Technical Appendix

Sampling and Data Collection for the 2002-03 Data

This project was initially piloted in Washington State during the 2002-03 school year. Eight assignments, four typical and four challenging, were collected from 48 teachers (24 English/language arts and 24 math) over the course of the school year. Student work associated with the teacher assignments was also collected three times during the year from a random sample of students predetermined by the researchers and blind to the teachers.

The teacher assignments and student work were scored in summer 2003 by using rubrics developed by the American Institutes for Research and SRI International, based on the rubrics used in the study of the Chicago Public Schools by Fred Newmann, Tony Bryk, and others (Newmann, Lopez, & Bryk, 1998; Bryk, Nagaoka, & Newmann, 2000; Newmann, Bryk, & Nagaoka, 2001).

School Selection

Eight public schools and five alternate public schools that planned to undergo conversion into smaller schools in 2003-04 were identified by Fouts and Associates to participate in the study. The rationale for school selection included a combination of the following school factors: (1) large size (allowing for a greater number of teachers eligible to participate), (2) reasonable likelihood of success converting to small schools (in the opinion of the team evaluating this reform effort), (3) history of administrative and teacher cooperation, (4) significant level of district support, (5) range in student ethnic diversity, and (6) range of geographic locations around the state. Two selected schools are in districts that have the Bill & Melinda Gates Foundation Model Districts grant. The remaining schools are involved in the Bill & Melinda Gates Foundation Achievers Program.

The eight schools that participated in the study are:

Clover Park High School
Henry Foss High School
North Central High School
Port Angeles High School
A. C. Davis High School
Foster High School
Mount Tahoma High School
West Valley High School

Teacher Selection

Teachers were eligible if they: (1) taught English or mathematics to sophomore students; (2) had a class that consisted of mostly sophomore students, and at least 25% of all sophomores took that level of coursework; and (3) were likely to be teaching the same or similar types of courses during the 2004-05 school year.

Overall, 48 teachers (24 English/language arts and 24 math) from the 8 schools initially agreed to participate in the study. However, one mathematics teacher declined any further participation after submitting only one assignment, and one English/language arts teacher submitted all teacher assignments but did not submit any student work. Depending on school size, four to eight teachers represented each school.

Student Selection

All teacher participants sent their sophomore students a letter describing the study and giving the students and their parents the opportunity to choose not to participate. Teachers submitted a class list to their building coordinator after removing the names of the nonsophomore students and the names of the students who opted out. Researchers used this list to randomly select those students who would participate in the study. Depending on the number of sophomores in a class, 6 to 12 students were randomly selected to be participants, and 2 alternates were selected if available. The names of the participating students and alternates were not revealed to the teachers.

Data Collection

Assignments were collected eight times over the course of the year—four times in the first semester and four times in the second semester. Student work was collected as well during three of the collection dates. Teachers turned in student work for all students in the class to a coordinator at the school, and the coordinator made copies of the work of the randomly selected sample students and sent them to the data collector. All three student work data collections occurred during the second quarter of the school year. A single quarter was chosen to ensure that selected students did not transfer out of classes during quarter or semester breaks and to maximize the likelihood that they would not move during this data collection. The collection dates for the assignments were:

Typical Assignments	*Challenging Assignments*
December 2 – 13, 2002[2]	September 9 – November 1, 2002
January 6 – 17, 2003[1]	November 11, 2002 – January 24, 2003[2]
March 3 – 14, 2003	February 3 – March 28, 2003
May 19 – 30, 2003	April 7 – May 30, 2003

Datasets

The data collection produced two datasets (one for assignments and one for student work) for each of the two subjects, English/language arts and math. Teacher feedback information is included in the student work

[1] Student work was collected in conjunction with these assignments.

database. The English/language arts assignment database has data on 177 assignments, and the student work database has data on 399 pieces of student work. The mathematics assignment database includes 184 mathematics assignments, and the student work database contains 425 pieces of student work. Each piece of student work can be linked to a teacher assignment.

Estimation Procedures

One goal of this project is to build on the work of Newmann et al. and their study of assignments and student work in the Chicago Public Schools (Newmann, Lopez, & Bryk, 1998; Bryk, Nagaoka, & Newmann, 2000; Newmann, Bryk, & Nagoaka, 2001). To that end, the estimation procedures described here are based on the procedures used by the Chicago researchers. There are two parts to these analyses. First, a Many-Facet Rasch Model (MFRM) analysis is used to combine the scores for the individual rubrics for each assignment (or piece of student work) into a single score of quality for that assignment. The second part of the analyses uses hierarchical linear modeling (HLM) conducted at the classroom level to examine the relationship between characteristics of the classroom (e.g., teacher background, student compositional variables) and the rigor of teacher assignments, as well as the relationships among the rigor of assignments, the quality of students' work, and jurisdiction-sponsored standardized tests. For the purposes of this report, only the procedures associated with the MFRM analysis will be discussed. (The procedures for the HLM analyses will be discussed in a subsequent paper to be written for the 2004 AERA conference in April.)

Analytic Approach Rationale: Many-Facet Rasch Measurement

Each assignment and piece of student work received a score on each of three to five rubrics (the number of rubrics depends on the subject area and whether the article being scored was an assignment or a piece of student work). The rubrics and score scales for them are shown in Table A.1.

Table A.1: Scoring Rubrics and Score Scales

English/Language Arts	
	Score Scales
Assignments[2]	
1. Construction of Knowledge	1-3
2. Elaborated Communication	1-4
3. Authentic Audience	1-3
4. Student Involvement	1-4
Student Work	
1. Construction of Knowledge	1-3
2. Elaborated Communication	1-4
3. Language Conventions	1-6
Teacher Feedback	1-4

Mathematics	
	Score Scales
Assignments	
1. Important Mathematical Content	1-4
2. Problem Solving and Reasoning	1-4
3. Effective Communication	1-3
4. Relevant Contexts and Connections	1-4
5. Student Involvement	1-4
Student Work	
1. Conceptual Understanding	1-4
2. Procedural Knowledge	1-4
3. Problem Solving and Reasoning	1-4
4. Effective Communication	1-4
Teacher Feedback	1-4

In addition, all assignments were scored again by a second scorer for each of the rubrics, as a check on the reliability of the scoring. Approximately 50% of the student work in English/language arts and 40% of mathematics work was also double-scored. As a result, there are quite a few pieces of data for each assignment and each piece of work from the multiple rubrics and scorers, and it is more useful for the analysis if the data can be combined into a single score for each assignment and a single score for each piece of student work.

[2] The ELA assignment and work rubrics also allowed scorers to indicate if there was insufficient information to assign a score.

What is the best way to go about combining the data? A simple average or the sum of the raw ratings is not adequate for these purposes, because of two factors that are sources of variability in the raw ratings. The first factor relates to differences in the severity of the scorers: one scorer may have higher standards than another. The second is associated with differences in the stringency of the rubrics. For example, it may be harder for an English/language arts assignment to achieve a top score in Construction of Knowledge than in Elaborated Communication.

Had *all* the assignments been rated by *all* the raters on *all the rubrics*, then the simple average of the ratings would have balanced out any differences in rater severity. However, such a massive rescoring activity was not feasible. Thus, we need to adjust statistically for the differences in the severity of the scorers and the stringency of the rubrics. We use the Many-Facet Rasch Measurement (Linacre, 1989a) technique to combine the individual raw scores from both scorers on each assignment or piece of work and, ultimately, to develop numeric scales to quantify the intellectual challenge of assignments and the overall quality of the student work. Scales are developed separately for the assignments and student work and for English/language arts (ELA) and mathematics.

The presence of differences in rater severity and in rubric stringency is one of the main reasons that this Many-Facet Rasch Model calibration step is important. It adjusts for rater severity, in terms of the estimated measure for each assignment or piece of student work. Likewise, it adjusts for the difficulty of the rubrics.

Theoretical Model

The Many-Facet Rasch Model used for assignments is:

$$\log\left(\frac{P_{nijk}}{P_{nji(k-1)}}\right) = B_n - C_i - D_j - F_{ik}$$

where
P_{nijk} is the probability of assignment n being given a rating of k on rubric i by scorer j
$P_{nij(k-1)}$ is the probability of assignment n being given a score of $k-1$ on rubric i by scorer j
B_n is the parameter for assignment n (quality of the assignment)
C_i is the parameter for rubric i (stringency of the rubric)
D_j is the parameter for scorer j (severity of the scorer)
F_{ik} is the parameter for receiving a rating of k relative to $k-1$ on rubric i (step difficulty).

The model for student work is:

$$\log\left(\frac{P_{nijk}}{P_{nji(k-1)}}\right) = B_n - C_i - D_j - F_{ik}$$

where
P_{nijk} is the probability of student work *n* being given a score of *k* on rubric *i* by scorer *j*
$P_{nij(k-1)}$ is the probability of student work *n* being given a score of *k-1* on rubric *i* by scorer *j*
B_n is the parameter for student work *n* (quality of the work)
C_i is the parameter for rubric *i* (stringency of the rubric)
D_j is the parameter for scorer *j* (severity of the scorer)
F_{ik} is the parameter for receiving a score of *k* relative to *k-1* on rubric *i* (step difficulty).

The product of the analysis is the measure of each element of three facets: the assignment rigor and authenticity (or student work quality), B_n; the rubric stringency, C_i; and the scorer severity, D_j (as well as the measure of step difficulty, F_{ik}, which is an output of the model in which we are less interested). The Many-Facet Rasch Model analysis corrects the estimates of assignment rigor and authenticity and the quality of student work for scorer severity and rubric difficulty. The Rasch-adjusted measures for assignments and student work, and their associated standard error estimates, will then be used as data for the HLM analyses. Parameters for ELA assignments, ELA student work, mathematics assignments, and mathematics student work will each have their own scales, which will not be linked to the scales of the others.

For example, at the end of the Many-Facet Rasch Model analysis, each scorer will have a parameter estimate to quantitatively represent his or her severity. Likewise, there will be a stringency parameter associated with each rubric and a parameter for each teacher assignment. All of these parameters are placed on a common scale so that they can be compared with each other.

There is also one teacher feedback rubric for student work. Because teacher feedback has only one rubric, it is not included in the Rasch measurement analysis but is examined and reported separately.

Rescaling

The default setting of the FACETS program (Linacre, 1989b), which performs the Rasch analysis, chooses the local origins of scales such that the mean calibrations of the scorers, the rubrics, and the scoring scale structure are all zero. As a result, the local origins of the quality of assignments and student work are defined by the model; that is, the mean of the rater measure is zero, the mean of the rubric measure is zero, and the mean of the step difficulties of each rubric is also zero. However, this means that the student work and assignment measures are not zero.

In other words, FACETS sets the origins as the following, in relation to the Rasch Model described above:

$$\sum_{i=1}^{m} C_i = 0\,,\ \sum_{j=1}^{12} D_j = 0\,, \text{ and } \sum_{k=1}^{\ell} F_{ik} = 0\,,$$

where *m* is the number of rubrics (which is different for each of the four groups: ELA assignments, ELA student work, mathematics assignments, and mathematics student work), 12 is the number of scorers, and ℓ is the number of score categories for rubric *i*. Because the means of *C*, *D*, and *F* are already determined, the mean of *B* (the assignment or student work parameter) is not constrained to equal zero.

Because the logit measure theoretically ranges from negative infinity to positive infinity, it is not a scale that is easy to interpret. For reporting purposes, we rescale the logit measure to a 0 to 10 scale. The transformation formula is:

Assignment (or student work) measure =
10 x (logit measure – min)/(max – min)

where logit measure is the original measure for either assignment or student work, min is the minimum value of the logit, and max is the maximum value of the logit. By the same operation, the estimated standard error is also transformed to the same scale by the formula:

Standard error = original standard error x 10 (max – min)

Overall Reliability of the Measures

One of the important questions to ask regarding this Many-Facet Rasch Model analysis is to what extent the estimated Rasch scores (based on raw scores assigned by using the scoring rubrics) provide reliable measures of the rigor and authenticity of teacher assignments and the quality of student work. This question can be answered by examining the reliability estimates produced by the FACETS program for the assignments and student work. The reliability calculated by FACETS is the Rasch equivalent to the KR-20 or Cronbach Alpha statistic, the ratio of the true variance to the observed variance. From the FACETS output, the reliability statistics are 0.85 for English/language arts assignments, 0.78 for English/language arts student work, 0.70 for mathematics assignments, and 0.62 for mathematics student work. Given the number of scored assignments and pieces of work and the number of score levels, these reliabilities are typical compared with other “tests” of similar length, while a reliability of 0.85 is considered very good by common standards.

We see evidence that English/language arts shows higher reliability than math, and assignments show higher reliability than student work. The reliability of the models ranges from mathematics student work, where the model accounts for 62% of the variance in the data, to ELA assignments, where the model accounts for 85% of the variance. The square root of the reliability is an estimate of the correlation between the true score and the observed score, ranging from 0.79 (i.e., the square root of the reliability for mathematics student work, 0.62) for mathematics student work to 0.92 (i.e., the square root of the reliability for ELA classroom assignments, 0.85) for ELA assignments. The reliability estimates indicate that the Rasch measures correlate highly with the true scores of the student works and assignments.

Facet 1: Rasch Scores as Valid Measures of Assignments and Student Work

On the basis of the high reliability coefficients discussed above, it appears that forming a single score for the assignments or student work based on the rubrics is reasonable. In addition to these reliability requirements, it is also important to examine how well the individual rubrics relate to one another: are measures that are expected to tap the same higher-order construct correlated, or do they seem to behave like independent measures? To assist in the interpretation of the Rasch measures, it is also important to understand how the individual rubrics correlate with the aggregate measure produced by the Rasch analysis.

To get a sense of the interdependence among the rubrics, we looked at correlations among the raw scores for each of the rubrics and the Rasch measures. Table A.2 shows the resulting correlation matrix.

The rubrics all have positive and highly statistically significant correlations with each other and with the Rasch measures, which increases our confidence that a high score on the Rasch measure tends to represent high raw scores on the individual rubrics that make up the scale. The Rasch measure correlates more highly with each of the raw ratings than the raw ratings correlate with each other, suggesting that the Rasch measure does a good job of summarizing the separate raw measures, and enabling us to use a single Rasch measure in place of the separate measures.

Table A.2: Correlations among Rasch Measures and Raw Scores on Individual Rubrics (p-values shown below correlations)

ELA Assignments N=177	Measure	CK	EC	AA	SI
Construction of Knowledge	0.67	1			
	<0.0001				
Elaborated Communication	0.80	0.58	1		
	<0.0001	<0.0001			
Authentic Audiences	0.58	0.26	0.32	1	
	<0.0001	0.0005	<0.0001		
Student Involvement in Crafting Assignments	0.66	0.29	0.46	0.37	1
	<0.0001	<0.0001	<0.0001	<0.0001	
ELA Student Work N=399	**Measure**	**CK**	**EC**	**LC**	
Construction of Knowledge	0.78	1			
	<0.0001				
Elaborated Communication	0.82	0.58	1		
	<0.0001	<0.0001			
Language Conventions and Resources	0.88	0.58	0.70	1	
	<0.0001	<0.0001	<0.0001		

Mathematics Assignments N=184	Measure	IMC	PS/R	EC	RC
Important Mathematics Content	0.65	1			
	<0.0001				
Problem Solving and Reasoning	0.74	0.39	1		
	<0.0001	<0.0001			
Effective Communication	0.55	0.27	0.37	1	
	<0.0001	0.0003	<0.0001		
Relevant Context and Real-World Connections	0.52	0.20	0.36	0.22	1
	<0.0001	0.0057	<0.0001	0.0029	
Student Involvement in Crafting Assignments	0.42	0.23	0.27	0.30	0.18
	<0.0001	0.0016	0.0002	<0.0001	0.0133
Mathematics Student Work N=425	**Measure**	**CU**	**PK**	**PS/R**	**EC**
Conceptual Understanding	0.63	1			
	<0.0001				
Procedural Knowledge	0.63	0.14	1		
	<0.0001	0.0036			
Problem Solving and Reasoning	0.61	0.52	0.12	1	
	<0.0001	<0.0001	0.0172		
Effective Communication	0.67	0.40	0.28	0.46	1
	<0.0001	<0.0001	<0.0001	<0.0001	

Facet 2: Rubric Stringency

The Rasch measure of the stringency of the rubrics indicates the difficulty level for each rubric; the higher the measure, the more difficult the rubric—that is, the harder it is to get a high score on that rubric. In this section, we discuss the rubric stringency measure.

English/Language Arts Assignments

The most stringent rubric for ELA assignments is Authentic Audiences and the least stringent rubric is Construction of Knowledge. The measures of the rubrics for the ELA assignments are:

Construction of Knowledge	-1.74
Elaborated Communication	-1.07
Authentic Audiences	1.84
Student Involvement in Crafting Assignments	0.98

English/Language Arts Student Work

The measures of the rubrics for ELA student work are:

Construction of Knowledge	-0.44
Elaborated Communication	-0.38
Language Conventions and Resources	0.83

Mathematics Assignments

The most stringent rubric for mathematics assignments is Student Involvement in Crafting the Assignments, with an MFRM parameter of 0.78. The least stringent is Problem Solving and Reasoning, with the lowest MFRM parameter of -0.78. The measures of rubrics for the mathematics assignments are:

Important Mathematical Content	-0.30
Problem Solving and Reasoning	-0.78
Effective Communication	-0.03
Relevant Contexts and Real-World Connections	0.34
Student Involvement in Crafting Assignments	0.78

Mathematics Student Work

The measures of the rubrics for mathematics student work are:

Conceptual Understanding	0.22
Procedural Knowledge	-1.87
Problem Solving and Reasoning	0.37
Effective Communication	1.28

Facet 3: Rater Severity

Table A.3 lists the rubrics and illustrates the differences in scores given by different scorers for each rubric for those assignments and student work that were scored by two raters. The first column shows the percentage of observations that received the same rating by the two scorers (i.e., perfect agreement), and the second column shows the percentage of observations that were scored within one point. The score scales for the different rubrics range from 3 to 6 points. The English/language arts scorers had perfect agreement of 60% or more on all four of the assignments rubrics, all three of the student work rubrics, and the teacher feedback score (column 1). They had at least 90% agreement within one point for all but one rubric (column 2). There was much more variation among the mathematics scorers, particularly for assignments, which ranged from 44% perfect agreement on Important Mathematical Content, to 91% perfect agreement on the Student Involvement rubric. However, the percentages of agreement within one point for mathematics scores are comparable to those of English/language arts. The teacher feedback ratings between scorers are quite consistent for mathematics, with 100% agreement within one point.

Differences among raters could be due to a variety of factors, such as the clarity and specificity of the scoring rubrics (including the clarity and specificity of the benchmarks used in the rubrics), the effectiveness of the scorer training, and the nature of the assignments themselves.

Table A.3: Agreement Rates on Assignments and Student Work, by Rubric (for assignments and work rated by two scorers)

English/Language Arts	**Perfect Agreement**	**Agreement within 1 point**
Assignments		
Construction of Knowledge	64%	96%
Elaborated Communication	60%	93%
Authentic Audience	69%	94%
Student Involvement in Crafting Assignments	67%	79%
Student Work		
Construction of Knowledge	63%	94%
Elaborated Communication	60%	90%
Language Conventions and Resources	69%	91%
Teacher Feedback	67%	95%

Mathematics	**Perfect Agreement**	**Agreement within 1 point**
Assignments		
Important Mathematical Content	44%	92%
Problem Solving and Reasoning	52%	88%
Effective Communication	72%	99%
Relevant Contexts and Real-world Connections	68%	93%
Student Involvement in Crafting Assignments	91%	99%
Student Work		
1. Conceptual Understanding	68%	89%
2. Procedural Knowledge	45%	79%
3. Problem Solving and Reasoning	83%	93%
4. Effective Communication	73%	99%
Teacher Feedback	92%	100%

From the MFRM, the scorer measures indicate the severity of each scorer. The more severe scorer gives consistently lower scores to assignments of the same quality, compared with the less severe scorer. The ranges of the scorer measures (i.e., the differences between the most and least severe scorers) are 0.83, 0.43, 1.65, and 0.98 standard deviations for English/language arts assignments, English/language arts student work, mathematics assignments, and mathematics student work, respectively. The ranges show that the raters for the mathematics teacher assignments have the largest disparity in severity, a difference of 1.65 standard deviations. The scores for the English/language arts student work have the smallest range in severity (i.e., 0.43 standard deviation). The wide disparity in the severity for the scores for mathematics teacher assignments may cause concerns about fairness in scoring. This is especially worrisome when not all scorers rate all assignments. In the future, the scorer training will attempt to close the severity gap. In addition, the FACETS output indicates that raters differ in severity at different rubric score levels. Thus, rater training will also focus on decreasing the possibility of differences between scorers and levels of scoring on various rubrics.

Future Data Collections

The schedule for future data collections is shown in Table A.4. In 2003-04 we are collecting assignments and work from 12 small start-up high schools and 4 additional pre-conversion schools. In 2004-05 we will again collect data from these same 12 small start-up high schools so we can examine changes over time in these schools. In addition, we will collect data from 8 large traditional high schools so these data can be compared data for to the 12 small start-ups. The 8 Washington pre-conversions from the pilot year will have converted into small learning communities and will be in their second year of conversion. We expect to collect data from 2 small learning communities for each of the original 8 pre-conversions and plan to use the same teachers, if possible.

In the final year of data collection, we will follow up on the 4 pre-conversions from Year 1; by then they will have been in conversion for over a year. As with the Washington schools, we will collect data from 2 of the smaller learning communities from each of the 4 originally large schools, and attempt to get assignments from the same teachers.

Table A.4: Data Collection Schedule

School Type	**Pilot Year (2002-03)**	**Year 1 (2003-04)**	**Year 2 (2004-05)**	**Year 3 (2005-06)**
Startups		12	12 (same as yr 1)	
Comparison			8	
Pre-conversion	8 WA	4 non-WA		
Conversion			16 WA[3]	8 non-WA[4]
TOTAL	**8**	**16**	**36**	**8**

[3] Two small learning communities from each of the eight Washington schools.

[4] Two small learning communities from each of the four converting schools outside Washington State.