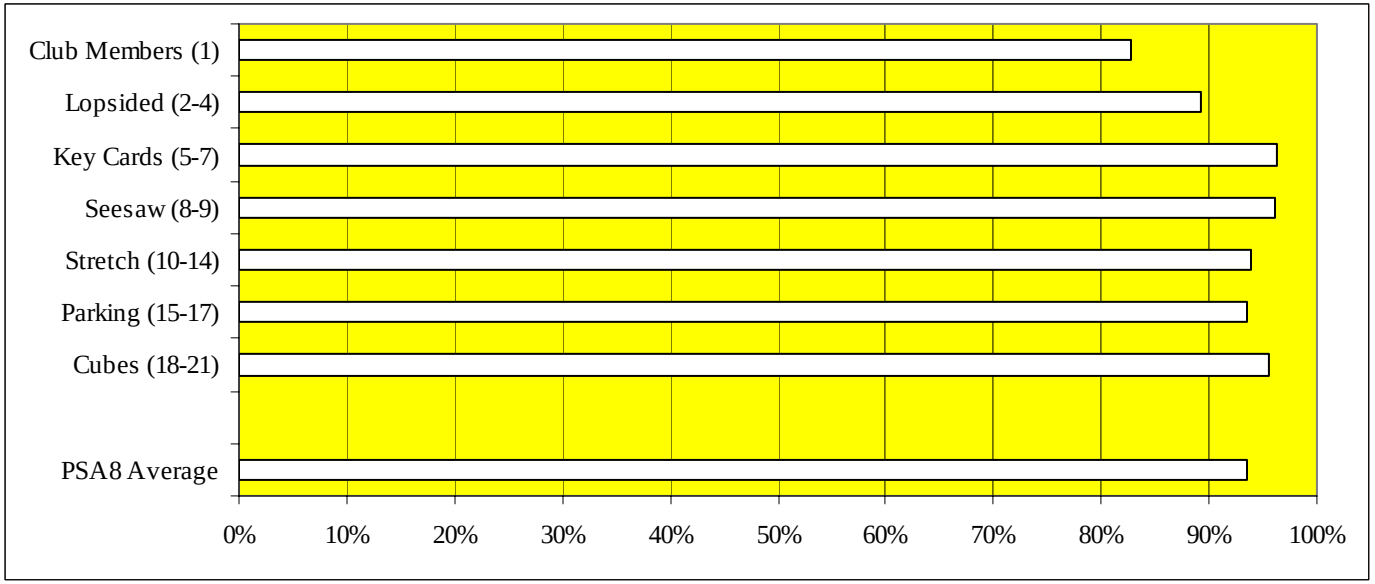

Figure 21. Interrater agreement on Grade 8 Problem Solving Assessment, by context.

All of the individual items had interrater agreement over 80%, and more than three-quarters 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 82.80% on Item 1 from "Club Members" context to a high of 98.15% on Item 12 from the "Stretch" context.

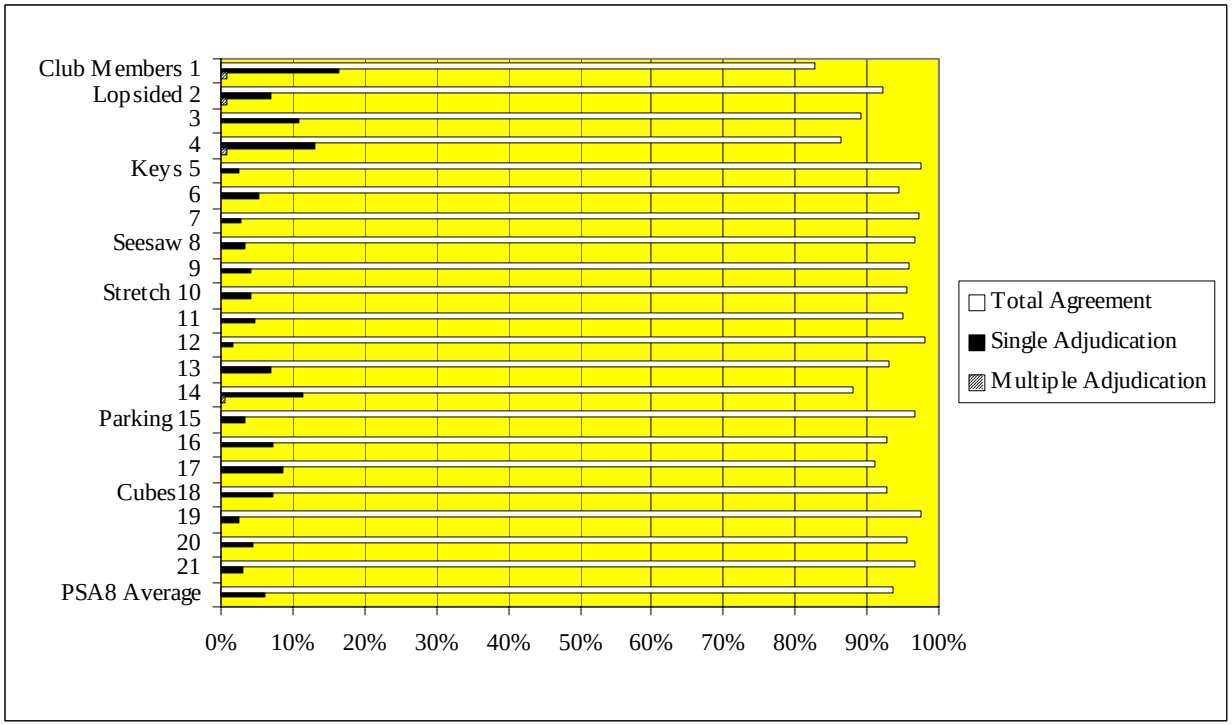

Figure 22. Interrater agreement on Grade 8 Problem Solving Assessment, by item.

The incidence of single adjudication was inversely proportional to the incidence of total agreement (see Figure 22 and Table B4 in the Appendix). The percentage of single adjudication ranged from a low of 1.59% on Item 12 from the "Stretch" context to a high of 16.40% on Item 1 from "Club Members" context. The incidence of multiple adjudication was very low. It ranged from 0% on 11 items (Item 3 from the "Lopsided" context, Items 5 and 7 from the "Key Cards" context, Items 8 and 9 from the "Seesaw" context, Item 13 from the "Stretch" context, Items 15 and 16 from the "Parking" context, and Items 18, 19, and 20 from the "Cubes" context) to a high of 0.79% on three items (Item 1 from the "Club Members" context and Items 2 and 4 from the "Lopsided" 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; and (e) many items with nonresponses or incorrect responses (Items 7 from the "Key Cards" context, Item 9 from the "Seesaw" context, and Item 21 from the "Cubes" context). The factor contributing to the lower interrater agreement (and higher adjudication) was subtleties in graphs/figures which students may not have marked clearly and which some teachers did not recognize in the first round of rating (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 (93.67%; see Figure 23 and Table B5 in the Appendix). Interrater agreement was over 80% on all of the contexts and over 90% on three-quarters of the contexts. The interrater agreement ranged from a low of 84.47% on "Club Members" context (Item 1) to a high of 97.20% on the "Seesaw" context (Items 8–9).

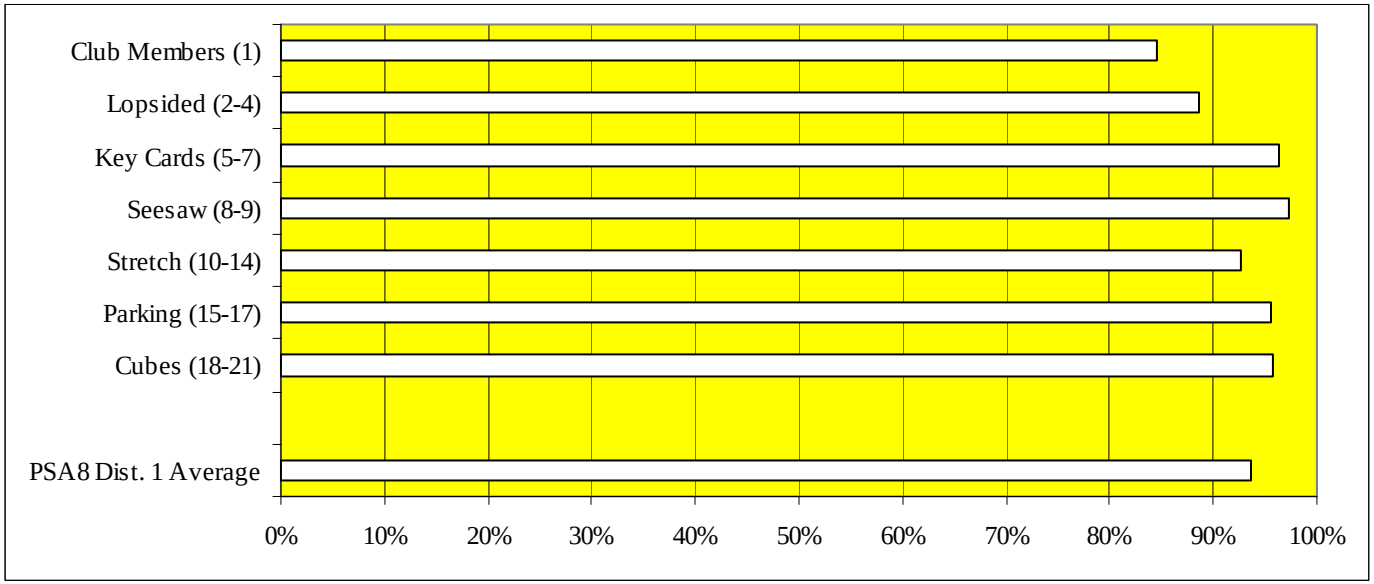

Figure 23. District 1 interrater agreement on Grade 8 Problem Solving Assessment, by context.

All of the individual items had interrater agreement over 80%, and all but five 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 83.23% on Item 14 from "Stretch" to a high of 99.38% on 2 items (Item 12 from the "Stretch" context and Item 15 from the "Parking" context).

Figure 24. District 1 interrater 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 (95.10%; see Figure 25 and Table B5 in the Appendix). Interrater agreement was over 80% on all contexts, and over 90% on six out of the seven of the contexts. The interrater agreement ranged from a low of 85.00% on the "Club Members" context (Item 1) to a high of 97.50% on the "Cubes" context (Items 18–21).

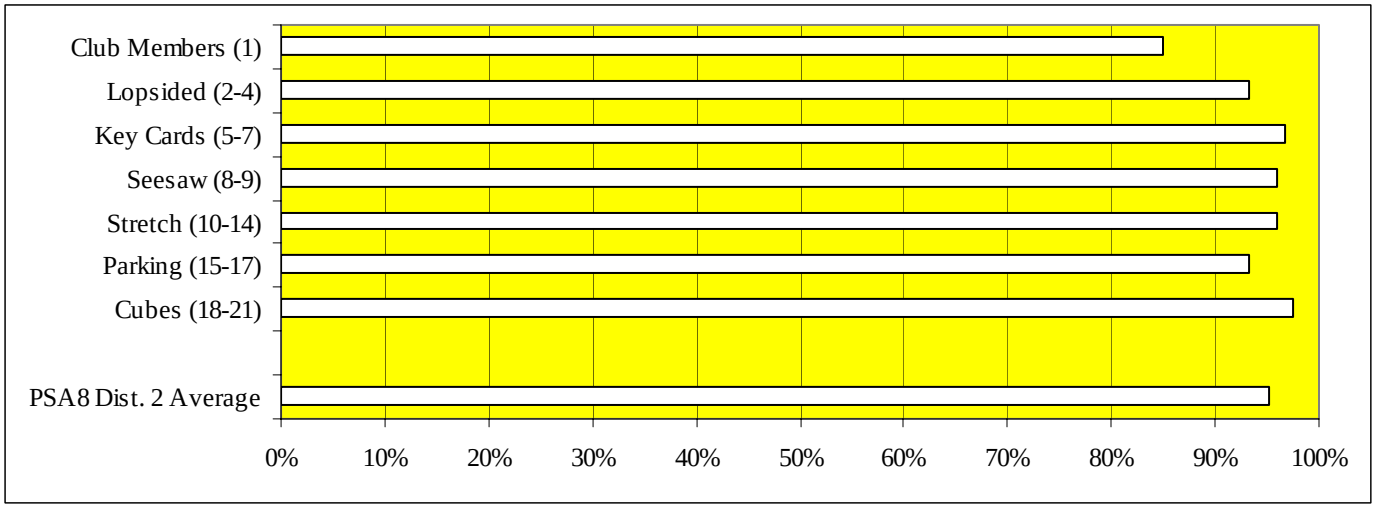

Figure 25. District 2 interrater agreement on Grade 8 Problem Solving Assessment, by context.

All of the individual items had interrater agreement over 80%, and 20 out of 21 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 85.00% on Item 1 from the "Club Members" to a high of 99.00% on 3 items (Item 7 from the "Key Cards" context, Item 11 from the "Stretch" context, and Item 19 from the "Cubes" context).

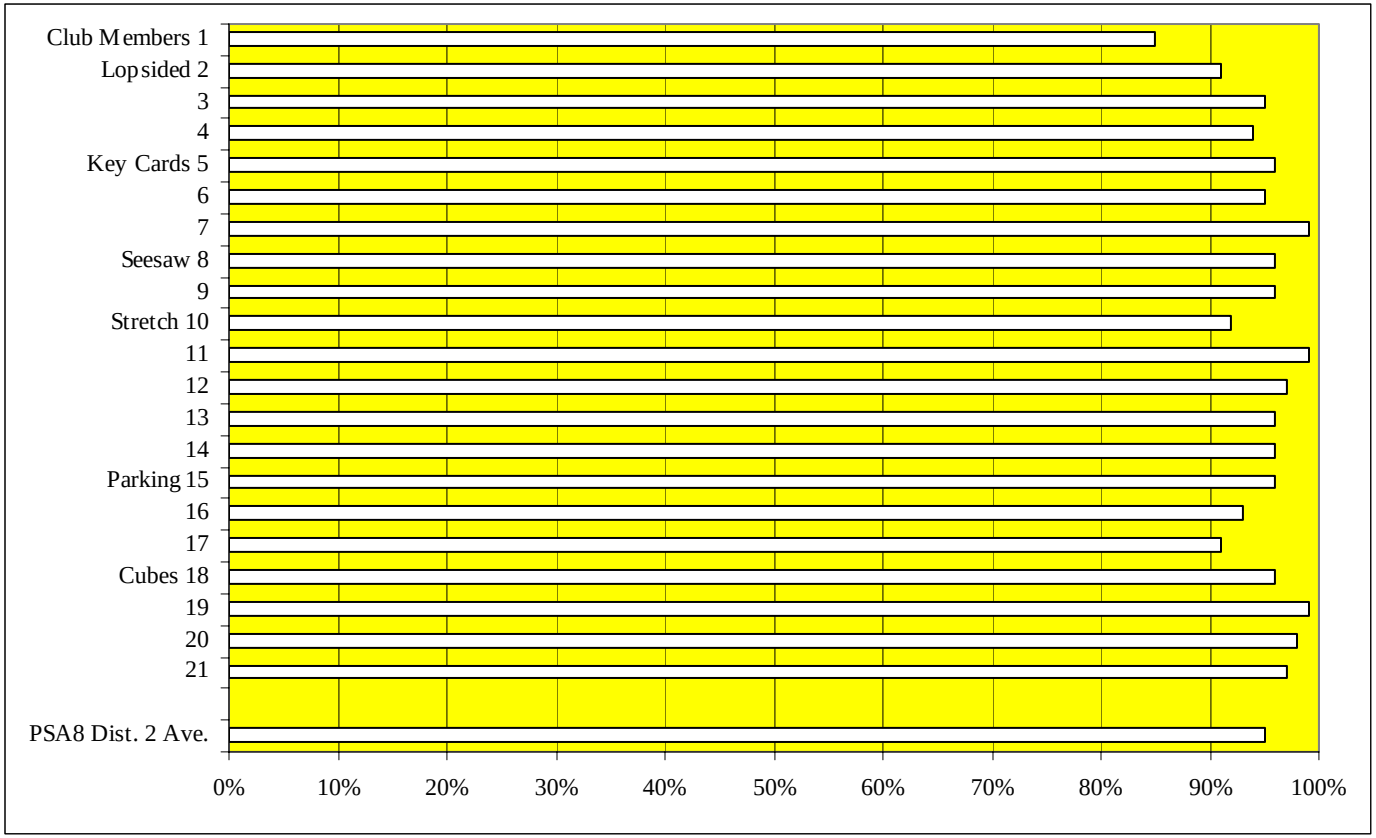

Figure 26. District 2 interrater agreement on Grade 8 Problem Solving Assessment, by item.

District 3. In District 3, the interrater agreement on the Grade 8 Problem Solving Assessment was high (90.21%; see Figure 27 and Table B5 in the Appendix). Interrater agreement was over 80% on all of the contexts and over 90% on four out of the seven contexts. The interrater agreement ranged from a low of 84.28% on the "Lopsided" context (Items 2–4) to a high of 96.86% on "Key Cards" context (Items 8–9).

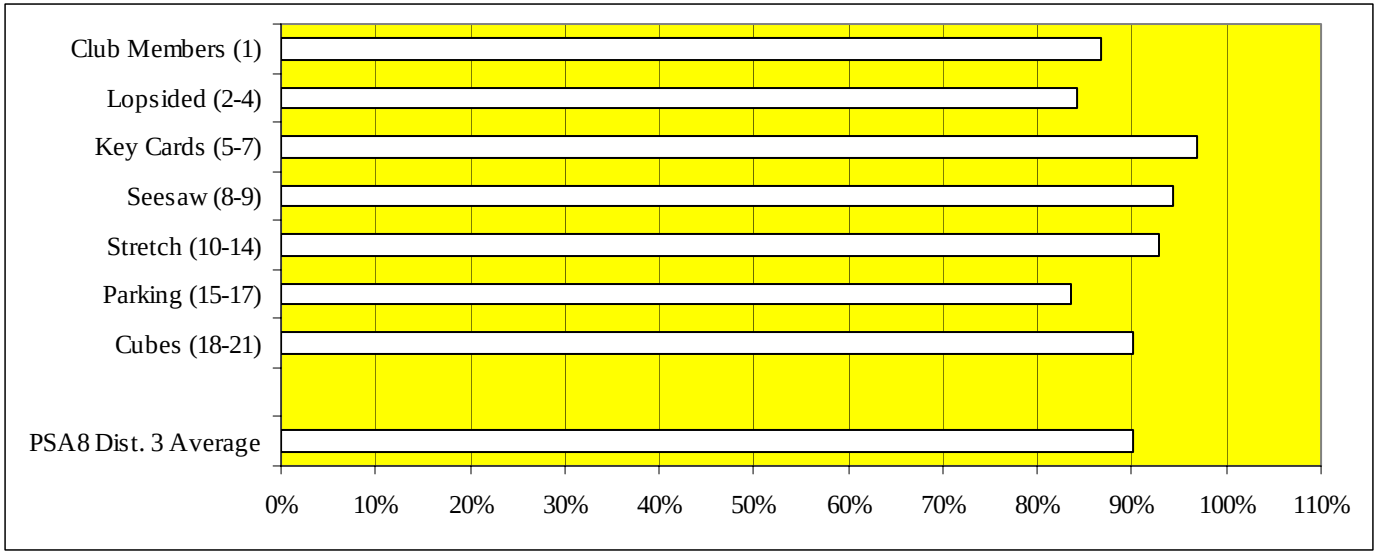

Figure 27. District 3 interrater agreement on Grade 8 Problem Solving Assessment, by context.

All but three of the individual items had interrater agreement over 80%, and more than 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 75.47% on Item 17 from the "Parking" context to a high of 100% on Item 10 from the "Stretch" context. The other two items with low interrater agreement were Items 4 and 3 from the "Lopsided" context at 77.36% and 79.25% respectively.

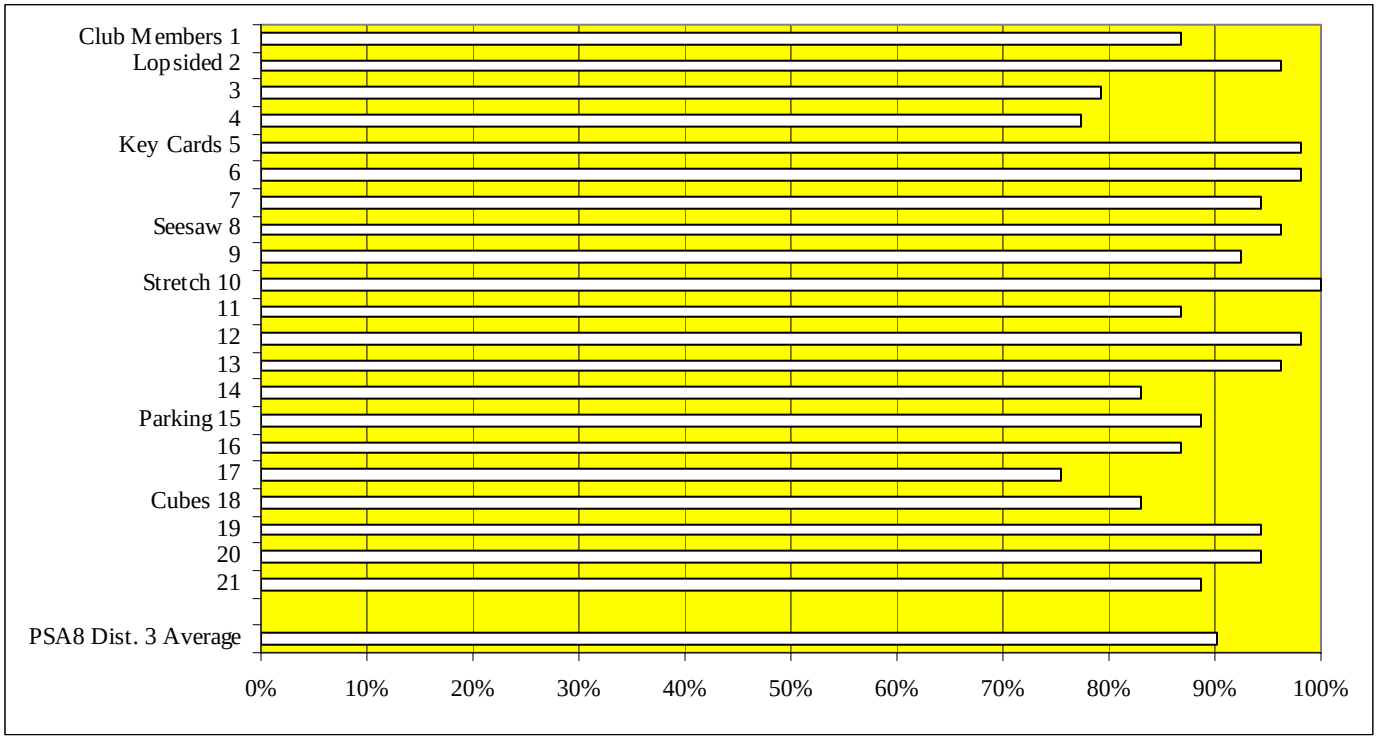

Figure 28. District 3 interrater agreement on Grade 8 Problem Solving Assessment, by item.

District 4. In District 4, the interrater agreement on the Grade 8 Problem Solving Assessment was high (93.60%; see Figure 29 and Table B5 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 71.88% on the "Club Members" context (Item 1) to a high of 96.88% on the "Cubes" context (Items 18–21).

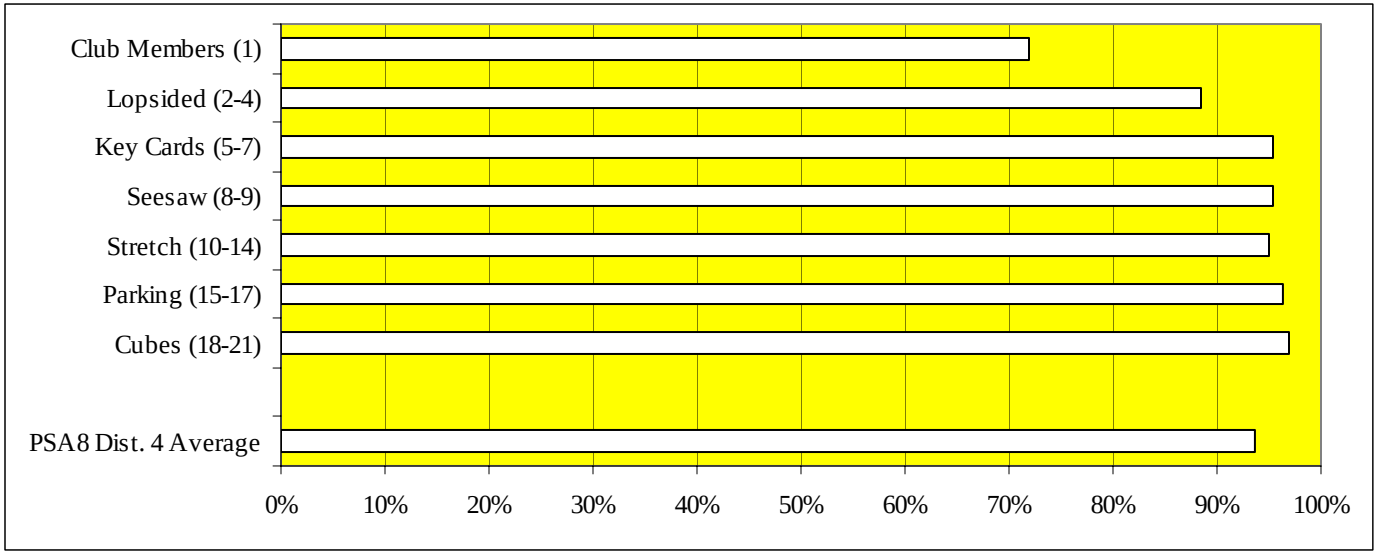

Figure 29. District 4 interrater agreement on Grade 8 Problem Solving Assessment, by context.

All but two of the individual items had interrater agreement over 80%, and 19 of the 21 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 71.88% on Item 1 from the "Club Members" to a high of 98.44% on 2 items (Items 19 and 21 from the "Cubes" context). The other item with low interrater agreement was Item 4 from the "Lopsided" context at 76.56%.

Figure 30. 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 31 and Table B5 in the Appendix). In District 1, interrater agreement was similar or average compared to the other districts. In District 2, interrater agreement was higher than other districts on the "Lopsided" context. In District 3, interrater agreement was much lower than the other districts (5% or greater) on the "Parking" context, lower on the "Lopsided" context, but higher on the "Club Members" context. In District 4, interrater agreement was much lower than the other districts (5% or greater) on the "Club Members" context.

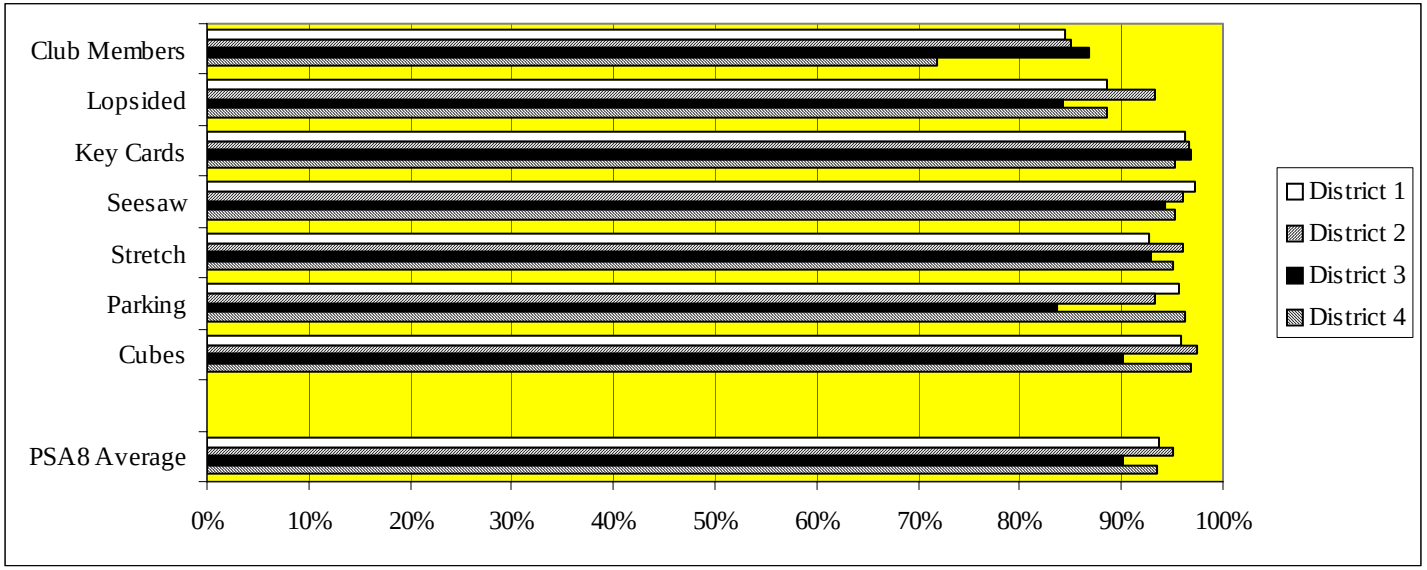

Figure 31. Across district interrater agreement on Grade 8 Problem Solving Assessment, by context.

Some individual items from each district have large (5% or greater) differences in interrater agreement (see Table 7 and Table B5 in the Appendix). In District 1, interrater agreement was lower than the other districts on Item 13 from the "Stretch" context and low on Item 14 from the "Stretch" context. In District 2, interrater agreement was lower than the other districts on Item 10 from the "Stretch" context, but higher on Item 4 from the "Lopsided" context, Item 7 from the "Key Cards" context, and Item 14 from the "Stretch" context. In District 3, interrater agreement was much lower than other districts on Items 3 and 4 from the "Lopsided" context, Item 11 from the "Stretch" context, all items (Items 15, 16 and especially 17) on the "Parking " context, and Items 18 and 21 from the "Cubes" context; low on Item 14 from the "Stretch" context; but higher than the other districts on Item 2 from the "Lopsided" context and Item 10 from the "Stretch" 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 and low on Item 4 from the "Lopsided" 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
Club Members	1	84.47%	85.00%	86.79%	***71.88%*** [7]
Lopsided	2	91.93%	91.00%	**96.23%** [8]	92.19%
	3	85.71%	95.00%	***79.25%***	96.88%
	4	88.20%	**94.00%**	***77.36%***	***76.56%***
Key Cards	5	98.76%	96.00%	98.11%	95.31%
	6	93.17%	95.00%	98.11%	93.75%
	7	96.89%	**99.00%**	94.34%	96.88%
Seesaw	8	97.52%	96.00%	96.23%	95.31%
	9	96.89%	96.00%	92.45%	95.31%
Stretch	10	96.27%	***92.00%***	**100.00%**	95.31%
	11	95.03%	99.00%	***86.79%***	95.31%
	12	99.38%	97.00%	98.11%	96.88%
	13	***89.44%***	96.00%	96.23%	95.31%
	14	***83.23%***	**96.00%**	***83.02%***	92.19%
Parking	15	99.38%	96.00%	***88.68%***	96.88%
	16	93.79%	93.00%	***86.79%***	95.31%
	17	93.79%	91.00%	***75.47%***	96.88%
Cubes	18	92.55%	96.00%	***83.02%***	96.88%
	19	97.52%	99.00%	94.34%	98.44%
	20	95.03%	98.00%	94.34%	93.75%
	21	98.14%	97.00%	***88.68%***	98.44%
Average		93.82%	92.55%	89.74%	93.82%

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

[8] 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 and blank answers.

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 very high (95.71%; see Figure 32 and Table B6 in the Appendix). Interrater agreement was over 80% on all of the contexts and over 90% on five out of the seven contexts. The interrater agreement ranged from a low of 85.61% on the "Lopsided" context (Items 2–4) to a high of 98.48% on the "Key Cards" context (Item 5–7).

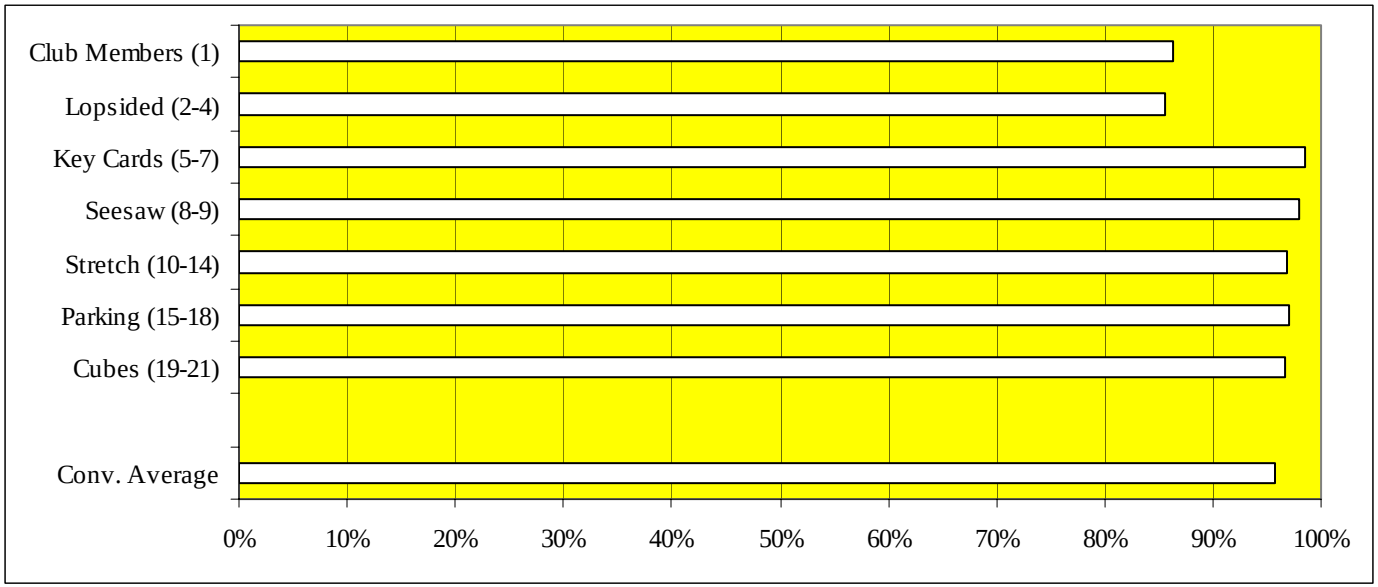

Figure 32. Interrater agreement on Grade 8 Problem Solving Assessment, by context: Conventional curricula.

All but one of the individual items had interrater agreement over 80%, and 18 out of 21 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 79.55% on Item 3 from the "Lopsided" context to a high of 100% on 5 items (Item 5 from the "Key Cards" context, Item 11 from the "Seesaw" context, Item 15 from the "Stretch" context, Item 17 from the "Parking" context, and Item 21 from the "Cubes" context).

Figure 33. 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 very high (93.96%; see Figure 34 and Table B6 in the Appendix). 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 82.34% on the "Club Members" context (Item 1) to a high of 96.03% on the "Seesaw" context (Items 8–9).

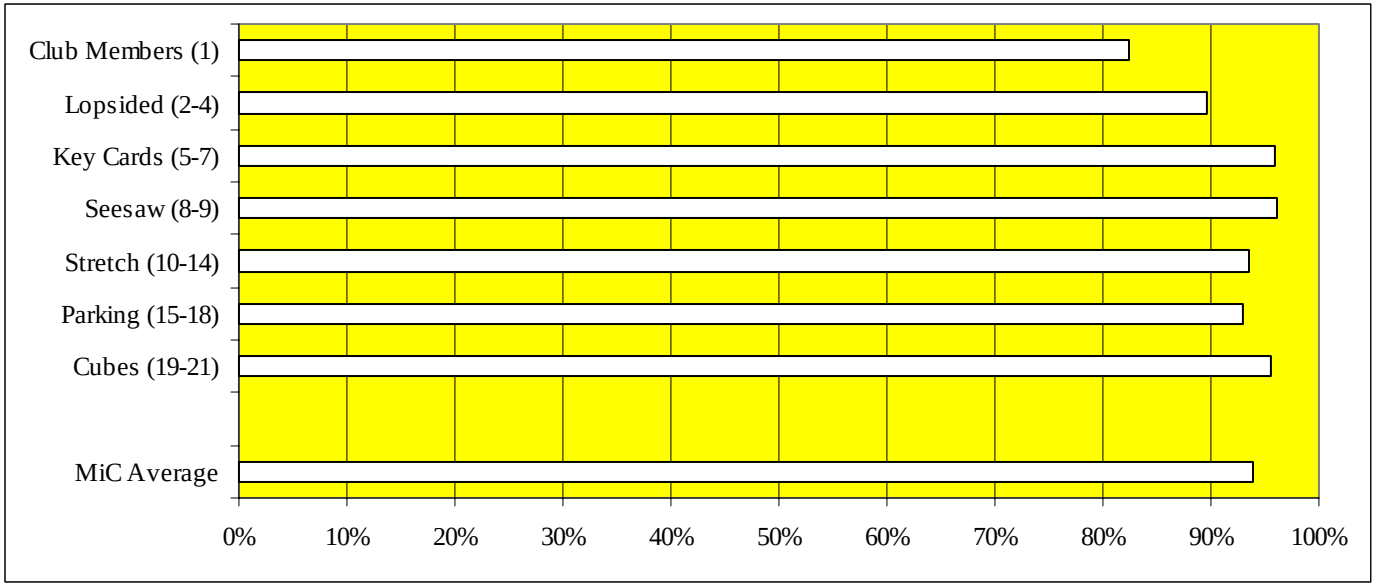

Figure 34. Interrater agreement on Grade 8 Problem Solving Assessment, by context: *Mathematics in Context* classes.

All of the individual items had interrater agreement over 80%, and over 90% on 18 out of 21 items (see Figure 35 and Table B6 in the Appendix). The interrater agreement on individual items ranged from a low of 82.34% on Item 1 from the "Club Members" context to a high of 98.20% on Item 12 from the "Stretch" context.

Figure 35. 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 36 and Table B6 in the Appendix). The average interrater agreement for conventional curricula was 95.71% and 93.96% for *Mathematics in Context* classes. The difference in interrater agreement was never greater than 5%.

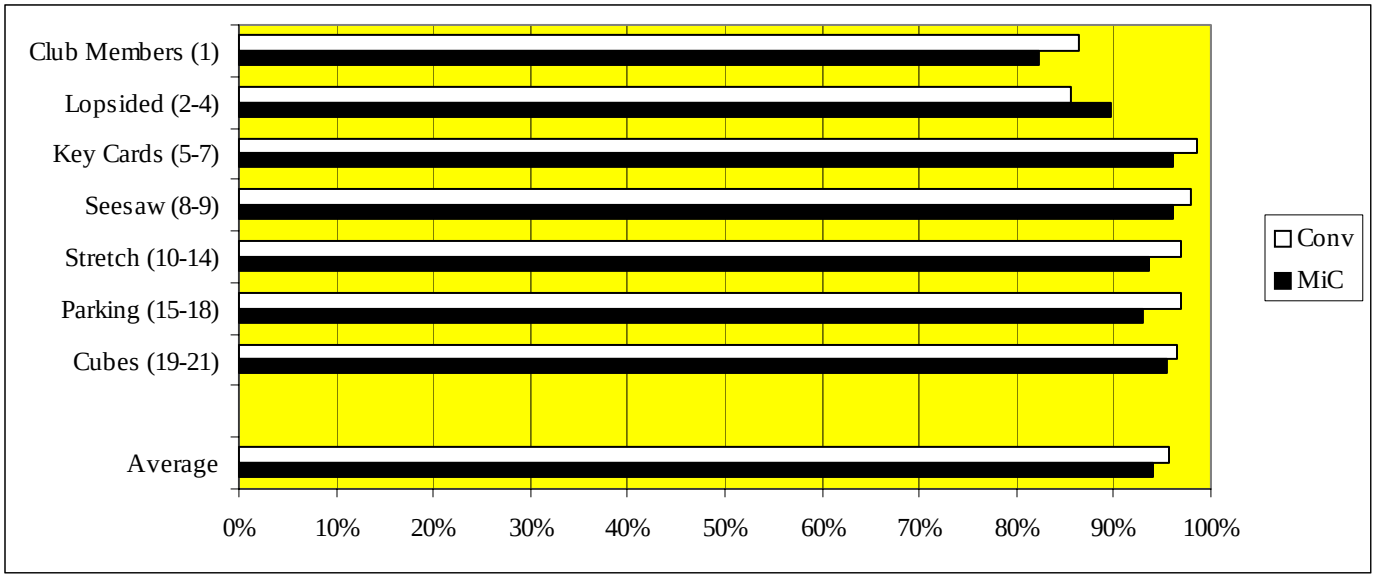

Figure 36. Interrater agreement on Grade 8 Problem Solving Assessment, by context: Conventional curricula and *Mathematics in Context* classes.

The interrater agreement on some individual items also revealed a difference between conventional curricula and *Mathematics in Context* classes (see Figure 37 and Table B6 in the Appendix). Assessments from conventional curricula had higher agreement (5% or greater) on Items 11, 13, and 14 from the "Stretch" context and Item 17 from the "Parking" context. Assessments from the *Mathematics in Context* classes had higher agreement on Items 3 and 4 from the "Lopsided" context.

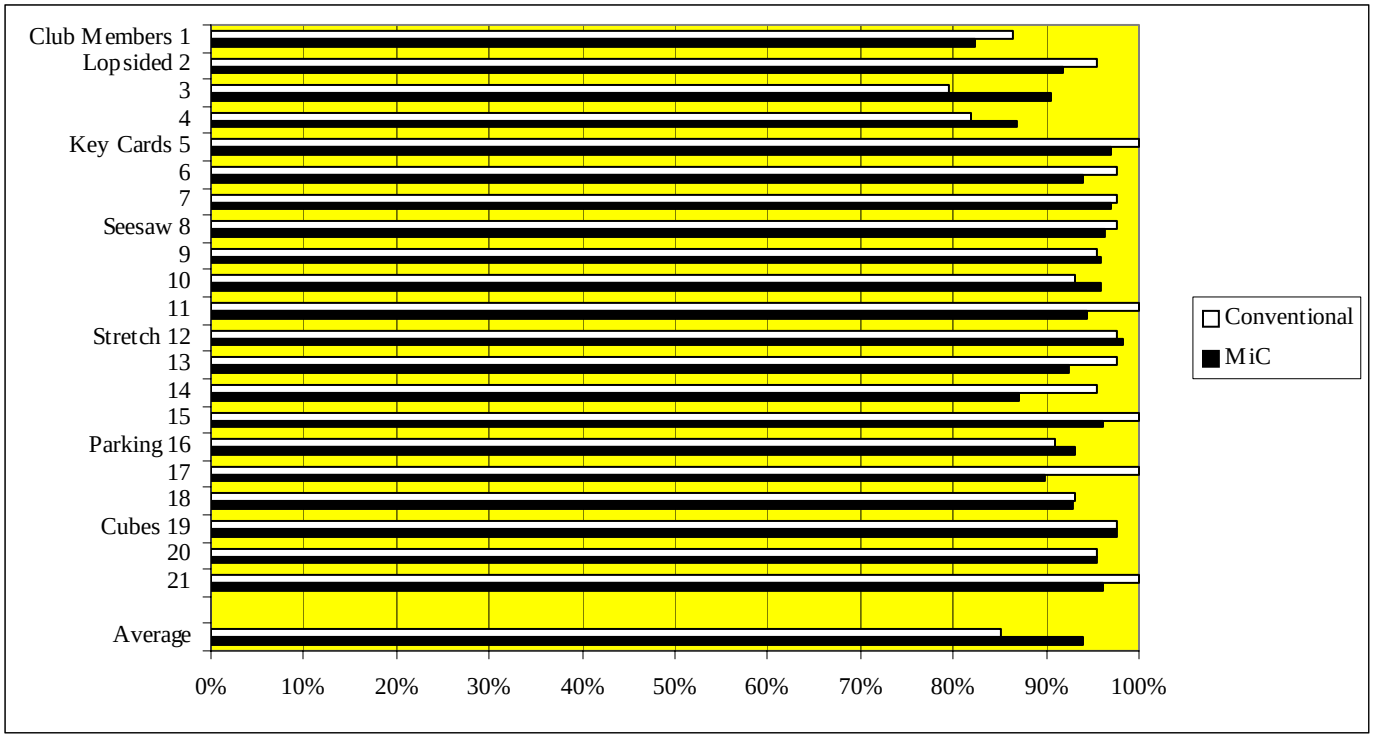

Figure 37. Interrater agreement on Grade 8 Problem Solving Assessment by item: Conventional and *Mathematics in Context* classes.

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.

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) well-defined and clarified rubrics; (c) effective scoring procedures; (d) items in which lower level of reasoning was elicited; and (e) proportion of student nonresponse.

The factor contributing to the lower interrater agreement (and higher adjudication) was subtleties in graphs/figures, which students might not have marked clearly and which some teachers did not recognize in the first round of rating.

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

The differences in interrater agreement across programs (conventional curricula or Mathematics in Context classes) were most likely due to differences in (a) content study teachers taught; (b) raters' interpretation of student work; and (c) proportion of student nonresponse and blank answers.
.

Interrater Reliability on External Assessments

All of the 2000 External Assessments were scored at two scoring institutes in the summer, 2000 (see Table A1 in the Appendix). In contrast to the Problem Solving Assessment, seven EA constructed-response items (anchor items) were repeated on the grade-specific assessments. Two other constructed-response items that appeared on one or two of the grade-specific assessments were also scored. For purposes of scoring, each set of constructed-response 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 many 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 7

Overall Interrater Reliability

The interrater agreement on the Grade 7 External Assessment was very high (92.26%; see Figure 38 and Appendix C1.) Interrater agreement was over 80% on seven out of the eight items.[9] All but two items had interrater agreement over 90%. The interrater agreement ranged from a low of 74.15% on Item 5 to a high of 99.72% on Item 22.

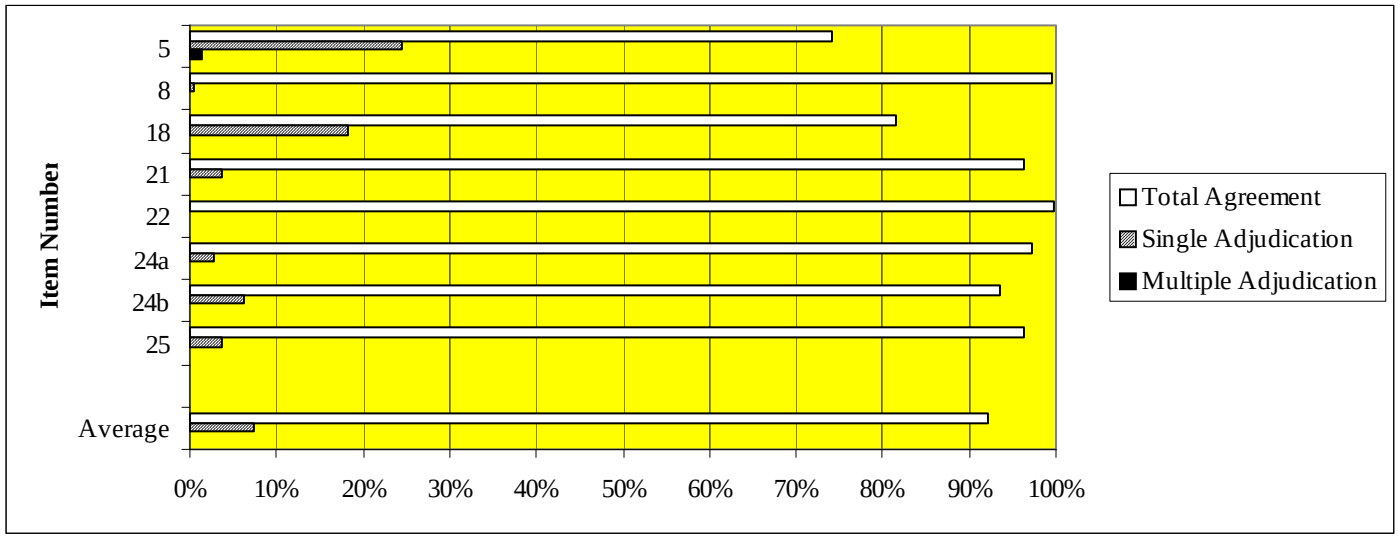

Figure 38. Interrater agreement on Grade 7 External Assessment, by item.

[9] 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 38 and Table C1 in the Appendix). The percentage of single adjudication ranged from a low of 0.28% on Item 22 to a high of 24.43% on Item 5. The incidence of multiple adjudication was very low ranging from 0% on Items 8, 21, 22, 24a, and 25 to a high of 1.42% 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 (Items 22, 24a, 24b, and 25). Factors contributing to the lower interrater agreement (and higher adjudication) on Items 5 and 18 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 (92.60%; see Figure 39 and Table C2 in the Appendix). Interrater agreement was over 80% on seven out of the eight items, and over 90% on three-quarters of the contexts. Interrater agreement ranged from a low of 68.93% on Item 5 to a high of 100% on the Items 8, 22, and 24a.

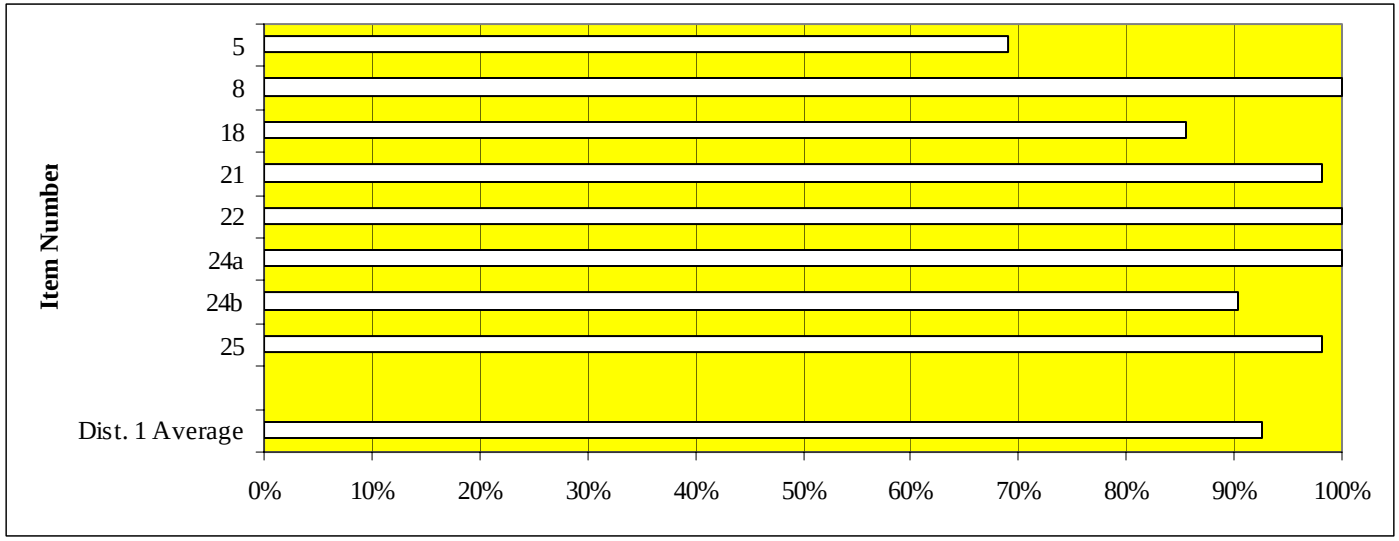

Figure 39. District 1 interrater agreement on Grade 7 External Assessment, by item.

District 2. In District 2, the interrater agreement on the Grade 7 External Assessment was very high (92.66%; see Figure 40 and Table C2 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 76.25% on Item 5 to a high of 100% on Items 8, 22, and 24b.

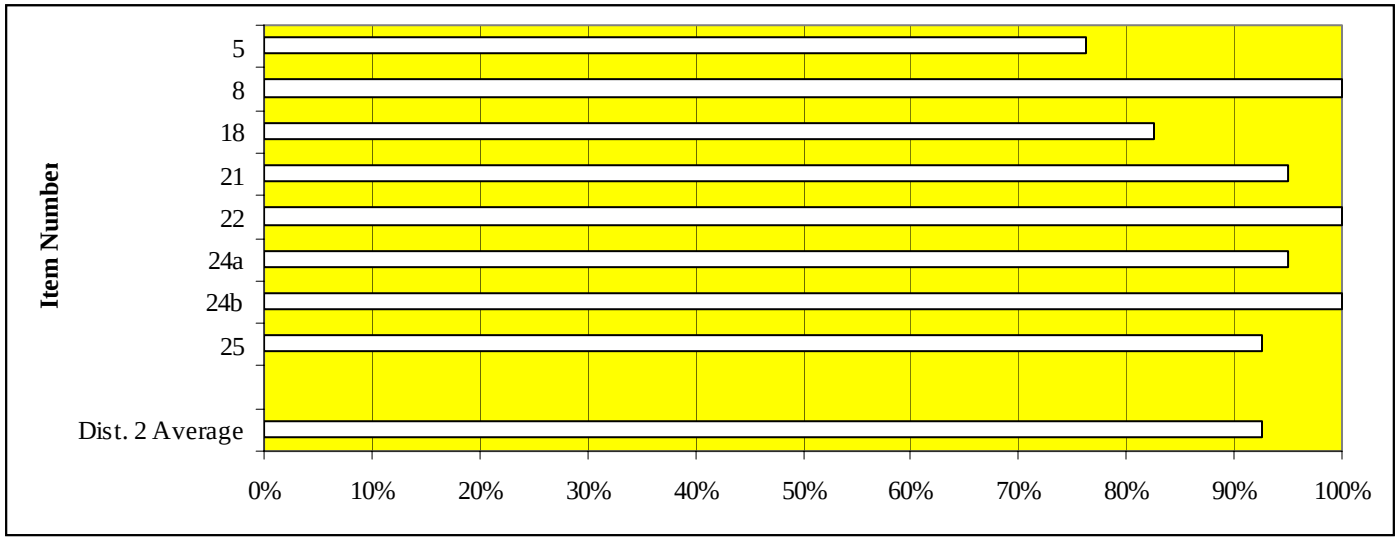

Figure 40. District 2 interrater agreement on Grade 7 External Assessment, by item.

District 3. In District 3, the interrater agreement on the Grade 7 External Assessment from District 3 was high (90.69%; see Figure 41 and Table C2 in the Appendix). Interrater agreement was over 80% on three-quarters of the items and over 90% on five out of the eight contexts. The interrater agreement ranged from a low of 75.51% on Items 5 and 18 to a high of 100% on Item 22.

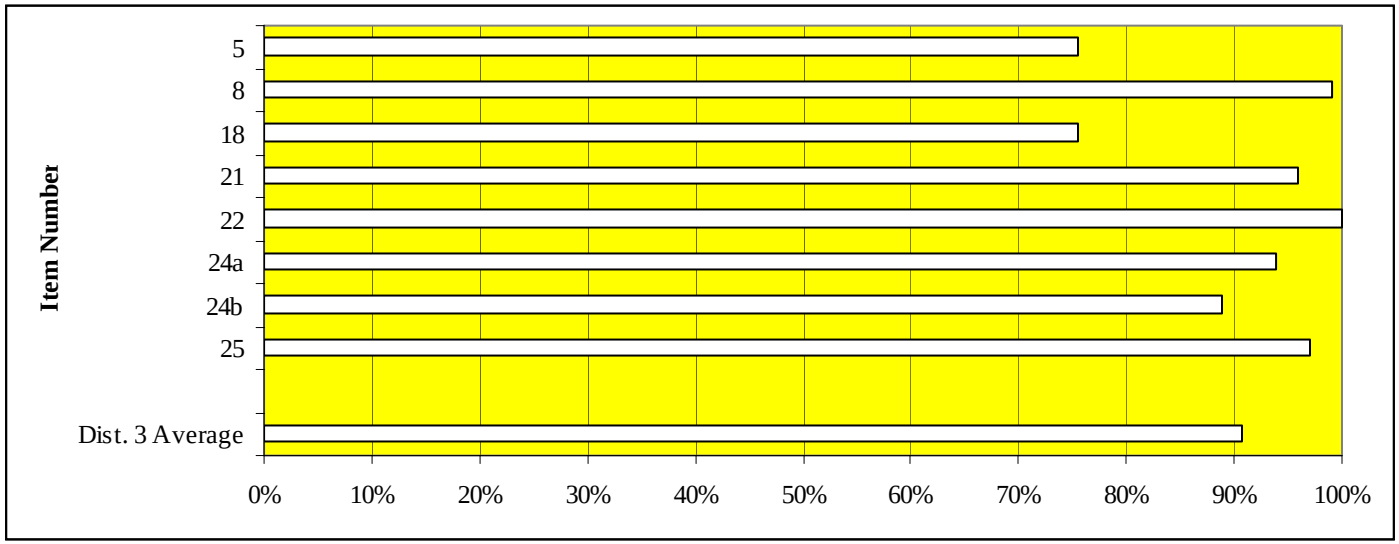

Figure 41. District 3 interrater agreement on Grade 7 External Assessment, by item.

District 4. In District 4, the interrater agreement on the Grade 7 External Assessment from District 4 was very high (93.49%; see Figure 42 and Table C2 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 77.46% on Item 5 to a high of 100% on Item 22.

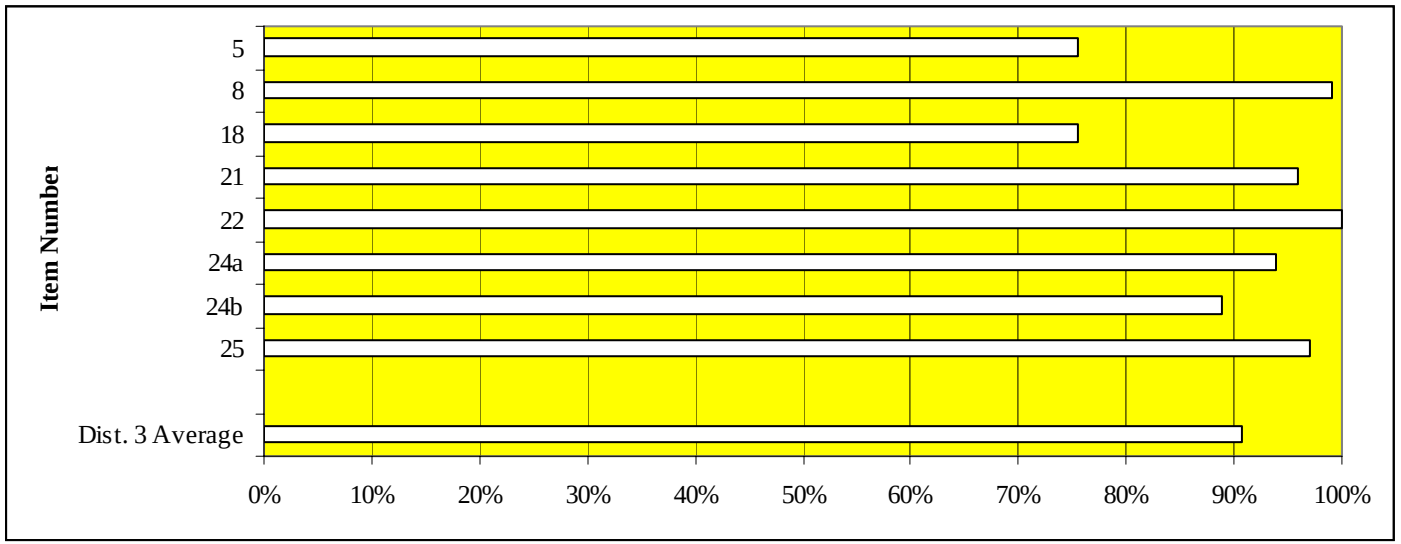

Figure 42. District 4 interrater agreement on Grade 7 External Assessment, by item.

Across districts. The interrater agreement across districts was very close on most items (see Figure 43 and Table C2 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, low on Item 24b, and high on Item 18. In District 2, interrater agreement was higher than other districts on Item 24b, lower than the other districts on Item 25, and low on Item 24a. In District 3, interrater agreement was much lower than other districts on Item 18 and low on Items 24a and 24b. In District 4, interrater agreement was high on Items 5, 24a and 24b.

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

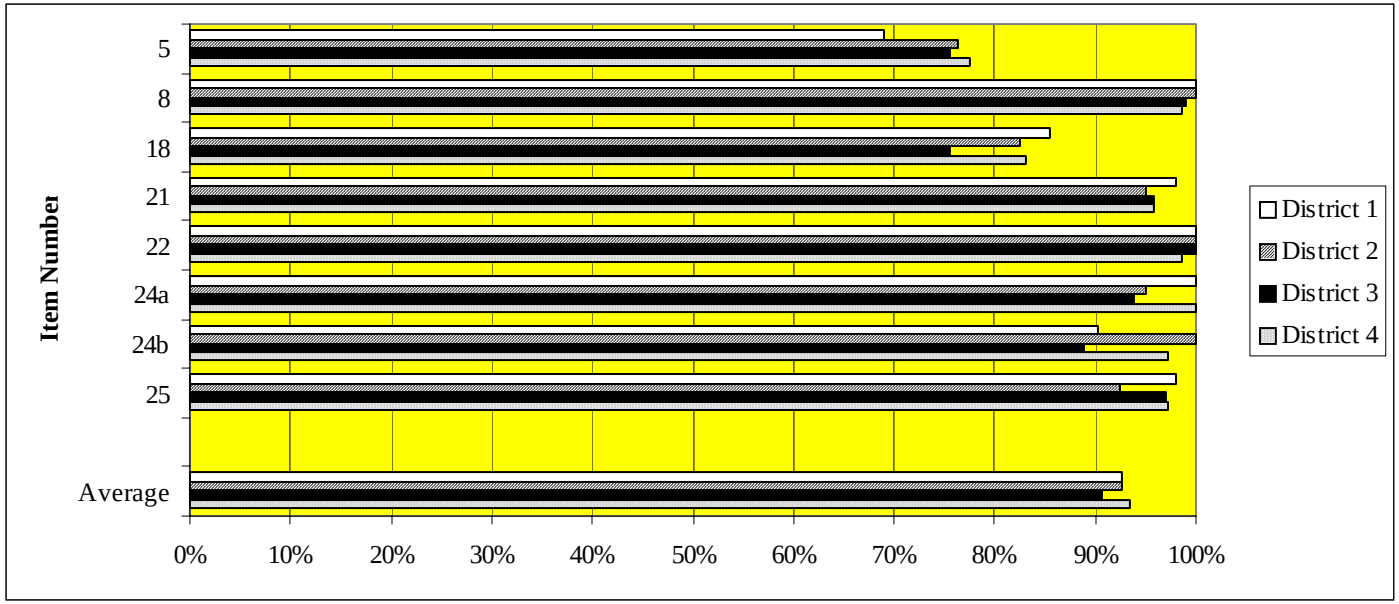

Figure 43. Across district interrater agreement on Grade 7 External Assessment, by item.

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 high (90.73%; see Figure 44 and Table C3 in the Appendix). Interrater agreement was over 80% on all but two items and over 90% on three-quarters of the items. The interrater agreement ranged from a low of 70.97% on Item 5 to a high of 100% on Items 8 and 22. The other item with low interrater agreement was Item 5 at 74.19%.

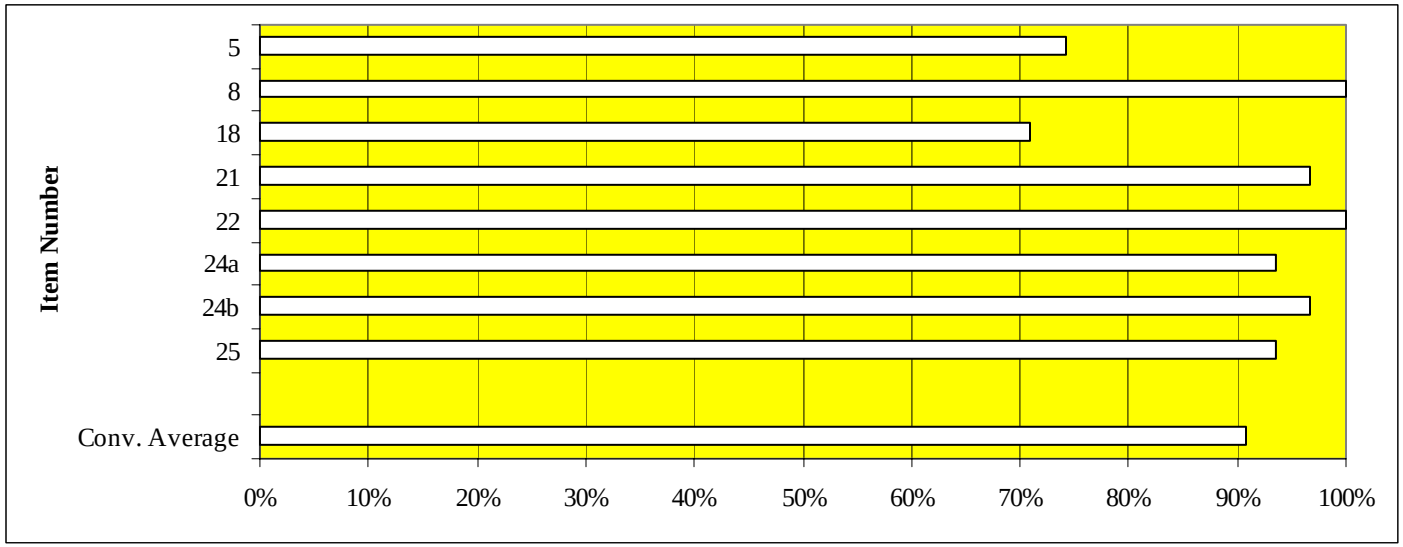

Figure 44. Interrater agreement on Grade 7 External Assessment, by item: Conventional curricula.

Mathematics in Context classes. The interrater agreement on the Grade 7 External Assessment from *Mathematics in Context* classes was very high (92.41%; see Figure 45 and Table C3 in the Appendix). Interrater agreement was over 80% on seven out of the eight items. Interrater agreement was over 90% on three-quarters of the items. The interrater agreement ranged from a low of 74.14% on Item 5 to a high of 99.69% on Item 22.

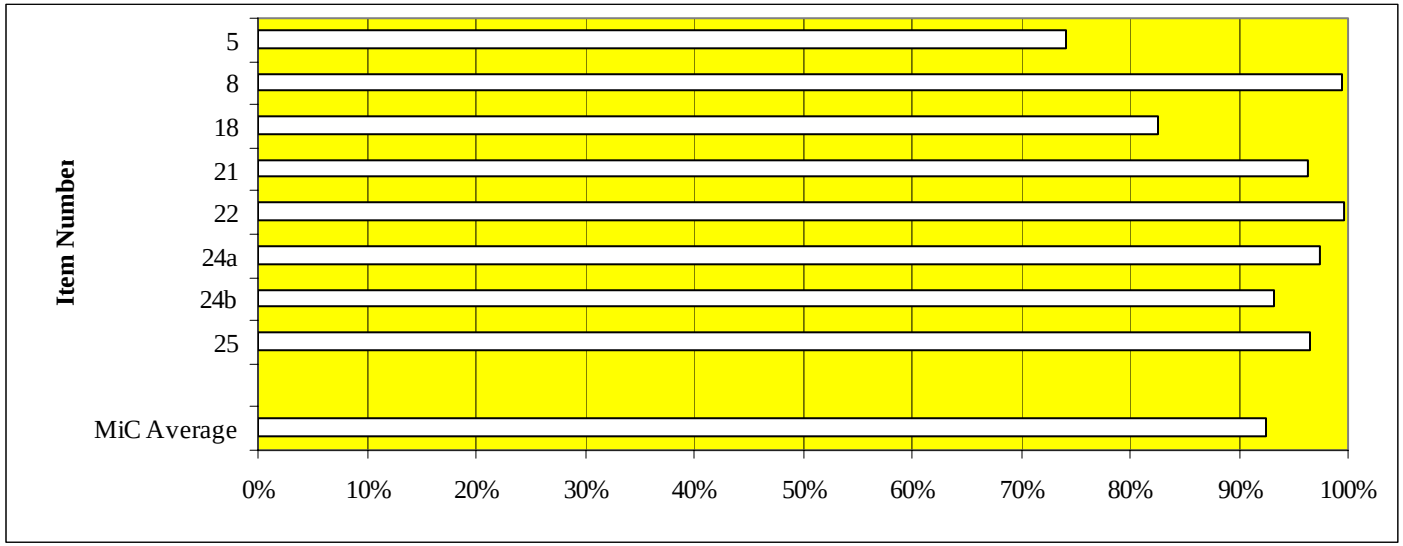

Figure 45. Interrater agreement on Grade 7 External Assessment, by item: *Mathematics in Context* classes.

Across program. Overall, the interrater agreement in conventional curricula and *Mathematics in Context* classes was similar (see Figure 46 and Table C3 in the Appendix). The average interrater agreement for conventional curricula was 90.73% and 92.41% for *Mathematics in Context* classes. The interrater agreement was higher (5% or greater) on assessments from the *Mathematics in Context* classes for Item 18.

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

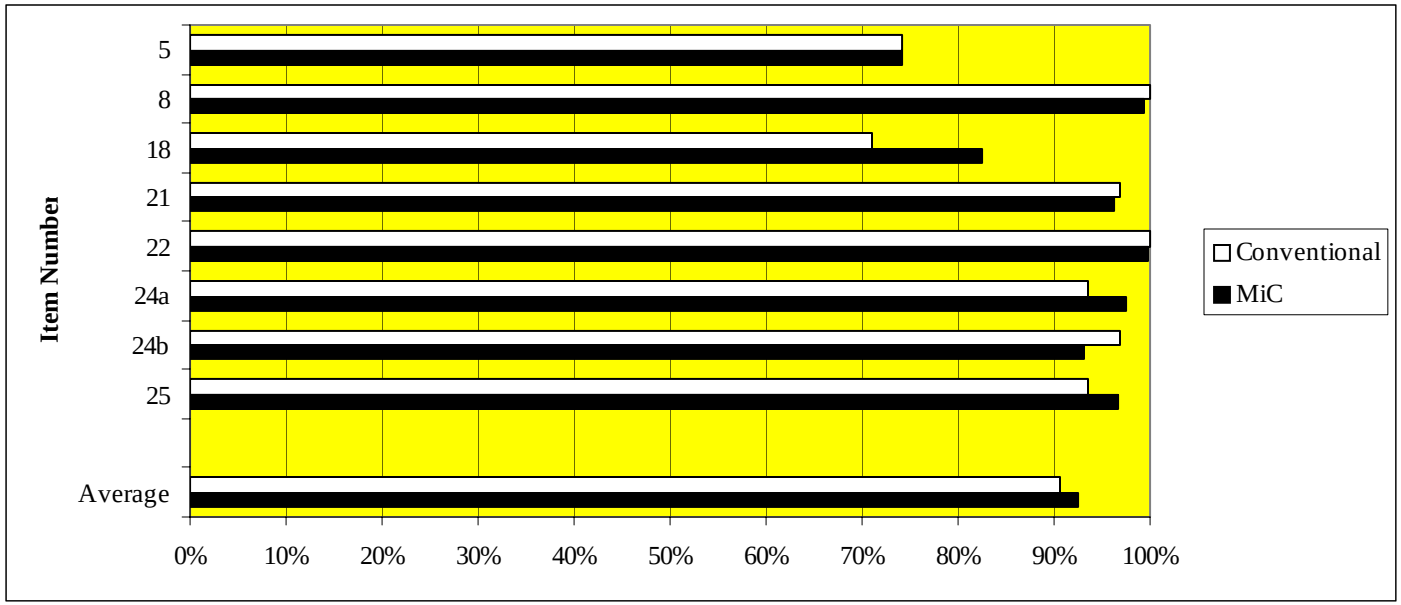

Figure 46. Interrater agreement on Grade 7 External Assessment, by item: Conventional curricula and *Mathematics in Context* classes.

Grade 8

Overall Interrater Reliability

The interrater agreement on the Grade 8 External Assessment was very high (92.78%; see Figure 47 and Appendix C4). Interrater agreement was over 80% on all of the items.[10] Interrater agreement was over 90% on seven out of the ten items. The interrater agreement ranged from a low of 81.25% on Item 1 to a high of 99.38% on Items 4 and 16.

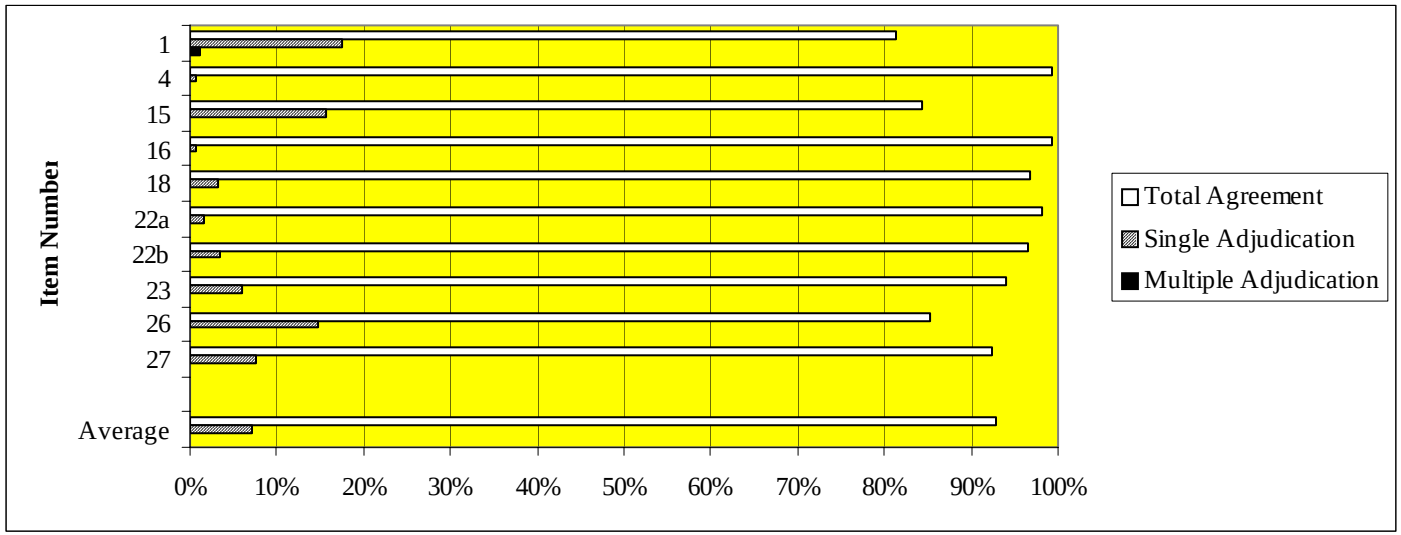

Figure 47. Interrater agreement on Grade 8 External Assessment, by item.

[10] 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 47 and Table C4 in the Appendix). The percentage of single adjudication ranged from a low of 0.63% on Items 4 and 16 to a high of 17.50% on Item 1. The incidence of multiple adjudication was very low ranging from 0% on 8 items (Items 4, 15, 16, 18, 22b, 23, 26, and 27) to a high of 1.25% on Item 1.

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 (92.35%; see Figure 48 and Table C5 in the Appendix). Interrater agreement was over 80% on all of the ten items. Interrater agreement was over 90% on seven out of ten items. The interrater agreement ranged from a low of 80.39% on Item 26 to a high of 100% on the Item 16.

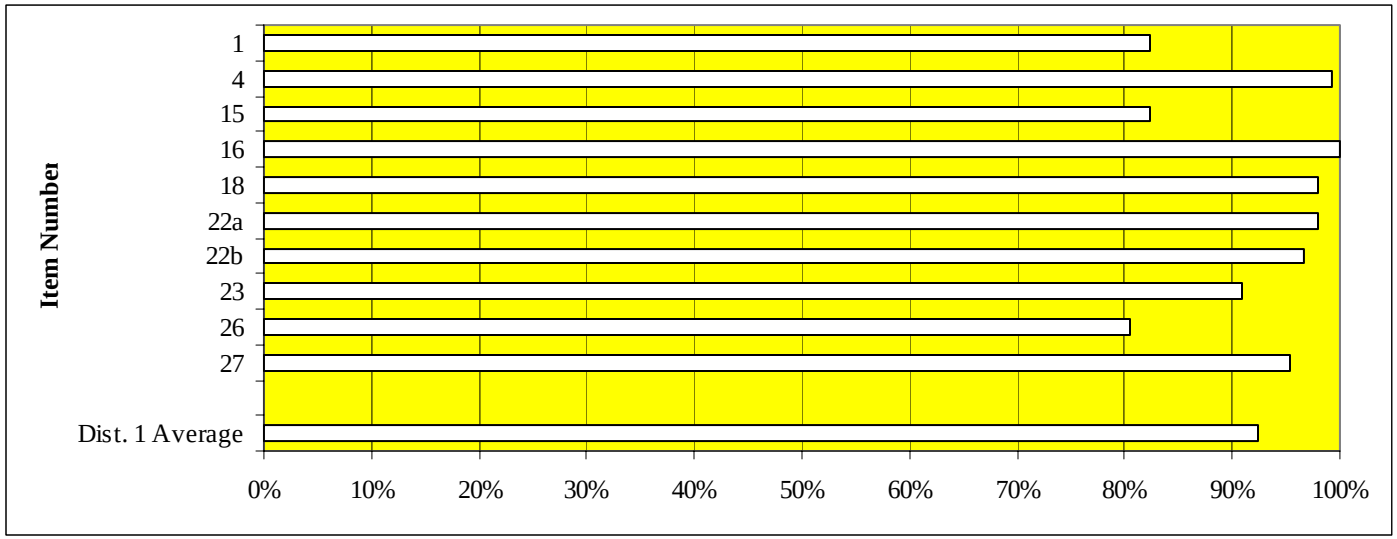

Figure 48. District 1 interrater agreement on Grade 8 External Assessment, by item.

District 2. In District 2, the interrater agreement on the Grade 8 External Assessment was very high (93.13%; see Figure 49 and Table C5 in the Appendix). Interrater agreement was over 80% on all of the ten items. Interrater agreement was over 90% on seven out of the ten items. The interrater agreement ranged from a low of 82.29% on Item 1 to a high of 98.96% on 3 items (Items 4, 16, and 22a).

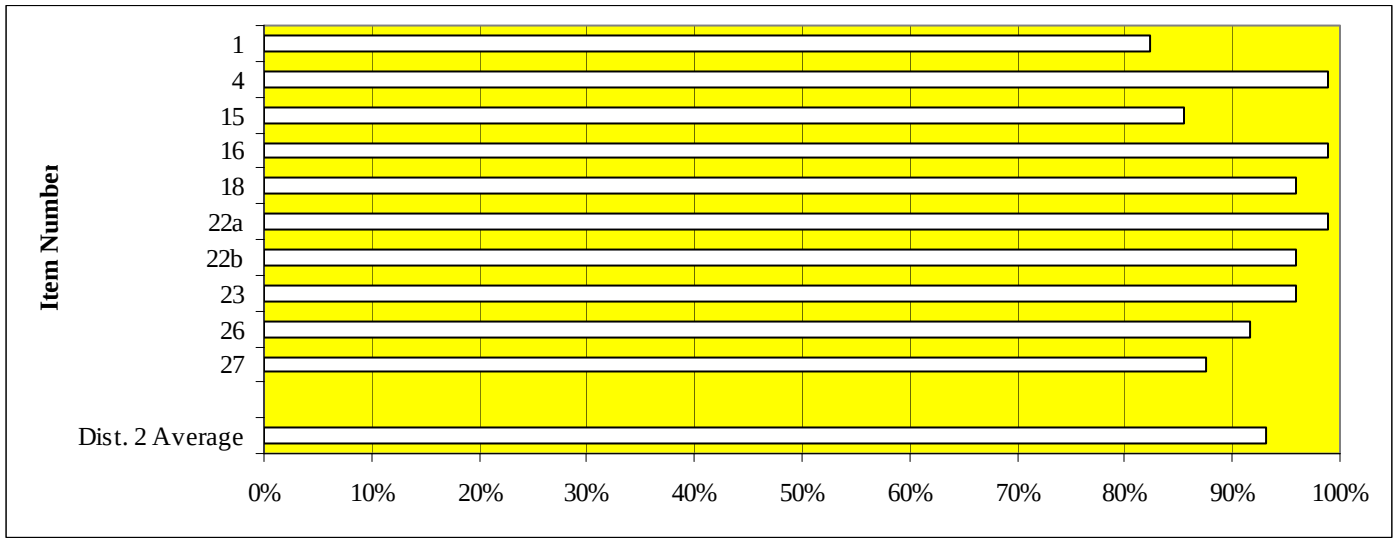

Figure 49. District 2 interrater agreement on Grade 8 External Assessment, by item.

District 3. In District 3,[11] the interrater agreement on the Grade 8 External Assessment from District 3 was high (86.67%; see Figure 50 and Table C5 in the Appendix). Interrater agreement was over 80% on nine out of the ten items. Half of the items had interrater agreement over 90%. The interrater agreement ranged from a low of 50.00% on Item 15 to a high of 100% on Items 4, 16, 18, 22a, and 23. The other item with low interrater agreement was Item 26 at 66.67%.

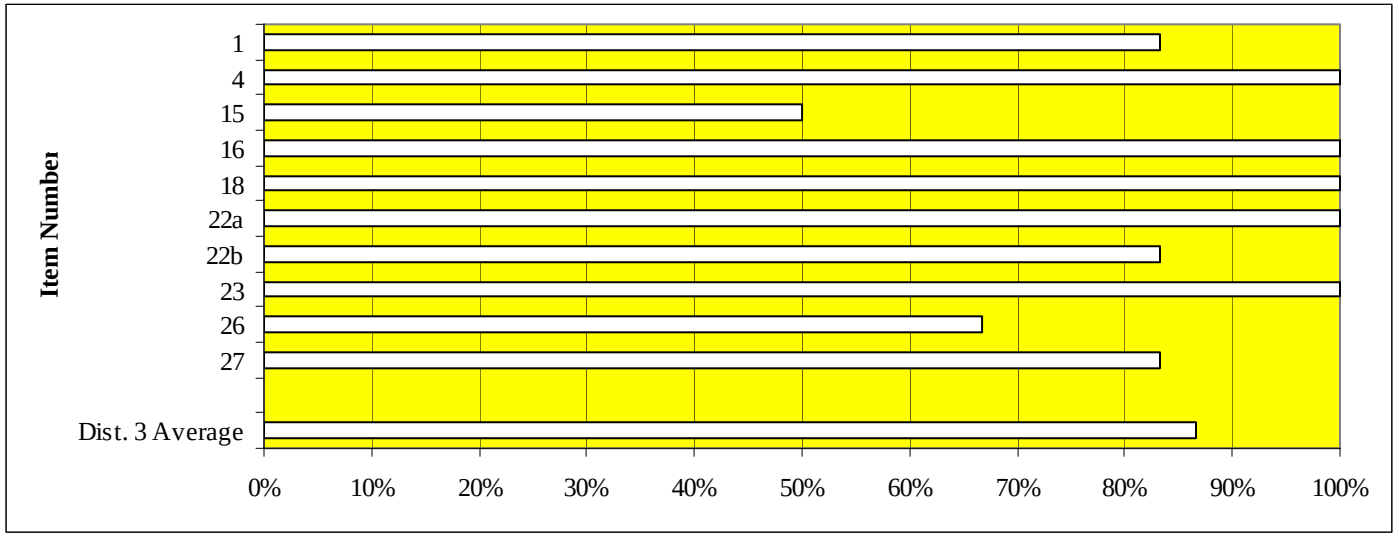

Figure 50. District 3 interrater agreement on Grade 8 External Assessment, by item.

[11] Only 6 External Assessments were received from the special education class from District 3.

District 4. In District 4, the interrater agreement on the Grade 8 External Assessment from District 4 was very high (93.85%; see Figure 51 and Table C5 in the Appendix). Interrater agreement was over 80% on nine out of the ten items, and over 90% on four-fifths of the items. The interrater agreement ranged from a low of 76.92% on Item 1 to a high of 100% on Item 4.

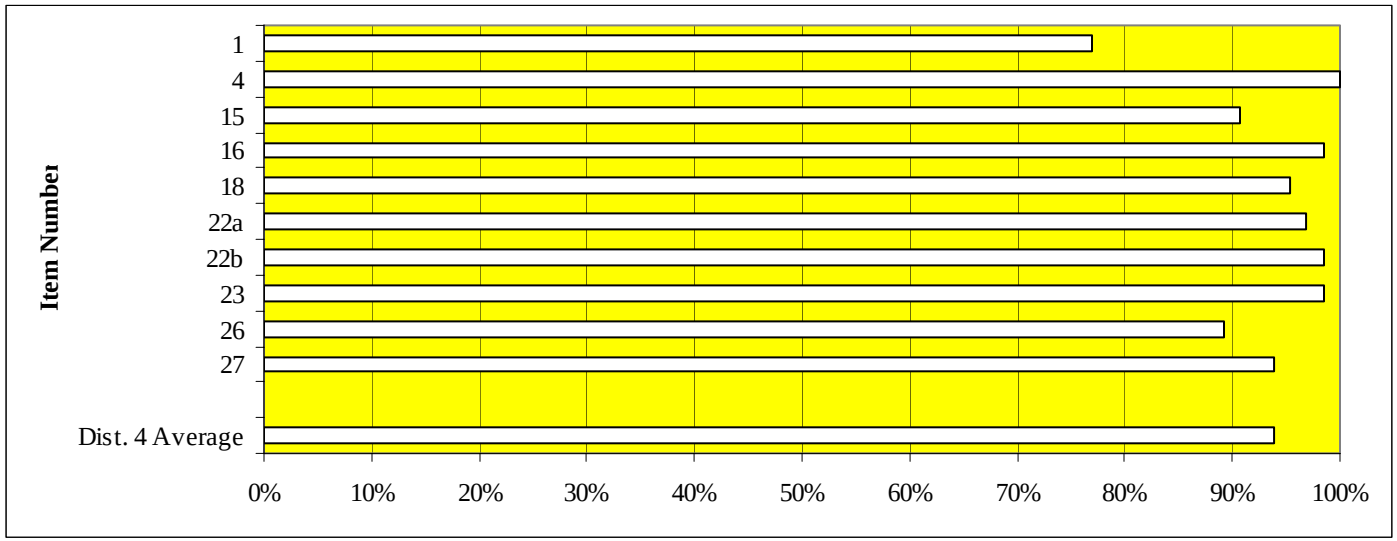

Figure 51. District 4 interrater agreement on Grade 8 External Assessment, by item.

Across districts. The interrater agreement across districts was very close on most items (see Figure 52 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 23. In District 2, interrater agreement was similar or average compared to the other districts. In District 3, interrater agreement was much lower on Items 15, 22b, 26, and 27; In District 4, interrater agreement was lower than in the other districts on Item 1 and higher on Item 15.

The large differences in interrater agreement were most likely due to (a) content study teachers taught and (b) proportion of nonresponse and incorrect responses. The large differences in interrater agreement was probably due to the fact that the only assessments received from District 3 were from the special education class

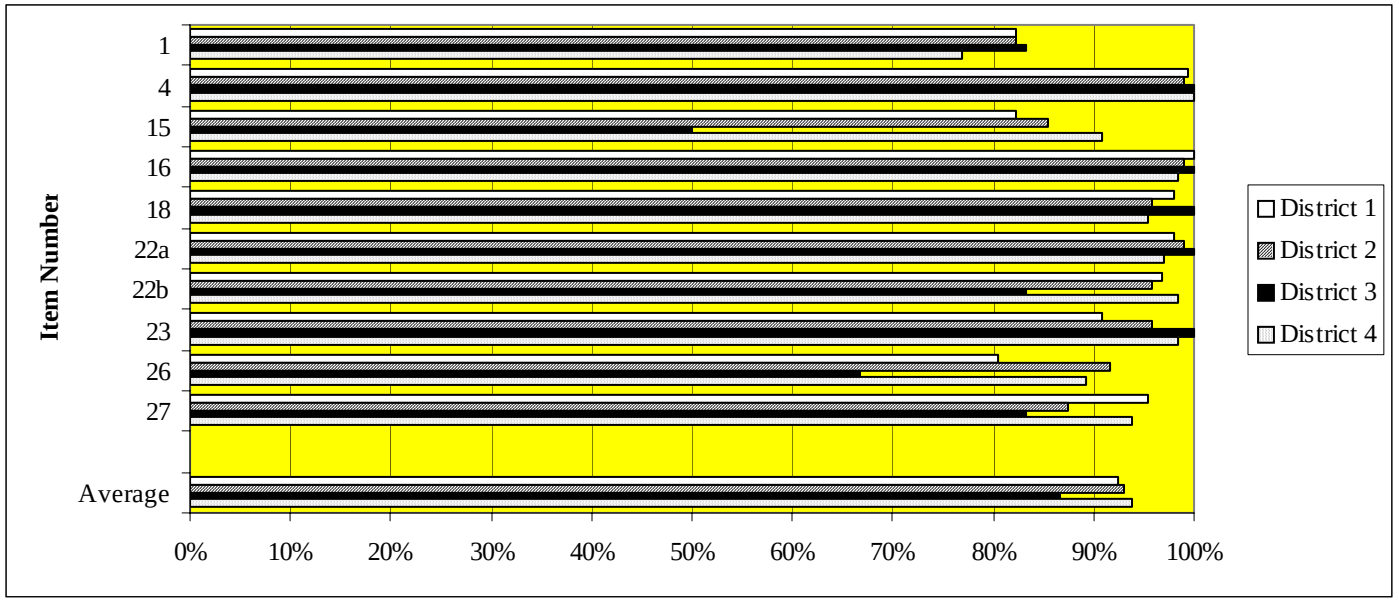

Figure 52. Across district interrater agreement on Grade 8 External Assessment, by item.

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 (94.86%; see Figure 53 and Table C6 in the Appendix). Interrater agreement was 80% or higher on all items. Interrater agreement was over 90% on four-fifths of the items. Interrater agreement ranged from a low of 80.00% on Item 1 to a high of 100% on Items 4, 16, 18, and 27.

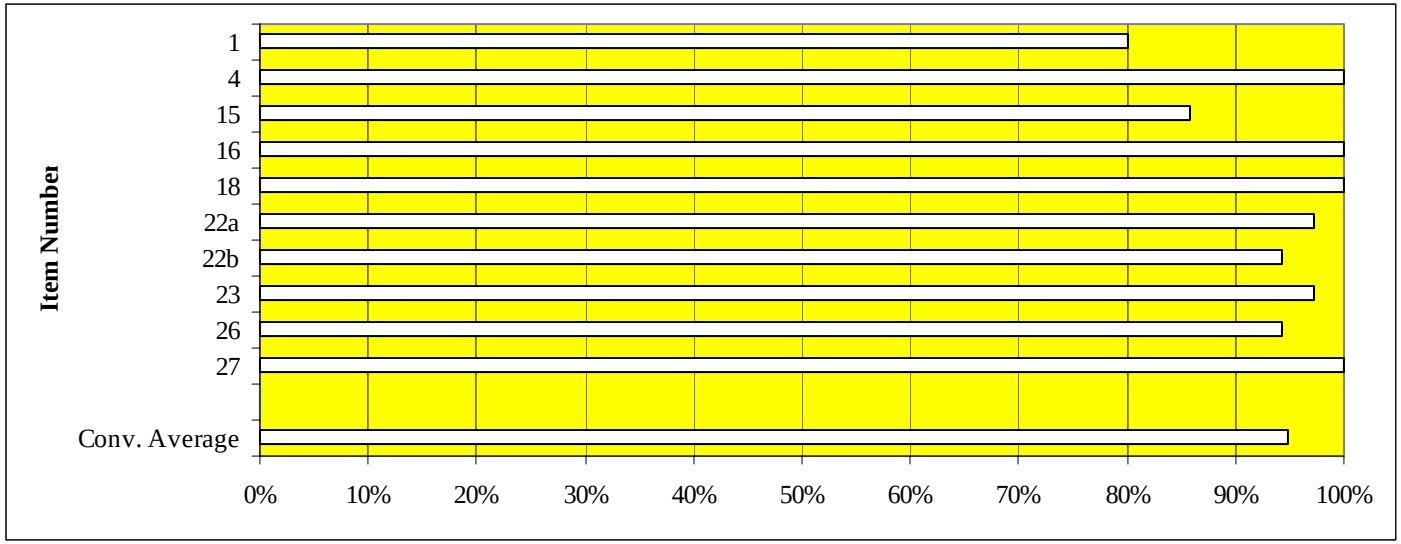

Figure 53. Interrater agreement on Grade 8 External Assessment, by item: Conventional curricula.

Mathematics in Context classes. The interrater agreement on the Grade 8 External Assessment from *Mathematics in Context* classes was very high (92.42%; see Figure 54 and Table C6 in the Appendix). Interrater agreement was over 80% on all of the items. More than two-thirds of the items had interrater agreement over 90%. The interrater agreement ranged from a low of 81.40% on Item 1 to a high of 99.30% on Items 4 and 16.

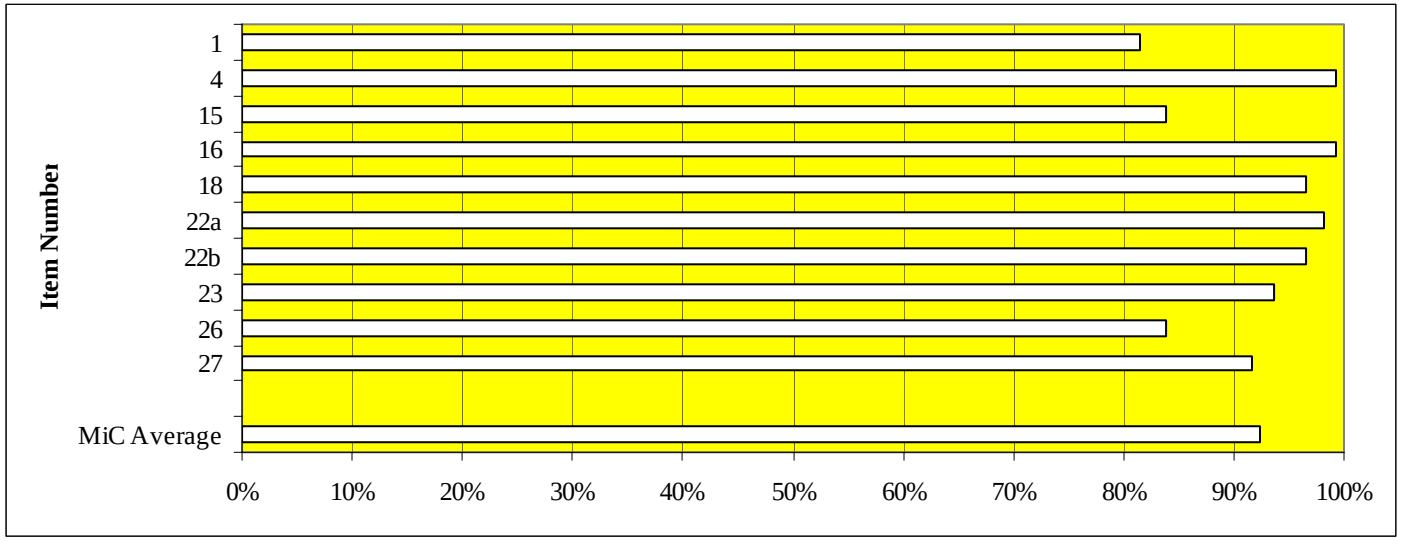

Figure 54. Interrater agreement on Grade 8 External Assessment, by item: *Mathematics in Context* classes.

Across program. Overall, the interrater agreement in conventional curricula and *Mathematics in Context* classes was similar (see Figure 55 and Table C6 in the Appendix). The average interrater agreement for conventional curricula was 94.86% and 92.42% 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 26 and 27.

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

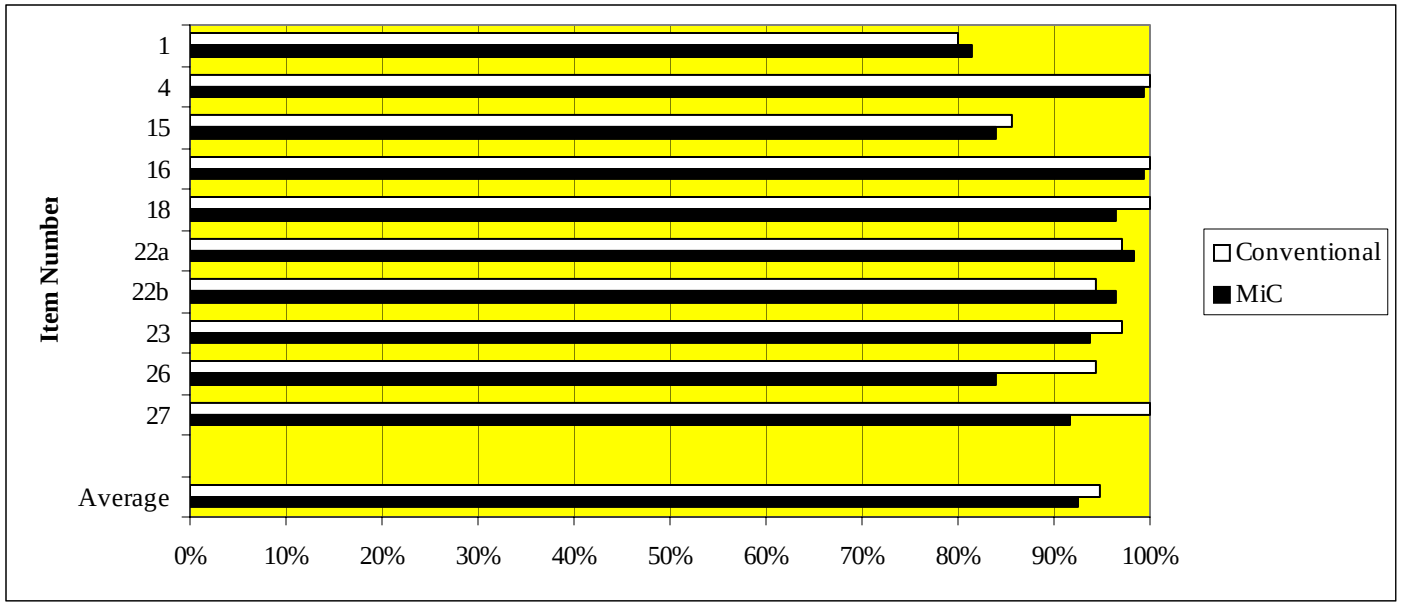

Figure 55. Interrater agreement on Grade 8 External Assessment, by item: Conventional curricula and *Mathematics in Context* classes.

Conclusion

By design, many of the items on the External Assessment were used at more than one grade level (see Figure 56 and Table C7 in the Appendix). Two factors led to higher interrater agreement. First, items eliciting lower level responses were less complex to score (Items 8 and 22 on EA7 and Items 4 and 18 on 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. Three items had lower interrater agreement (Items 5 and 18 on EA7 and Items 1, 26 and 15 on EA8) because of difficulties with the open-ended format and multiple scoring criteria.

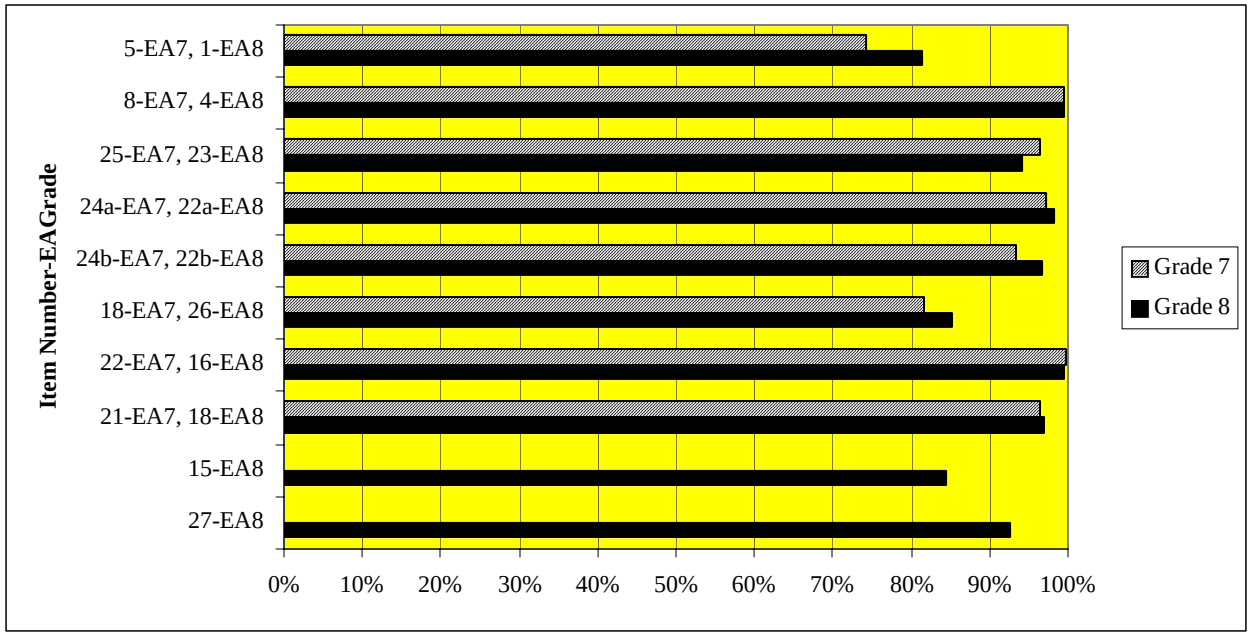

Figure 56. Interrater agreement of items from the Grades 7 and 8 from the 1999–2000 External Assessment.

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) effective scoring procedures; (c) less complex rubrics which could not be changed; and (d) 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.

Interrater agreement on EA anchor items improved at succeeding grade level assessments, indicating increasing rater competence as they worked through identical items on the Grade 7 and Grade 8 External Assessments.

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.

Very little difference in interrater agreement between conventional curricula and *Mathematics in Context* classes was apparent. The differences most likely were due to the content sudy teachers taught.

Appendix A

Interrater Reliability by Scoring Institute and by Rater

Interrater Agreement, By Scoring Institute and 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 2	PSA-7 (85) &	4	18	A	306	82.70%	59	15.95%	5	1.35%	370
PSA-7	(5/24/00-	PSA-8 (99)			B	296	88.89%	32	9.61%	5	1.50%	333
PY #15-21,	5/25/00)				C	77	78.57%	20	20.41%	1	1.02%	98
PYG #22-26,					D	71	78.02%	19	20.88%	1	1.10%	91
PSA-8					E	290	86.57%	38	11.34%	7	2.09%	335
CM #1,					F	265	82.55%	55	17.13%	1	0.31%	321
ST #10-14					G	289	88.11%	38	11.59%	1	0.30%	328
					H	375	89.07%	41	9.74%	5	1.19%	421
					I	258	90.85%	24	8.45%	2	0.70%	284
					J	346	90.58%	34	8.90%	2	0.52%	382
					K	226	87.60%	32	12.40%	0	0.00%	258
				Total:	11	2799		392		30		3221
				Average:			**86.90%**		**12.17%**		**0.93%**	
2	District 4	PSA-8 (42)	1	1	L	30	72.09%	11	25.58%	1	2.33%	42
PSA-8	(6/8/00)				M	19	77.36%	8	22.64%	0	0.00%	27
CM #1					N	11	73.33%	3	20.00%	1	6.67%	15
				Total:	3	60		22		2		84
				Average:			**71.43%**		**26.19%**		**2.38%**	
3	Madison-1	PSA-7 (318),	13	51	O	2030	94.03%	126	5.84%	3	0.14%	2159
PSA-7	(7/24/00-	PSA-8 (374),			P	1581	92.29%	129	7.53%	3	0.18%	1713
BF #1-7,	7/28/00)	EA-7 (315), &			Q	1532	93.53%	105	6.41%	1	0.06%	1638
PN #8-10,		EA-8 (308)			R	2777	92.14%	234	7.76%	3	0.10%	3014
A #11-14,					S	1423	90.75%	139	8.86%	6	0.38%	1568
PY #15-21,					T	1467	91.52%	133	8.30%	3	0.19%	1603
PYG #22-26,					U	1521	92.97%	115	7.03%	0	0.00%	1636
PSA-8					V	1694	92.92%	124	6.80%	5	0.27%	1823
CM #1,					W	2478	91.81%	214	7.93%	7	0.26%	2699
LP #2-4,					X	1544	91.04%	146	8.61%	6	0.35%	1696
KC #5-7,					Y	1405	92.68%	107	7.06%	4	0.26%	1516
SS #8-9,					Z	1593	91.34%	143	8.20%	8	0.46%	1744
ST #10-14,					AA	2066	93.74%	133	6.03%	5	0.23%	2204
PK #15-17,					AB	1496	92.98%	110	6.84%	3	0.19%	1609
CU #18-21,					AC	1050	91.30%	99	8.61%	1	0.09%	1150
EA-7					AD	1353	90.62%	136	9.11%	4	0.27%	1493
#24a+b,					AE	2250	92.94%	167	6.90%	4	0.17%	2421
EA-8					AF	148	83.15%	28	15.73%	2	1.12%	178
#22a+b				Total:	18	29260		2360		66		31686
				Average:			**92.34%**		**7.45%**		**0.21%**	

Interrater Agreement, By Scoring Institute and Rater

Table A1 *(continued)*
Interrater Reliability by Scoring Institute and by Rater

Institute (Contexts Scored)*	Location (Date)	Assessments Rated *(N)*	Contexts *(N)*	Items in Contexts *(N)*	Rater	Agreement *(N)*	Agreement *%*	Single Adjudication *(N)*	Single Adjudication *%*	Multiple Adjudication *(N)*	Multiple Adjudication *%*	Student Responses Rated *(N)*
4	Madison-2	EA-7 (315) &	8	14	AG	681	93.42%	46	6.31%	2	0.27%	729
EA-7	(8/7/00-	EA-8 (308)			AH	615	89.13%	75	10.87%	0	0.00%	690
#5,	7/8/00)				AI	809	92.67%	63	7.22%	1	0.11%	873
#8,					AJ	579	91.18%	53	8.35%	3	0.47%	635
#18,					AK	589	93.20%	42	6.65%	1	0.16%	632
#21,					AL	530	87.17%	76	12.50%	2	0.33%	608
#22,					AM	562	90.94%	55	8.90%	1	0.16%	618
#25,					AN	632	93.35%	43	6.35%	2	0.30%	677
EA-8					AO	485	95.28%	24	4.72%	0	0.00%	509
#1,					AP	545	94.13%	33	5.70%	1	0.17%	579
#3,					AQ	559	90.45%	59	9.55%	0	0.00%	618
#15,					AR	467	91.57%	42	8.24%	1	0.20%	510
#16,					AS	394	91.42%	36	8.35%	1	0.23%	431
#18,					AT	461	88.15%	61	11.66%	1	0.19%	523
#23,					AU	638	89.99%	67	9.45%	4	0.56%	709
#26,				Total:	15	8546		775		20		9341
#27				**Average:**			**91.49%**		**8.30%**		**0.21%**	

* Context Key:
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-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 1999-2000 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	354	346	97.74%	8	2.26%	0	0.00%
	2	354	306	86.44%	48	13.56%	0	0.00%
	3	354	320	90.40%	34	9.60%	0	0.00%
	4	354	307	86.72%	47	13.28%	0	0.00%
	5	354	325	91.81%	27	7.63%	2	0.56%
	6	354	325	91.81%	27	7.63%	2	0.56%
	7	354	331	93.50%	23	6.50%	0	0.00%
	Total	2478	2260		214		4	
	Average			**91.20%**		**8.64%**		**0.16%**
The Pentagon	8	354	304	85.88%	48	13.56%	2	0.56%
	9	354	285	80.51%	67	18.93%	2	0.56%
	10	354	336	94.92%	18	5.08%	0	0.00%
	Total	1062	925		133		4	
	Average			**87.10%**		**12.52%**		**0.38%**
Airships	11	354	344	97.18%	10	2.82%	0	0.00%
	12	354	263	74.29%	77	21.75%	14	3.95%
	13	354	314	88.70%	39	11.02%	1	0.28%
	14	354	325	91.81%	29	8.19%	0	0.00%
	Total	1416	1246		155		15	
	Average			**87.99%**		**10.95%**		**1.06%**
Pyramids	15	354	276	77.97%	76	21.47%	2	0.56%
	16	354	323	91.24%	30	8.47%	1	0.28%
	17	354	345	97.46%	9	2.54%	0	0.00%
	18	354	283	79.94%	67	18.93%	4	1.13%
	19	354	311	87.85%	42	11.86%	1	0.28%
	20	354	327	92.37%	27	7.63%	0	0.00%
	21	354	324	91.53%	30	8.47%	0	0.00%
	Total	2478	2189		281		8	
	Average			**88.34%**		**11.34%**		**0.32%**
Playgrounds	22	354	347	98.02%	7	1.98%	0	0.00%
	23	354	335	94.63%	19	5.37%	0	0.00%
	24	354	284	80.23%	65	18.36%	5	1.41%
	25	354	336	94.92%	18	5.08%	0	0.00%
	26	354	326	92.09%	28	7.91%	0	0.00%
	Total	1770	1628		137		5	
	Average			**91.98%**		**7.74%**		**0.28%**
PSA7	Total	9204	8248		920		36	
	Average			**89.61%**		**10.00%**		**0.39%**

Interrater Agreement, By Scoring Institute and Rater

Table B2

Interrater Agreement by District for 1999-2000 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	100	98	98.00%	86	84	97.67%	97	96	98.97%	71	68	95.77%
	2	100	90	90.00%	86	75	87.21%	97	86	88.66%	71	55	77.46%
	3	100	96	96.00%	86	74	86.05%	97	87	89.69%	71	63	88.73%
	4	100	93	93.00%	86	76	88.37%	97	81	83.51%	71	57	80.28%
	5	100	95	95.00%	86	78	90.70%	97	92	94.85%	71	60	84.51%
	6	100	92	92.00%	86	80	93.02%	97	88	90.72%	71	65	91.55%
	7	100	92	92.00%	86	82	95.35%	97	91	93.81%	71	66	92.96%
	Total	700	656		602	549		679	621		497	434	
	Average			**93.71%**			**91.20%**			**91.46%**			**87.32%**
The Pentagon	8	100	87	87.00%	86	71	82.56%	97	82	84.54%	71	64	90.14%
	9	100	89	89.00%	86	71	82.56%	97	73	75.26%	71	52	73.24%
	10	100	95	95.00%	86	81	94.19%	97	92	94.85%	71	68	95.77%
	Total	300	271		258	223		291	247		213	184	
	Average			**90.33%**			**86.43%**			**84.88%**			**86.38%**
Airships	11	100	99	99.00%	86	84	97.67%	97	90	92.78%	71	71	100.00%
	12	100	73	73.00%	86	58	67.44%	97	72	74.23%	71	60	84.51%
	13	100	93	93.00%	86	73	84.88%	97	84	86.60%	71	64	90.14%
	14	100	91	91.00%	86	82	95.35%	97	87	89.69%	71	65	91.55%
	Total	400	356		344	297		388	333		284	260	
	Average			**89.00%**			**86.34%**			**85.82%**			**91.55%**
Pyramids	15	100	76	76.00%	86	59	68.60%	97	82	84.54%	71	59	83.10%
	16	100	94	94.00%	86	72	83.72%	97	91	93.81%	71	66	92.96%
	17	100	99	99.00%	86	82	95.35%	97	94	96.91%	71	70	98.59%
	18	100	91	91.00%	86	55	63.95%	97	81	83.51%	71	56	78.87%
	19	100	95	95.00%	86	56	65.12%	97	94	96.91%	71	66	92.96%
	20	100	94	94.00%	86	76	88.37%	97	89	91.75%	71	68	95.77%
	21	100	91	91.00%	86	74	86.05%	97	92	94.85%	71	67	94.37%
	Total	700	640		602	474		679	623		497	452	
	Average			**91.43%**			**78.74%**			**91.75%**			**90.95%**
Playgrounds	22	100	98	98.00%	86	82	95.35%	97	97	100.00%	71	70	98.59%
	23	100	98	98.00%	86	75	87.21%	97	94	96.91%	71	68	95.77%
	24	100	89	89.00%	86	69	80.23%	97	75	77.32%	71	51	71.83%
	25	100	97	97.00%	86	78	90.70%	97	91	93.81%	71	70	98.59%
	26	100	98	98.00%	86	77	89.53%	97	89	91.75%	71	62	87.32%
	Total	500	480		430	381		485	446		355	321	
	Average			**96.00%**			**88.60%**			**91.96%**			**90.42%**
PSA7	Total	2600	2403		2236	1924		2522	2270		2568	2389	
	Average			**92.42%**			**86.05%**			**90.01%**			**93.03%**

Interrater Agreement, By Scoring Institute and Rater

Table B3

Interrater Agreement by Program for 1999-2000 Grade 7 Problem Solving Assessments

Context	Item Number	Conventional Curricula			Mathematics in Context		
		Assessments	Agreement		Assessments	Agreement	
		(N)	*(N)*	*%*	*(N)*	*(N)*	*%*
Baby Feeding	1	33	32	96.97%	321	314	97.82%
	2	33	32	96.97%	321	274	85.36%
	3	33	29	87.88%	321	291	90.65%
	4	33	30	90.91%	321	277	86.29%
	5	33	31	93.94%	321	294	91.59%
	6	33	32	96.97%	321	293	91.28%
	7	33	32	96.97%	321	299	93.15%
	Total	231	218		2247	2042	
	Average			**94.37%**			**90.88%**
Pentagon	8	33	32	96.97%	321	272	84.74%
	9	33	28	84.85%	321	257	80.06%
	10	33	31	93.94%	321	305	95.02%
	Total	99	91		963	834	
	Average			**91.92%**			**86.60%**
Airships	11	33	32	96.97%	321	312	97.20%
	12	33	24	72.73%	321	239	74.45%
	13	33	30	90.91%	321	284	88.47%
	14	33	31	93.94%	321	294	91.59%
	Total	132	117		1284	1129	
	Average			**88.64%**			**87.93%**
Pyramids	15	33	25	75.76%	321	251	78.19%
	16	33	28	84.85%	321	295	91.90%
	17	33	33	100.00%	321	312	97.20%
	18	33	22	66.67%	321	261	81.31%
	19	33	27	81.82%	321	284	88.47%
	20	33	32	96.97%	321	295	91.90%
	21	33	30	90.91%	321	294	91.59%
	Total	231	197		2247	1992	
	Average			**85.28%**			**88.65%**
Playgrounds	22	33	32	96.97%	321	315	98.13%
	23	33	30	90.91%	321	305	95.02%
	24	33	28	84.85%	321	256	79.75%
	25	33	29	87.88%	321	307	95.64%
	26	33	32	96.97%	321	294	91.59%
	Total	165	151		1605	1477	
	Average			**91.52%**			**92.02%**
PSA7	Total	858	774		8346	7474	
	Average			**90.21%**			**89.55%**

Interrater Agreement, By Scoring Institute and Rater

Table B4

Interrater Agreement on 1999-2000 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	378	313	82.80%	62	16.40%	3	0.79%
	Total	378	313		62		3	
	Average			**82.80%**		**16.40%**		**0.79%**
Lopsided	2	378	349	92.33%	26	6.88%	3	0.79%
	3	378	337	89.15%	41	10.85%	0	0.00%
	4	378	326	86.24%	49	12.96%	3	0.79%
	Total	1134	1012		116		6	
	Average			**89.24%**		**10.23%**		**0.53%**
Key Cards	5	378	368	97.35%	10	2.65%	0	0.00%
	6	378	357	94.44%	20	5.29%	1	0.26%
	7	378	367	97.09%	11	2.91%	0	0.00%
	Total	1134	1092		41		1	
	Average			**96.30%**		**3.62%**		**0.09%**
Seesaw	8	378	365	96.56%	13	3.44%	0	0.00%
	9	378	362	95.77%	16	4.23%	0	0.00%
	Total	756	727		29		0	
	Average			**96.16%**		**3.84%**		**0.00%**
Stretch	10	378	361	95.50%	16	4.23%	1	0.26%
	11	378	359	94.97%	18	4.76%	1	0.26%
	12	378	371	98.15%	6	1.59%	1	0.26%
	13	378	352	93.12%	26	6.88%	0	0.00%
	14	378	333	88.10%	43	11.38%	2	0.53%
	Total	1890	1776		109		5	
	Average			**93.97%**		**5.77%**		**0.26%**
Parking	15	378	365	96.56%	13	3.44%	0	0.00%
	16	378	351	92.86%	27	7.14%	0	0.00%
	17	378	344	91.01%	33	8.73%	1	0.26%
	Total	1134	1060		73		1	
	Average			**93.47%**		**6.44%**		**0.09%**
Cubes	18	378	351	92.86%	27	7.14%	0	0.00%
	19	378	369	97.62%	9	2.38%	0	0.00%
	20	378	361	95.50%	17	4.50%	0	0.00%
	21	378	365	96.56%	12	3.17%	1	0.26%
	Total	1512	1446		65		1	
	Average			**95.63%**		**4.30%**		**0.07%**
PSA8	Total	7938	7426		495		17	
	Average			**93.55%**		**6.24%**		**0.21%**

Interrater Agreement, By Scoring Institute and Rater

Table B5

Interrater Agreement by District for 1999-2000 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	161	136	84.47%	100	85	85.00%	53	46	86.79%	64	46	71.88%
	Total	161	136		100	85		53	46		64	46	
	Average			**84.47%**			**85.00%**			**86.79%**			**71.88%**
Lopsided	2	161	148	91.93%	100	91	91.00%	53	51	96.23%	64	59	92.19%
	3	161	138	85.71%	100	95	95.00%	53	42	79.25%	64	62	96.88%
	4	161	142	88.20%	100	94	94.00%	53	41	77.36%	64	49	76.56%
	Total	483	428		300	280		159	134		192	170	
	Average			**88.61%**			**93.33%**			**84.28%**			**88.54%**
Key Cards	5	161	159	98.76%	100	96	96.00%	53	52	98.11%	64	61	95.31%
	6	161	150	93.17%	100	95	95.00%	53	52	98.11%	64	60	93.75%
	7	161	156	96.89%	100	99	99.00%	53	50	94.34%	64	62	96.88%
	Total	483	465		300	290		159	154		192	183	
	Average			**96.27%**			**96.67%**			**96.86%**			**95.31%**
Seesaw	8	161	157	97.52%	100	96	96.00%	53	51	96.23%	64	61	95.31%
	9	161	156	96.89%	100	96	96.00%	53	49	92.45%	64	61	95.31%
	Total	322	313		200	192		106	100		128	122	
	Average			**97.20%**			**96.00%**			**94.34%**			**95.31%**
Stretch	10	161	155	96.27%	100	92	92.00%	53	53	100.00%	64	61	95.31%
	11	161	153	95.03%	100	99	99.00%	53	46	86.79%	64	61	95.31%
	12	161	160	99.38%	100	97	97.00%	53	52	98.11%	64	62	96.88%
	13	161	144	89.44%	100	96	96.00%	53	51	96.23%	64	61	95.31%
	14	161	134	83.23%	100	96	96.00%	53	44	83.02%	64	59	92.19%
	Total	805	746		500	480		265	246		320	304	
	Average			**92.67%**			**96.00%**			**92.83%**			**95.00%**
Parking	15	161	160	99.38%	100	96	96.00%	53	47	88.68%	64	62	96.88%
	16	161	151	93.79%	100	93	93.00%	53	46	86.79%	64	61	95.31%
	17	161	151	93.79%	100	91	91.00%	53	40	75.47%	64	62	96.88%
	Total	483	462		300	280		159	133		192	185	
	Average			**95.65%**			**93.33%**			**83.65%**			**96.35%**
Cubes	18	161	149	92.55%	100	96	96.00%	53	44	83.02%	64	62	96.88%
	19	161	157	97.52%	100	99	99.00%	53	50	94.34%	64	63	98.44%
	20	161	153	95.03%	100	98	98.00%	53	50	94.34%	64	60	93.75%
	21	161	158	98.14%	100	97	97.00%	53	47	88.68%	64	63	98.44%
	Total	644	617		400	390		212	191		256	248	
	Average			**95.81%**			**97.50%**			**90.09%**			**96.88%**
PSA8	Total	3381	3167		2100	1997		1113	1004		1344	1258	
	Average			**93.67%**			**95.10%**			**90.21%**			**93.60%**

Interrater Agreement, By Scoring Institute and Rater

Table B6

Interrater Agreement by Program for 1999-2000 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	44	38	86.36%	334	275	82.34%
	Total	44	38		334	275	
	Average			**86.36%**			**82.34%**
Lopsided	2	44	42	95.45%	334	307	91.92%
	3	44	35	79.55%	334	302	90.42%
	4	44	36	81.82%	334	290	86.83%
	Total	132	113		1002	899	
	Average			**85.61%**			**89.72%**
Key Cards	5	44	44	100.00%	334	324	97.01%
	6	44	43	97.73%	334	314	94.01%
	7	44	43	97.73%	334	324	97.01%
	Total	132	130		1002	962	
	Average			**98.48%**			**96.01%**
Seesaw	8	44	43	97.73%	334	322	96.41%
	9	44	42	95.45%	334	320	95.81%
	Total	352	345		2672	2566	
	Average			**98.01%**			**96.03%**
Stretch	10	44	41	93.18%	334	320	95.81%
	11	44	44	100.00%	334	315	94.31%
	12	44	43	97.73%	334	328	98.20%
	13	44	43	97.73%	334	309	92.51%
	14	44	42	95.45%	334	291	87.13%
	Total	220	213		1670	1563	
	Average			**96.82%**			**93.59%**
Parking	15	44	44	100.00%	334	321	96.11%
	16	44	40	90.91%	334	311	93.11%
	17	44	44	100.00%	334	300	89.82%
	Total	132	128		1002	932	
	Average			**96.97%**			**93.01%**
Cubes	18	44	41	93.18%	334	310	92.81%
	19	44	43	97.73%	334	326	97.60%
	20	44	42	95.45%	334	319	95.51%
	21	44	44	100.00%	334	321	96.11%
	Total	176	170		1336	1276	
	Average			**96.59%**			**95.51%**
PSA8	Total	1188	1137		9018	8473	
	Average			**95.71%**			**93.96%**

Interrater Agreement, By Scoring Institute and Rater

Appendix C

Interrater Reliability–External Assessment

Table C1
Interrater Agreement on 1999-2000 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	352	261	74.15%	86	24.43%	5	1.42%
8	352	350	99.43%	2	0.57%	0	0.00%
18	352	287	81.53%	64	18.18%	1	0.28%
21	352	339	96.31%	13	3.69%	0	0.00%
22	352	351	99.72%	1	0.28%	0	0.00%
24a	352	342	97.16%	10	2.84%	0	0.00%
24b	352	329	93.47%	22	6.25%	1	0.28%
25	352	339	96.31%	13	3.69%	0	0.00%
Total	2816	2598		211		7	
Average			**92.26%**		**7.49%**		**0.25%**

Interrater Agreement, By Scoring Institute and Rater

Table C2

Interrater Agreement by District for 1999-2000 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	103	71	68.93%	80	61	76.25%	98	74	75.51%	71	55	77.46%
8	103	103	100.00%	80	80	100.00%	98	97	98.98%	71	70	98.59%
18	103	88	85.44%	80	66	82.50%	98	74	75.51%	71	59	83.10%
21	103	101	98.06%	80	76	95.00%	98	94	95.92%	71	68	95.77%
22	103	103	100.00%	80	80	100.00%	98	98	100.00%	71	70	98.59%
24a	103	103	100.00%	80	76	95.00%	98	92	93.88%	71	71	100.00%
24b	103	93	90.29%	80	80	100.00%	98	87	88.78%	71	69	97.18%
25	103	101	98.06%	80	74	92.50%	98	95	96.94%	71	69	97.18%
Total	824	763		640	593		784	711		568	531	
Average			**92.60%**			**92.66%**			**90.69%**			**93.49%**

Interrater Agreement, By Scoring Institute and Rater

Table C3
Interrater Agreement by Program for 1999-2000 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	31	23	74.19%	5	321	238	74.14%
8	31	31	100.00%	8	321	319	99.38%
18	31	22	70.97%	18	321	265	82.55%
21	31	30	96.77%	21	321	309	96.26%
22	31	31	100.00%	22	321	320	99.69%
24a	31	29	93.55%	24a	321	313	97.51%
24b	31	30	96.77%	24b	321	299	93.15%
25	31	29	93.55%	25	321	310	96.57%
Total	248	225		Total	2568	2373	
Average			**90.73%**	**Average**			**92.41%**

Interrater Agreement, By Scoring Institute and Rater

Table C4
Interrater Agreement on 1999-2000 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	320	260	81.25%	56	17.50%	4	1.25%
4	320	318	99.38%	2	0.63%	0	0.00%
15	320	270	84.38%	50	15.63%	0	0.00%
16	320	318	99.38%	2	0.63%	0	0.00%
18	320	310	96.88%	10	3.13%	0	0.00%
22a	320	314	98.13%	5	1.56%	1	0.31%
22b	320	309	96.56%	11	3.44%	0	0.00%
23	320	301	94.06%	19	5.94%	0	0.00%
26	320	273	85.31%	47	14.69%	0	0.00%
27	320	296	92.50%	24	7.50%	0	0.00%
Total	3200	2969		226		5	
Average			**92.78%**		**7.06%**		**0.16%**

Interrater Agreement, By Scoring Institute and Rater

Table C5

Interrater Agreement by District for 1999-2000 Grade 8 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)*	*%*
1	153	126	82.35%	96	79	82.29%	6	5	83.33%	65	50	76.92%
4	153	152	99.35%	96	95	98.96%	6	6	100.00%	65	65	100.00%
15	153	126	82.35%	96	82	85.42%	6	3	50.00%	65	59	90.77%
16	153	153	100.00%	96	95	98.96%	6	6	100.00%	65	64	98.46%
18	153	150	98.04%	96	92	95.83%	6	6	100.00%	65	62	95.38%
22a	153	150	98.04%	96	95	98.96%	6	6	100.00%	65	63	96.92%
22b	153	148	96.73%	96	92	95.83%	6	5	83.33%	65	64	98.46%
23	153	139	90.85%	96	92	95.83%	6	6	100.00%	65	64	98.46%
26	153	123	80.39%	96	88	91.67%	6	4	66.67%	65	58	89.23%
27	153	146	95.42%	96	84	87.50%	6	5	83.33%	65	61	93.85%
Total	1530	1413		960	894		60	52		650	610	
Average			**92.35%**			**93.13%**			**86.67%**			**93.85%**

Table C6
Interrater Agreement by Program for 1999-2000 Grade 8 External Assessments

Costructed-Response Item	**Conventional** Assessments *(N)*	Total Agreement *(N)*	*%*	Costructed-Response Item	***Mathematics in Context*** Assessments *(N)*	Total Agreement *(N)*	*%*
1	35	28	80.00%	1	285	232	81.40%
4	35	35	100.00%	4	285	283	99.30%
15	35	30	85.71%	15	285	239	83.86%
16	35	35	100.00%	16	285	283	99.30%
18	35	35	100.00%	18	285	275	96.49%
22a	35	34	97.14%	22a	285	280	98.25%
22b	35	33	94.29%	22b	285	275	96.49%
23	35	34	97.14%	23	285	267	93.68%
26	35	33	94.29%	26	285	239	83.86%
27	35	35	100.00%	27	285	261	91.58%
Total	350	332		Total	2850	2634	
Average			**94.86%**	**Average**			**92.42%**

Interrater Agreement, By Scoring Institute and Rater

Table C7

Interrater Agreement for 1999-2000 External Assessment by Question Across Grades 7 and 8

Context	7th Grade		8th Grade		Average Grades 7 & 8 Agreement
	Item	Agreement	Item	Agreement	
1	*5*	74.15%	*1*	81.25%	77.53%
2	*8*	99.43%	*4*	99.38%	99.40%
3	*25*	96.31%	*23*	94.06%	95.24%
4a	*24a*	97.16%	*22a*	98.13%	97.62%
4b	*24b*	93.47%	*22b*	96.56%	94.94%
5	*18*	81.53%	*26*	85.31%	83.33%
6a	*22*	99.72%	*16*	99.38%	99.55%
7	*21*	96.31%	*18*	96.88%	96.58%
8	--	--	*15*	84.38%	84.38%
9	--	--	*27*	92.50%	92.50%