Assessing the Quality of Reading Comprehension Assignments and Student Work

Lindsay Clare Matsumura, Mikyung Kim Wolf, Amy Crosson

Allison Levison, Maureen Peterson, Lauren Resnick

University of Pittsburgh, Learning Research and Development Center

Brian Junker

Carnegie Mellon University

Paper presented at the Annual Meeting of the American Educational Research Association Meeting, San Diego, April 2004

Contact Information:

Lindsay Clare Matsumura

University of Pittsburgh, Learning Research and Development Center

3939 O'Hara Blvd. Room 801

Pittsburgh, PA 15260

(412) 624-7594

lclare@pitt.edu

ABSTRACT

This study presents preliminary findings from research developing an instructional quality assessment (IQA) that could be used to monitor the influence of reform initiatives on students' learning environments and to guide professional development efforts within a school or district. This paper focuses specifically on the portion of the IQA used to evaluate the quality of teachers' reading comprehension assignments and student work. While results are limited due to a very small sample of participating teachers (N = 13, 52 assignments), results indicated a moderate level of inter-rater agreement and a good degree of consistency for the dimensions measuring academic rigor, but not the clarity of the teachers' expectations. The rigor of the assignments collected from teachers also was associated with the rigor of observed instruction. While our rubrics appeared to capture meaningful differences in quality between assignments in terms of the learning opportunities afforded to students, collecting four assignments from teachers did not yield a stable estimate of quality. Individual teachers varied quite a bit with regard to the quality of the assignments they submitted—some poor, some fair and some good. This suggests that the way in which assignments are collected from teachers should be revised. Implications for professional development also are discussed.

As described in the introductory paper of this symposium (Junker et al, 2004), the primary goal of the Instructional Quality Assessment (IQA) toolkit is to develop a set of measurement tools that provide a rich picture of instructional quality, and have the potential to serve as a *learning tool* (Sheppard, 2000) for district and school personnel (principals, teachers, etc.). Without sacrificing richness, however, the IQA also is intended to be a toolkit that is reasonably parsimonious to use. In other words, our goal is to create a set of measures that can be used to assess instructional quality within a reasonable period of time and at a reasonable cost. These somewhat competing goals form the central challenge of our project: How can a measure of instructional quality be created that is

"rich" in content and at the same time "lean" enough to be used in large scale research and evaluation studies?

To address this challenge, we measured instructional quality from multiple perspectives. As described in Wolf et al's (2004) paper earlier in this symposium, teachers were observed once in their classroom. This observation is intended to provide, among other things, insight into students' opportunities to develop their academic language skills and engage with rigorous content material. A single observation, however, does not reveal enough about a teacher's classroom practice to be considered an adequate assessment of instructional quality. We decided, therefore, to also collect a sample of assignments with student work from teachers in order to gain a more multi-faceted perspective on the quality of instruction.

This paper describes our work so far looking at the quality of the assignments we collected from teachers. This work was conducted in two content areas: reading comprehension and mathematics. The purpose of this paper is to describe the work undertaken in reading comprehension, and specifically, students' responses to literature[1]. While other aspects of reading comprehension instruction also are important to study, for example, the support students receive to decode text, develop their vocabulary, etc., we focused on students' responses to literature as this would be a likely area (or genre) for them to demonstrate higher level academic skills.

The first part of this paper describes the research and theories that underlie the development of our rubrics, or how we went about determining the degree to which an assignment task supports students' engagement in meaningful, challenging work. Preliminary results from a small pilot study then are described that focus on the technical quality of these rubrics. Assignment quality also is investigated from a more qualitative perspective to look at the degree to which our rubrics may capture important differences in students' opportunities to learn. Specifically, the following questions are addressed:

1. How reliable and independent are the classroom assignment rating scales?
2. How many assignments and raters might be needed to obtain a stable estimate of the quality of classroom practice?
3. What is the relation of the classroom assignment ratings and observed instruction?
4. What was the quality of the assignments we collected? Do differentially rated assignments (that is, assignments that are considered to be of high, medium, or low quality) provide qualitative differences in students' opportunity to learn?

Assignment Tasks as an Indicator of Instructional Quality

"People learn by doing" is an old and familiar maxim. A more up-to-date version, informed by 30 plus years of research on learning and instruction, adds a role for teachers: "People learn by doing with guidance and assistance." This

[1] Our work conducted in mathematics will be described in the following paper by Boston et al. (2004)

view of learning is rooted in the theoretical work of Lev Vygotsky and is supported by research focused on children's development across diverse cultures. This body of research indicates that children learn skills—such as weaving, sewing, cooking, etc.--by jointly participating in activities with an adult (or other more capable peer). These adult mentors "scaffold" children's participation in the activity by orienting children to the overall goals of a task, breaking the activity down into manageable parts, and focusing children's attention and actions on the steps required to complete the activity. Adult mentors also support and guide children's participation in an activity by demonstrating and modeling the act to be performed, and "marking critical discrepancies between what the child has produced and the idealized version of the activity" (cited in Rogoff, p. 94). Through engaging children in the appropriate handling of a task, adults "create situations in which children can extend current skills and knowledge to a higher level of competence" (Rogoff, 1990, p. 93). In other words, adults open "zones of proximal development" for children by allowing them to do with assistance what they would not be able to do on their own (Vygotsky, 1978).

In order to become powerful abstract thinkers and consumers of texts, students need the opportunity to participate in social interactions where analytical and abstract thinking is modeled for them and where they have the opportunity to practice their emerging skills in this area. Verbal interactions, or classroom conversations are of critical importance for providing students with

the opportunity to be exposed to modeling of these types of skills, as well as to practice their thinking and reasoning skills and get immediate feedback on their efforts. But classroom conversations alone, however excellent they may be, are not enough to develop students' comprehension skills. Students also need the opportunity to apply their newly emerging thinking and reasoning skills in written forms as well. This is important for monitoring student learning. Additionally, students who have not had the opportunity to develop these skills (the ability to write about text in meaningful ways) are at a distinct disadvantage academically—a disadvantage that limits their chances of being successful in high school, being accepted to a college, and completing college-level course work.

Assignment tasks provide insight into the level and type of support (or scaffolding) a teacher provides to students, and so can an important source of information for assessing students' opportunity to learn academic skills. Assignment tasks can be a window into the degree to which a teacher makes a task accessible to students (e.g. breaks down the steps of a task or provides explicit directions for how to complete each step), communicates performance expectations (e.g. demonstrates an idealized version of the act to be performed), and provides feedback to students on their efforts (e.g. marks critical features of discrepancies between what a child has produced and the ideal solution).

Assignment tasks also can provide insight into students' opportunities to learn skills and content that are germane to a specific discipline. As described

earlier, children learn through joint-participation in real activities. They learn to weave by weaving (with assistance), and to cook by cooking (with assistance). By the same token, children learn academic skills by being assisted to engage in the work of real scholars. This could mean using mathematics to solve real world problems that contain multiple solutions (as a scientist or an engineer might). This could also mean synthesizing, analyzing, interpreting, and evaluating information from texts (as would be required in college).

The question is, how does one determine the degree to which an assignment supports and guides students to develop higher-level academic skills? What would one look for in an assignment to indicate that students had been exposed to a high quality classroom learning environment?

Defining Assignment Quality

To answer this question we drew in part on research investigating best practices for teaching reading comprehension and research investigating assignment quality in English language arts, in addition the Principles of Learning (Institute for Learning, 2002). The following sections provide a brief overview of this research.

Some elements of effective reading comprehension instruction. The ability to comprehend text is a very complex process. In order to become proficient readers students have to master a number of interrelated skills. These include the ability to construct mental models of a text at various levels, for example, understanding how clauses are related, or how events in a text are temporally

sequenced. Besides understanding the specific events or ideas represented in a text, however, proficient readers also are able to construct meaning beyond what is represented on the written page. In other words, they are able to apply higher-order thought processes to infer meaning beyond the surface-level features of the text.

Borrowing from Bloom (1956) the complexity of thought processes one could use to infer meaning from a text could be described in three general levels (described in Snow, 2002, p. 109). At the lowest level are *recognition* and *recall* or the ability to identify specific content verbatim and to reproduce (remember and retrieve) specific content that was explicitly mentioned in a text. The second level is termed *comprehension* and includes the ability to generate a mental model of a text by summarizing, paraphrasing, explaining, or translating a text. At the highest level of complexity are the *application* of knowledge from a text to solve a problem not mentioned in the text, and the *analysis* of a text into its constituent parts that are linked back to each other in new ways. The ability to *synthesize* or construct new patterns or structures from the events in a text, and the *evaluation* of a text based on an external criteria or standard also are considered to be high-level thinking skills.

Effective reading comprehension instruction supports students to answer higher-level questions about a text (Snow, 2002). In addition to understanding the surface level features of a story (i.e. constructing a mental model of the events), effective reading comprehension instruction provides students with an

opportunity to construct meaning beyond what is represented on the page. Instruction of this type guides students to analyze, synthesize, evaluate or apply knowledge from a text in the service of more deeply comprehending what they read.

Another element of effective reading comprehension instruction concerns the curricular materials used by a teacher. Ideally students would be exposed to a curriculum that is in depth and challenging, and exposes students to a wide variety of genres (Snow, 2002). This includes having students read texts that contain themes or ideas that are complex enough (or illustrate sufficient "grist") to support meaningful writing topics and classroom discussions (Beck, McKeown, Hamilton, & Kucan, 1997). Such texts could convey information to students about other places, times, or cultures. These texts also could contain interesting dilemmas where there is no obvious right or wrong answer, or other themes that broaden students' thinking. Grist could also be evidenced in the writer's craft, for example, in the language use, vocabulary and organizational structures employed by an author.

Finally, effective reading comprehension instruction exposes students to a wide variety of intentionally applied comprehension strategies (National Reading Panel, 2000). Specifically, effective teachers guide the reader and/or model for the reader the actions that the reader needs to take to improve comprehension. She does this by clearly explaining the reason for the task (and standards for completing the task), breaking the task down in smaller parts for

their students and activating prior knowledge. Teachers then have students practice these strategies and provide them with assistance and feedback on their performance until the student internalizes the skill and is able to independently carry out the comprehension task on their own. This type of explicit instruction appears to be especially beneficial for lower-achieving students (Snow, 2002, p. 33).

Research in assignment quality. Most of the research on assignment quality has included a focus on students' opportunity to apply higher-level thinking skills. For example, Stein, Smith, Hennigsen, and Silver (2000) considered mathematics tasks that required students to recall information only, or apply an algorithm or procedure without any reference to an underlying mathematical concept to represent a lower level of cognitive demand. Tasks that had students apply procedures and engage with the underlying conceptual ideas, or that required students to apply complex problem-solving (i.e. non-algorithmic) thinking were considered to represent a higher level of cognitive demand.

With specific regard to English language arts, two separate efforts comprise the bulk of the research in assignment quality: studies connected by the Chicago Consortium for Quality Schools, and the National Center for Research in Evaluation, Standards and Student Testing at UCLA. Fred Newmann, Anthony Bryk and their colleagues in Chicago defined high quality assignments as "authentic intellectual work." They operationalized the characteristics of

authentic intellectual work in three scales. The first scale, construction of knowledge, focused on the extent to which an assignment required students to organize, interpret, evaluate, or synthesize prior knowledge to solve new problems. The second scale, disciplined inquiry, focused on the extent to which the assignment required students to use a prior knowledge base, strive for in-depth understanding, and express their ideas with elaborated communication. The last scale, value beyond school, focused on the applicability of the task to life outside the school setting (Newmann, et al., 2001).

Similar to Newmann and Bryk's work, (Clare) Matsumura (the first author of this paper) and her colleagues at CRESST looked at assignment quality in terms of its level of cognitive challenge, or the degree to which students had the opportunity to apply higher order reasoning, engage with academic content material and produce extended responses. They did not consider the applicability of a task to contexts outside of schools. They also looked at how clearly a teacher articulated the specific skills, concepts or content knowledge students were to gain from completing the assignment in order to ascertain teachers' intentions for a task (specific learning versus activity for activity's sake). The clarity and specificity of the grading criteria used to assess students' work, and the alignment between the learning goals and the assignment task, and the learning goals and the grading criteria, were considered as well. The purpose of these dimensions was to produce more diagnostic information about assignment quality that could be used to guide professional development efforts. In other

words, to consider at what point in the assignment activity (e.g. the conception, implementation, assessment of student performance, etc.) teachers might need additional support (Matsumura, Garnier, Pascal, & Valdés, 2002; Clare & Aschbacher, 1999).

Results from both of these two projects indicated that students produced higher quality work and scored higher on standardized tests of achievement when they were exposed to higher-quality assignments (see Newmann, Bryk, & Nagaoka, 2001; Clare & Aschbacher, 2001; Matsumura, Garnier, Pascal & Valdés, 2002). This is after controlling for students' SES, ethnicity, language status and prior level of achievement. These results supported the decision to include measures of assignment quality in the IQA toolkit as it appears that teachers' assignments have the potential to yield important information about the quality of classroom practice that is associated with differential student achievement.

IQA assignment quality dimensions. As described in Junker et al (2004), the IQA rubrics are structured around the Principles of Learning (Institute for Learning, 2002). These principals are a comprehensive, standards-based framework that includes both instructional processes and the external supports intended to support high-quality teaching and learning. They are comprised of nine interrelated constructs (see Appendix A).

We focused on two Principals of Learning in developing our rubrics that we believed would be most proximal to assignment quality. These are *academic rigor* and *clear expectations*. The principal of learning, *academic rigor*, holds that

student success depends on their exposure to a rich knowledge core that is organized around the mastery of major concepts. This curriculum also should provide students with the regular opportunity to pose and solve problems, formulate hypotheses, justify their reasoning, construct explanations, interpret text, and test their own understanding. Additionally, students should have the opportunity to construct their own understandings of concepts based on the synthesis of several sources of information including their experiences outside of school.

The second principal of learning upon which we based on work, *clear expectations*, holds that students need to have access to the performance expectations for their work. Teachers can communicate these expectations to students by teachers posting or distributing standards and rubrics, for example, or by discussing with students the criteria for work that meets a specific standard. Providing students with models of high quality work that outline a sequence of expected concepts and skills students are to master in the process of accomplishing a larger standard, and discussing these models with students also are important for communicating expectations to students. Other important means for making expectations clear to students include involving students in judging their own work with respect to the standards, and communicating to parents what students are supposed to accomplish.

For the IQA toolkit these principals of learning were operationalized in five rubrics and a checklist for evaluating assignment quality. Specifically, to

assess the degree to which an assignment promoted *academic rigor in a thinking curriculum* we looked at the rigor of the text used for an assignment in terms of the complexity of its themes and content. Similar to the research in assignment quality described earlier, we also looked at the degree to which an assignment provided students with an opportunity to develop their analysis and interpretation skills, and engage with the deeper meanings of a text (i.e. go beyond describing surface-level details). The degree to which students were supported to realize the potential of the task during implementation, that is, that the collection of student work evidenced that students had analyzed and interpreted the deeper meanings of the text and supported their responses with evidence, also was assessed. Additionally, we looked at the rigor of a teacher's expectations for the quality of student work. By this we meant the degree to which a teacher's expectations (expressed in her grading criteria or assignment directions) supported students to apply higher-level comprehension skills and support their responses with extensive evidence from a text.

To assess the degree to which a teacher communicated *clear expectations* to students regarding the quality of their work we looked first at the specificity and amount of information a teacher provided to students for what they would need to do to successfully complete the task. We also considered the teacher's efforts (based on self-reported information) to ensure that all students had access to the performance expectations for a task.

The following sections describe the results of a small pilot study investigating the technical quality of these rubrics in terms of interrater reliability, stability of the assignment ratings and relation to observed instruction. These results must be interpreted with a great deal of caution, however, as they were based on a very small sample of teachers. These analyses have utility, however, for providing information that could be used to guide future development work. Additionally, variation in assignment quality is explored from a more qualitative perspective in order to take a closer look at the degree to which our ratings capture meaningful distinctions in students' opportunity to develop higher-level academic thinking and writing skills.

Methods

Sample

Second- and fourth-grade teachers (N = 30) were recruited from 11 elementary schools across two demographically similar school districts[2]. Of these 30 teachers, 14 participated in the reading comprehension portion of the study, and 13 of these teachers turned in assignments with samples of student work. These schools served a diverse population of students (26% African American, 6% Asian, 47% Latino, 15% white, 6% other) 20% of whom were English language learners. Teachers who participated in the study had been teaching for an average of 14 years, and had been at their school an average of 4 years.

[2] One third-grade teacher and one fifth-grade teacher were recruited as well.

Procedures

District personnel suggested schools that might be interested in participating in the study. A member of the IQA research team then contacted the principals of these schools. Principals who were interested in participating were asked to explain the study to all of the second- and fourth-grade teachers at their school.

In April (2003) a member of the IQA research team visited each school to discuss the study with interested teachers and distribute the assignment collection materials. Teachers were asked to submit four reading comprehension assignments—two recent assignments and two assignments they considered to be challenging for their students (N = 52 assignments). For each assignment teachers filled out a two-page cover sheet describing the assignment task, their assessment criteria for grading student work and how they shared these criteria with students. Teachers also submitted six samples of student work for each task--two samples of work they considered to be of high, medium and low quality respectively. The assignments were collected later in May when they classroom was observed. Teachers were given $100 gift certificate as a token of appreciation for completing the assignment coversheets and assembling the samples of student work.

The assignments were rated by graduate students who were recruited to participate in the data collection and were not part of the team who developed the rubrics (N = 2). We hired "naïve" raters in order to assess the quality of our

rating training program and rubrics (as evidenced by the degree to which people who were external to the project could agree on the different ratings). The assignments also were rated by members of the IQA development team (N = 4), though these ratings were not included in the analyses reported here. The assignments were randomly ordered for scoring and were rated independently by each of the naïve raters (N = 52 assignments). The dimension measuring the academic rigor of the text was rated a few weeks later by these same raters, after the research team located the relevant texts. Because of difficulty locating the books or articles, it was possible to rate this dimension for only 37 assignments.

Measures

Assignment quality is assessed on a four-point scale (1 = poor, 4 = excellent) for the following dimensions, with the exception of the dimension measuring the rigor of the text which is assessed on a three-point scale (1 = poor, 3 = excellent):

- Rigor of the Text – The purpose of this dimension is to measure the degree to which the text that is the focus of a reading comprehension assignment contains literary or informational content that is complex and engaging enough to warrant extended writing. Additionally, this dimension considers the richness and variety of the language (vocabulary and sentence structures) in the text. To receive a high score on this dimension, a text would have to contain a complex plot or elaborated information, and the text would have to contain rich or highly specific vocabulary.

• Potential of the Task – The purpose of this dimension is to describe the degree to which an assignment provides students with an opportunity to develop their analysis and interpretation skills and to engage with the deeper meanings of a text. Specifically, this dimension considers the extent to which students are supported to apply higher-level skills in the service of deepening their comprehension of a text, as opposed to recalling, describing, or identifying basic information. To receive a high score on this dimension, students would be required to go beyond surface-level description, detail, or theme identification, and to engage with subtle nuances of the text or the overarching or larger significance of the work (e.g., discussion of story themes) with the opportunity to develop and elaborate their ideas. Additionally, the task would require students to provide evidence from a text to support their ideas.

• Implementation of the Task – The purpose of this dimension is to describe the degree to which students are supported to realize the potential of the task during implementation. To receive a high score on this dimension, the collection of student work would evidence that students analyzed and interpreted the deeper meanings of the text and that students provided extensive evidence for their positions. Additionally, the collection of student work would demonstrate that students were supported to develop and elaborate their ideas through extended written response.

• Rigor of Expectations – The purpose of this dimension is to describe the degree to which a teacher's expectations for the quality of students' work

support students to analyze and interpret the deeper meanings of a text. An assignment that received a high score for this dimension would focus on students' attainment of these higher-level skills.

• Clarity and Detail of the Expectations – The purpose of this dimension is to assess the specificity and elaborateness of a teacher's expectations for the quality of students' work for the assignment task. A high score for this dimension would indicate that a teacher provided a great deal of information to students for what they would need to do to successfully complete the task. Each of the teacher's criteria for success would be clearly articulated, and within these criteria, detail would be provided for the varying levels of success (e.g. what a student would need to do to get an A, a B, etc.).

In addition to the five-point scales, teachers also completed a checklist reporting how they shared their expectations to students (e.g. discussed criteria in class, posted criteria charts, shared models of high quality work, etc.).

Analyses

Descriptive statistics were used to characterize the teachers' assignments. Cohen's kappa coefficients were calculated to investigate the level of agreement between five raters on each dimension when controlled for chance agreement. Cronbach's alpha coefficients were calculated to estimate the consistence of these ratings at the teacher level. Correlations also were computed to measure the strength of agreement between the rater pair.

Generalizability studies were conducted to investigate whether the design based on four raters and the collection of four assignments from teachers yielded a stable estimate of the overall quality of assignments. Finally, correlations were computed at the teacher level to investigate the interrelationship of the assignment ratings, and the relation of the assignment ratings to observed instruction.

Results

The results of this study (limited by the small sample size) are presented in the following sections organized by each of the research questions.

How reliable and independent are the classroom assignment rating scales?

To address the first part of this question we investigated the interrater reliability of the rating scales, the degree to which different people can independently look at the same phenomenon (in this case teachers' assignments) and agree on a score. The percent agreement between raters was calculated on assignment ratings within each grade level. Results indicated that there was a fair level of agreement between the two raters who scored the assignments for the dimensions measuring the academic rigor of the assignments (see Table 1). The percent agreement ranged from 81.1% to 63.5%, and the correlation between raters ranged from ($\underline{r}$ = .83 to $\underline{r}$ = .81) for each of the dimensions measuring the academic rigor of the assignment task. The dimension measuring clear expectations, however, had poor inter-rater agreement (44.2%), and a relatively low correlation between the raters ($\underline{r}$ = .56).

Cohen's kappa coefficients were calculated to investigate the level of agreement when controlling for chance agreement. Significant kappas for each of the academic rigor dimensions indicated that the level of rater agreement was better than chance. The magnitude of the kappas ranged from .51 to .66, indicating a moderate level of agreement between the raters for these dimensions. The exception to this pattern was, again, the dimension measuring clear expectations (kappa = .24).

Cronbach's alpha coefficients also were calculated to investigate the consistency of the ratings within each assignment for each dimension. This statistic considers the trend in rater agreement, and ranged from .88 to .91, confirming a high degree of consistency within each dimension for each assignment. The clarity of the expectations rubric showed a lower level of consistency (.71).

Table. 1

Inter-rater reliability of assignment ratings for the reading comprehension assignments (*N* = 52 assignments)

Dimension	% of exact agreement	Spearman *r*	Kappa	Alpha
AR0: Grist*	81.1	.76	.66	.87
AR1: potential	71.2	.84	.59	.91
AR2: implementation	69.2	.84	.56	.88
AR3: expectations	63.5	.83	.51	.90
CE: clarity of expectations	44.2	.56	.24	.71

*Note: *N* = 37 for this dimension. It was not possible to rate the rigor of the text for every assignment

To investigate the independence of the assignment ratings, we examined the relation of the different scales to each other. Our reasons for this were twofold: First, evaluating large-scale reform efforts can be quite costly, so it is imperative that measurement tools be as efficient and streamlined as possible. We examined the interrelation of the rating scales, therefore, to reduce possible redundancy in our rating scheme by investigating whether certain scales may be so highly correlated with one another that they could be eliminated. Additionally, we were interested in looking at the interrelation of the scales within each construct/principle of learning (to how consistent these were with one another), as well as the relationship of these constructs (academic rigor and clear expectations) to each other.

Results indicated that most of the dimensions measuring academic rigor were significantly associated with one another, specifically, the potential and implementation of the assignment tasks ($r = .70$, $p < .05$) and the potential of the task and the rigor of the expectations ($r = .85$, $p < .01$). The exception to this pattern was the dimension measuring the academic rigor of the texts read by students for the assignment. This dimension was not significantly associated with any of the other assignment quality rubrics—within academic rigor or those rubrics measuring the clarity of the expectations.

The two dimensions measuring clear expectations (the clarity of the expectations and the communication of the expectations to students) were

significantly associated (r = .82, p < .01). Additionally, the dimensions measuring the clarity and rigor of the expectations for an assignment task (r = .58, p < .05) and the rigor and communication of the expectations to students were significantly associated (r = .82, p < .01). For the most part, however, the two constructs (clear expectations and academic rigor) did not show a high level of association.

Table 2

Interrelation of assignment ratings for the reading comprehension assignments (*N* = 52)

	Rigor of text	Potential of task	Implementation of task	Rigor of expectations	Clarity of expectations	Comm. of expectations
Rigor of the Text	1	.35	.20	.40	.45	.45
Potential of task		1	.70*	.85**	.25	.51
Implementation of task			1	.49	.01	.31
Rigor of Expectations				1	.58*	.78**
Clarity of Expectation					1	.82**
Comm. of expectations						1

*p < .05, **p < .01

How many assignments and raters would be needed to obtain a stable estimate of the quality of classroom practice?

Generalizability and decision studies were conducted to determine how many raters and assignments might be necessary to obtain a stable estimate of the quality of classroom practice. Results indicated that our design based on two raters and four teacher assignments yielded a generalizability coefficient of only .44 (.80 and above is considered to be good). As shown in Table 3, 39.9% of the variance was explained by the interaction of teacher by assignment type, far surpassing the variation between teachers (9.2% of the total variance). In other words, individual teachers tended to submit assignments of varying quality. Results of decision studies indicated that increasing the number of assignments (or raters) did not greatly increase the stability of the estimate. For example, collecting six assignments from teachers yielded an estimated G-coefficient of only .48.

Table 3

Estimates of Variance Components for the Reading Comprehension Assignments (N = 13).

Source of Variation	Estimated Variance Component [a]	Percentage of Total Variance
Teacher	0.103	9.2
Rater	0	0.0
Assignment Type	0	0.0
Rubric	0.008	0.7
Teacher x Rater	0.016	1.4
Teacher x Assignment Type	0.445	39.9
Teacher x Rubric	0.138	12.4
Rater x Assignment Type	0.004	0.4
Rater x Rubric	0.004	0.4
Assignment Type x Rubric	0.022	2.0
Teacher x Rater x Assignment Type	0	0.0
Teacher x Assignment Type x Rubric	0	0.0
Rater x Assignment Type x Rubric	0	0.0
Teacher x Rater x Assignment Type x Rubric, Error	0.375	33.6

[a]. Negative variance component was set to zero.

What is the relation of the classroom assignment ratings and observed instruction?

To address our third research question, we compared the ratings of assignment quality to the quality of a teacher's observed instruction. The purpose of this was to assess the degree to which the classroom assignment ratings yielded meaningful and appropriate information about students' learning environments that were commensurate with other measures of quality practice.

Results indicated that the degree to which students were asked to analyze and interpret text (potential of the task) and the rigor of a teacher's expectations for student work were associated with the rigor of the observed lesson ($r = .66$, $p < .01$ and $r = .60$, $p < .05$ respectively). Contrary to expectations, however, the implementation of the classroom task was not associated with the level of observed rigor in the observation.

What was the quality of the assignments we collected? Do differentially rated assignments (that is, assignments that are considered to be of high, medium, or low quality) provide qualitative differences in students' opportunity to learn?

The quality of the assignments we collected for this pilot ranged from poor to excellent (and as described earlier, this was true even within some of the same classrooms). As shown in Table 4, however, on average the assignments were considered to be of fair quality (i.e. were rated a '2' on a four-point scale) on all of the dimensions. The exception to this was the dimension measuring the rigor of the text. This dimension was assessed on a three-point scale, so a mean score of (2.38) indicates a somewhat higher level of quality than the other dimensions.

Table. 4

Description of reading comprehension assignments (*N* = 52)

AR Dimension	Mean	SD	Range
Academic rigor of the text*	2.38	.72	1-3
Potential of the task	2.31	.98	1-4
Implementation of the task	1.79	.89	1-3
Rigor of the expectations	2.29	1.07	1-4
Clarity of the expectations	2.39	1.02	1-4

*Note: *n* = 37 for this dimension as it was not possible to rate the rigor of the text for every assignment. Also, this dimension was assessed on a three point scale, as opposed to a four-point scale.

Teachers also completed a checklist describing how they shared their expectations for quality work with students. Results indicated that the teachers for half of the assignments (51.9%) reported that they discussed their criteria for high quality work with students in class and shared models of high quality work with them in advance of their completing the assignment. For slightly more than a quarter of the assignments (26.9%) the teachers reported that they discussed their criteria for high quality work with the students, but did not provide them with models of high quality assignments. Finally, the teachers for nearly a quarter of the assignments (21.2%) reported that they did not share their criteria for assessing students' work with students in advance of their completing the assignment.

We took a closer look at assignment quality to investigate whether our ratings captured meaningful differences in students' opportunities to learn. The following sections describe three assignments—one each considered to be of low (a '1'), fair ('2') and excellent ('4') quality with specific regard to academic rigor.

An assignment considered to be of poor quality. For this assignment fourth-grade students read an excerpt from a book about gorillas by Seymour Simon (Harper Collins) and generated a series of questions about the text. The teacher's expectations for high quality work for this assignment were as follows:

> Students were asked to jot down any initial questions that they had. Two formats were presented. Students selected which format they preferred. As students read they were expected to jot answers to questions and generate new questions.
>
> High performance – students generated a reasonable number of questions (at least 5) and were able to answer if answer was present in the text. Students demonstrated high performance if they asked questions about things they didn't already know. Several questions were also critical.
> Middle performance – Students asked at least five questions. A few were obvious. Students answered the questions. Most answers were copied verbatim from the text.
> Low performance – Few questions, many questions had obvious answers. Missed many answers presented in the text.

These criteria were not explicitly shared with students, but reportedly were modeled for students during a mini-lesson that followed this assignment. The following is an example of student work considered by the teacher to be of high quality for the class for this task.

Gorillas
Q: What kind of food do they eat?
A: A gorillas grab plants to shake fruits, females pick fruit, a large gorilla may climb a tree to reach a favorite snack.
Q: Where do they live?
A: Heavily forested areas in Africa.
Q: Do they have lots of hairs?
A: …black to grayish brown.
Q: How long do they live?
A:
Q: Where in the world do they live?
A: Western tropical rain forest in West Africa.
Q: Do they have toes and fingers?
A: Arms and two legs, and head, five fingers and five toes.
Q: How many teeth do they have all together?
A: 32
Q: Do they eat insects?
A:
Q: Are they shy?
A: Gorillas are shy.
Q: How many pounds do they weigh?
A: Four hundred pounds.

This assignment was considered to be of a basic quality with regard to academic rigor. The questions generated by students were very similar (nearly identical). Furthermore, the questions generated by students, even those students whose work was considered by the teacher to be of high quality, required only basic recall of isolated facts. In contrast to the teacher's stated expectations for the task, students were clearly not guided to generate or answer questions that required them to think "critically," or even to know very much about gorillas beyond very basic, surface details (e.g. that they have five fingers and toes, 32 teeth, etc.).

Generating questions about what you read is a strategy for helping students comprehend text. It does not appear from the students' work, however, that this strategy was applied in a way that deepened students' grasp of the content material or developed their analytical thinking and writing skills. An assignment that would have provided students with a greater opportunity to develop their academic thinking and writing skills would have guided students to read and synthesize more in depth information about gorillas, and/or (for example) apply this information to considering the larger issue of saving endangered species, or generating solutions to problems such as poaching and deforestation. As the assignment was, however, students were provided with little or no opportunity to develop higher-level comprehension skills.

A fair quality assignment. This fifth-grade assignment was considered to illustrate a fair level of quality (a '2') for academic rigor. For this assignment students wrote a summary of the book "Jumanji" by Chris Van Allsburg. The teacher's expectations for this assignment were as follows:

> Students used a rubric and criteria for writing summaries as a guide and to know what was expected of them. Students were to write a summary including all important events from the story (without writing every little detail). They had to include names of important characters and write events that took place, therefore showing comprehension of the text.

The teacher also used the following rubric to assess students' work. These criteria reportedly were developed with the students in class, and students also were provided with models of high quality work. This rubric is as follows:

4 – My summary is explicit and in my own language. I mentioned important character names and important events. I included appropriate details and vocabulary.
3 – My summary is adequate and in my own language. I mentioned important characters' names and important events. I included some details and vocabulary.
2 – I have a partial summary. It is generally in my own language. I included some important characters and events. My summary may have some misinterpretation.
1 – I wrote down one or two events in my own language. I may have copied the text or may have some incorrect information.

The following is an example of student work considered by the teacher to be of high quality for the class.

Jumanji

This story is about two kids names Peter and Judy. Their parents go out so they stay home alone. They get bored so they go outside and find a board game called Jumanji. They play theh game and find out that the game has some crazy twists, like if they land on a lion a lion appears! At the end, Judy lands on Jumanji and all of the mess is gone.

This assignment provides students with a greater opportunity to develop their comprehension skills than the previous assignment (which received a score of '1' for most of the academic rigor dimensions). Rather than recall basic isolated facts about a story, students were supported to create coherent summaries of the book (i.e. create a coherent mental model of what they read). This is a complicated text, and summarizing it would not necessarily be a simple task. Students (even those whose work was considered to be of high quality), however, wrote only single paragraph responses that lacked detail from the story. Furthermore, students were not supported to engage with the deeper

themes, or message of the story, such as, in the case of this story, the importance of perseverance and courage. This is clear both in the teacher's expectations (that focused on students recalling the names of characters and primary events in the story) and in the quality of students' work.

An exemplary assignment. This assignment illustrates an exemplary level of academic rigor and was assigned the highest scores for each of the rubrics measuring this construct. For this third-grade assignment students read two chapters from "The Prince" by Niccolo Machiavelli (Bantam Books). Students then wrote an essay describing the primary message (or lesson) Machiavelli was attempting to convey in the story. To prepare students to complete the assignment, the teacher reported that she took several weeks discussing and analyzing the book with students. Her expectations for the quality of students' work were as follows:

> My expectations were for my students to get Machiavelli's true message to the readers. I also guided the students in the analysis of literature and the process to follow. Focus and deep concentration was needed to understand the work. Most of the children understood the literary work. Most students realized that attacking a serious writing slowly and carefully is not so difficult after all and can be quite interesting. My "checklist" was used, and vocabulary lists were used.
>
> I distinguished between high, middle and low performance work in the compositions by the students' analysis of the actions taken in the chapters…by the [main character], in the order in which these actions were discussed, by the conclusions arrived at by the students, by the mentioning of the most important points of the readings.

The following is an excerpt from a student's essay considered by the teacher to be of medium quality for this assignment.

"The Prince" by Nicolo Machiavelli

The reason we are studying this book is because it tells us that we have to watch out for people that act like the prince. People think that Machiavelli is the prince because he wrote all these bad deeds, but he is not, but he is trying to tell us that people like Sadam Hussan act just like the prince because he wants to start war…

…Agathocles had everything planned…ahead of time. Every thing was planned nothing was left to chance. Agathocles won dominion and boldness, greatness of spirit, but he did not win glory or respect. Agathocles went straight into training of evil doings. One of the evil doings is kill one's fellow citizens. This training of evil doing is for gain of power and land…

If Agathocles was always evil to the max all of the people would rebel. But if he does some nice thing the people will say well, he is not so bad. But it didn't matter to Agathocles, he just wanted to have lots of power.

Now this prince acts just like Agathocles…

This assignment required students to read a very challenging text, and write an extensive essay (several pages) that focused on the author's intention for writing the story. The teacher's expectations for this assignment were quite high and focused on students' grasp of the "true message" of the text, in addition to the inclusion of important actions in the story. This is reflected in the quality of students' work. As shown in the excerpt from a student's essay above, considered by the teacher to be of medium quality for this assignment, students did in fact write about the meaning of the story (e.g., "If Agathocles was always evil to the max all of the people would rebel. But if he does some nice things the people will say, he is not so bad.") Because students were supported to engage with challenging material and talk about the author's purpose in writing the story with details from the text, this assignment was considered to illustrate a high level of academic rigor.

Summary and Conclusions

In summary, while these results are quite limited by the small sample size, the reliability of the dimensions measuring the academic rigor of an assignment (as assessed by the level of interrater agreement) appeared to be moderate overall, and showed a good level of internal consistency. The dimension measuring the clarity of a teacher's expectations (clear expectations), in contrast was poor. Additional development work will need to be undertaken to revise this rubric in order to improve interrater agreement rubric. It is possible that including more benchmark samples of clear expectations in the rater training program could help improve the reliability of that dimension as well.

Most of the dimensions measuring academic rigor were significantly and positively associated with one another, specifically, the potential and implementation of the assignment tasks, and the potential of the task and the rigor of the expectations. The exception to this pattern was the dimension measuring the academic rigor of the texts read by students for the assignment. This dimension was not significantly associated with any of the other assignment quality rubrics—within academic rigor or those rubrics measuring the clarity of the expectations suggesting that this rubric provides unique information regarding instructional quality.

We took a closer look at the assignments we received from teachers to better understand why the rigor of the text was not associated with the other academic rigor dimensions. It appears that some teachers who received low

scores for the potential and implementation of their assignment tasks had assigned high quality texts for their students. It did not appear that any of the teachers who received high scores for these dimensions had assigned low quality texts. In other words, it is unlikely that a low quality text would provide the material necessary for a high quality response to literature. At the same time, teachers did not always exploit the potential of the texts they assigned to students by supporting them to analyze and interpret what they read at a deep level. This was illustrated in the assignment described earlier in this paper where students read a text about gorillas, but only were asked to generate simple questions about the text. It is possible then, that a high quality text could be considered a necessary, but on its own insufficient factor for a high-quality response to literature assignment task. This issue would need to be explored in future research, however, to draw more definitive conclusions.

The two dimensions measuring clear expectations (the clarity of the expectations and the communication of the expectations to students) also were significantly associated. The dimensions measuring the clarity and rigor of the expectations for an assignment task and the rigor and communication of the expectations to students were significantly associated as well. For the most part, however, the two constructs (clear expectations and academic rigor) did not show a high level of association. This suggests that they are measuring two independent facets of instructional quality. These results also raise questions as to the grouping of the different dimensions, notably, if the rubric measuring the

rigor of a teacher's expectations is best situated with the academic rigor dimensions, or if it should be grouped with the clear expectations rubrics.

We also compared our ratings of assignment quality to the quality of observed instruction in order to assess the degree to which the classroom assignment ratings yield information about students' learning environments that were commensurate with other measures of quality practice. Results indicated that the degree to which students were asked to analyze and interpret text (potential of the task) and the rigor of a teacher's expectations for student work were associated with the rigor of the observed lesson. Contrary to expectations, however, the implementation of the classroom task was not associated with the level of observed rigor in the observation. It is not clear to us why this was the case, and raises questions about how we defined potential and implementation in the reading comprehension assignments. These constructs appear to be more difficult to disaggregate in this content area (as opposed to mathematics). Again, future research is necessary with larger samples of classrooms to draw more definitive conclusions.

Generalizability studies were conducted to determine how many raters and assignments would be necessary to obtain a stable estimate of the quality of classroom practice. Results indicated that our design based on two raters and four teacher assignments did not yield a stable estimate of quality (G-coefficient = .44). This means that individual teachers provided assignments that varied in quality—some poor, some fair, and some good.

We returned to the original portfolios submitted by teachers to gain a better understanding of what this variation meant. In fact, while a few teachers did submit assignments that were consistent in quality (e.g. four poor quality assignments), most of the teachers showed quite a bit of variation in their portfolios. For example, the fourth-grade teacher who submitted the assignment about gorillas described earlier (rated a '1' for academic rigor) also submitted a reading response journal assignment where students wrote a letter to the teacher describing what they were reading (scored a '2' as this was mostly a summary of surface-level events), an analysis of a character's traits with some evidence from the text (scored a '3'), and an assignment where students drew their impressions of a text (scored a '1').

It is possible (even probable) that teachers would submit assignments that were more consistent in quality if we asked for all challenging or all typical work. Considering that we asked for both, our results are hardly surprising. Being more specific with regard to the type of reading comprehension assignment we ask teachers to submit (e.g. all responses to literature) likely would increase the stability of our ratings at the teacher level. Our results from this study are interesting, however, for showing the wide degree of variation within classrooms in students' opportunities to develop their comprehension skills, and more importantly, suggests that teachers may have a broad (perhaps fuzzy) idea of how to support student development in this area.

Finally, while additional refinement of the rating scales is needed, it appears that our ratings of individual assignments captured meaningful differences in students' opportunities to development higher-level thinking and writing skills. Furthermore, it appears from the quality of the assignments we collected that there is room for improvement in many of the assignments given to students within classrooms. Specifically, as found in other research on reading comprehension instruction, it appears that the tasks assigned to students do not frequently provide students with an opportunity to develop more complex thinking skills, or apply strategies in a way that supports students to look beyond the surface-level features of a text. Future research is being planned that will focus on the relation of these ratings to student achievement. Looking at the relation of specific dimensions to student learning will help us further refine our scales, and will provide additional information as well that could be useful in terms of developing the IQA as a tool for instructional leadership, professional development and (teacher) self-assessment. These issues will be explored further in the last paper of this symposium by Crosson et al.

References

Beck, I., McKeown, M., Hamilton, R., & Kucan, L. (1997). Questioning the author. Newark, DE: International Reading Association.

Clare, L., & Aschbacher, P. (2001). Exploring the technical quality of using assignments and student work as indicators of classroom practice. Educational Assessment, 7(1).

Crosson, A., Junker, B. W., Matsumura, L. C., & Resnick, L. B. (2003). Developing an Instructional Quality Assessment. Paper presented at the Annual Meeting of the American Educational Research Association, April 2003, Chicago IL.

Institute for Learning (IFL, 2002). Principles of Learning. Overview available at http://www.instituteforlearning.org/pol3.html. University of Pittsburgh, Pittsburgh PA: Author.

Matsumura, L.C., Garnier, H., Pascal, J., & Valdés, R. (2001). Measuring instructional quality in accountability systems: Classroom assignments and student achievement. Educational Assessment, 8(3), 207-229.

National Reading Panel (2000). Teaching children to read: An evidence based assessment of the scientific research literature on reading and implications for reading instruction. Washington, DC: National Institute of Child Health and Human Development.

Newmann, F.M., Lopez, G., & Bryk, A.S. (1998). The quality of intellectual work in Chicago schools: A baseline report. Chicago: Consortium on Chicago School Research.

Newmann, F.M., Bryk, A.S., & Nagaoka, J.K. (2001). Authentic intellectual work and standardized tests: Conflict or coexistence? Chicago: Consortium on Chicago School Research.

Pressley, M. (1998). Reading instruction that works: The case for balanced instruction. New York: Guilford Press.

Resnick, L. & Nelson-LeGall, S. (1999). Socializing intelligence. In L. Smith, J. Dockrell, & P. Tomlinson (Eds.), Piaget, Vygotsky and beyond. London: Routledge.

Rogoff, B. (1992). Apprenticeship in thinking. New York: Oxford University Press.

Sheppard, L. (2000). The role of assessment in a learning culture. Educational Researcher, 29(7), 4-14.

Snow, C. (2002). Reading for understanding: Toward and R&D program in reading comprehension. RAND Reading Study Group. Santa Monica, CA: RAND.

Stein, M. K., Smith, M. S., Henningsen, M. A., & Silver, E. A. (2000). Implementing standards-based mathematics instruction: A casebook for professional development. New York, NY: Teachers College Press.

Storms, B.A., Riazantseva, A, & Gentile, C. (2000). Focusing in on content and communication (writing assignments that work). California English, 5(4), 26-27.

Vygotsky, L.S. (1978). Mind in society: The development of higher psychological processes. Cambridge, MA: Harvard University Press.

Appendix A: The Principles of Learning

The Principles of Learning are comprised of nine interrelated constructs. The first Principle of Learning, *organizing for effort*, is based on the assumption that nearly all students who put forth a focused effort over a sustained period of time can and should have the opportunity to master a challenging curriculum. Standards that clearly set out what students are supposed to be able to do in every subject area, and that set forth specified tasks (e.g. reading a certain number of books, writing certain types of papers, etc.) students are to accomplish are critical for supporting an effort-based school system. Apportioning instructional time and resources (including parent involvement) in such a way that students who are having academic difficulty have access to additional support also is of critical importance for supporting an effort-based learning culture.

Holding *clear expectations* for student learning and communicating these expectations to students is the second Principle of Learning. This can be accomplished, for example, by teachers making their scoring criteria accessible to students by posting or distributing standards and rubrics, discussing with students the criteria for work that meets a specific standard. Providing students with models of high quality work that outline a sequence of expected concepts and skills students are to master in the process of accomplishing a larger standard, and discussing these models with students also are important for communicating expectations to students. Other important means for making

expectations clear to students include involving students in judging their own work with respect to the standards, and communicating to parents what students are supposed to accomplish.

Providing students with *fair and credible evaluations* are important for supporting an effort-based learning culture. These types of assessments should be aligned with standards and should provide meaningful information to teachers, parents, colleges, and future employers about what individual students know and can do. To be fair, these types of assessments should be ones that students can prepare for, meaning that they are tightly aligned with the everyday curricula to which students are exposed.

In this same vein, frequent *recognition of accomplishment* is important for motivating students to put forth high levels of effort over time. This accomplishment must be authentic in that it should represent real achievement in learning. Additionally, this recognition should be organized around clearly demarcated progress points en route to accomplishing larger learning standards.

Students also must have the opportunity to engage with academically rich content material and to develop their thinking skills in order to achieve at high levels. This Principle of Learning holds that students must be exposed to a rich knowledge core that is organized around the mastery of major concepts. This curriculum also should provide students with the regular opportunity to pose and solve problems, formulate hypotheses, justify their reasoning, construct explanations, interpret text, and test their own understanding. Additionally,

students should have the opportunity to construct their own understandings of concepts based on the synthesis of several sources of information including their experiences outside of school.

In order for classroom talk to promote learning it must be accountable to a learning community. This idea is captured in the principle of learning, *accountable talk*. This means that students and teachers should respond and build on each other's contributions and work together to understand each other's positions through clarifying and summarizing each other's arguments. Students (and teachers) also should make specific use of knowledge and provide discipline specific evidence to support claims and arguments. Parallel to the principle of learning *academic rigor*, these types of conversations should press students to synthesize information from several sources, construct explanations, conjectures and hypotheses, and test their own understanding of concepts (e.g. by asking questions, comparing ideas, etc.). Students also should be encouraged to challenge the quality of each other's evidence and reasoning.

One way to support the development of higher-order thinking skills and an effort-based learning culture is to model and hold students accountable for using intelligent habits of mind. This principle of learning, *socializing intelligence*, asserts that intelligence can be viewed as much as a learned behavior as an innate trait. In other words, students can and should be taught to think more intelligently by learning to regularly employ problem-solving and reasoning skills. This can be accomplished by communicating to students that they are able

to become even more competent learners through the persistent application of strategies, and holding students accountable for analyzing problems, asking questions, and getting information.

To take responsibility for the quality of their thinking, students also need to develop *self-management of learning* skills, that is, they need to develop the meta-cognitive skills necessary to monitor their own learning. For example, students need to learn to regularly check their own understanding by restating ideas and concepts in their own words, asking themselves questions, an checking new information against their own background knowledge. These skills also can be developed by encouraging students to assess their own work. Teachers play an integral role in the acquisition of these skills by actively modeling these skills for students during lessons. For example, as described before, by thinking out loud and articulating their own problem solving strategies and processes.

The final principle of learning, *learning as apprenticeship*, focuses on the importance of creating opportunities for students to participate in guided learning. This principle of learning underscores the importance of children being exposed to high quality models for learning, and the importance of children receiving specific coaching and mentoring toward the accomplishment of learning goals.