Overview of the Instructional Quality Assessment

Brian Junker[1], Lindsay Clare Matsumura[2], Amy Crosson[2], Mikyung Kim Wolf[2], Allison Levison[2], Yanna Weisberg[3], and Lauren Resnick[2]

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

Contact Information:

Brian W. Junker
Department of Statistics
Carnegie Mellon University
232 Baker Hall
Pittsburgh PA 15213
brian@stat.cmu.edu

Lindsay Clare Matsumura
Learning Research and Development Center
University of Pittsburgh
3939 O'Hara St., Room 801
Pittsburgh PA 15260
lclare@pitt.edu

[1] Department of Statistics, Carnegie Mellon University
[2] Learning Research and Development Center, University of Pittsburgh
[3] Department of Psychology, Carnegie Mellon University

Overview of the Instructional Quality Assessment

Abstract

Educators, policy-makers, and researchers need to be able to assess the efficacy of specific interventions in schools and school districts. While student achievement is unquestionably the bottom line, it is essential to open up the educational process so that each major factor influencing student achievement can be examined; indeed as a proverb often quoted in industrial quality control goes, “That which cannot be measured, cannot be improved”. Instructional practice is certainly a central factor: if student achievement is not improving, is it because instructional practice is not changing, or because changes in instructional practice are not affecting achievement? A tool is needed to provide snapshots of instructional practice itself, before and after implementing new professional development or other interventions, and at other regular intervals to help monitor and focus efforts to improve instructional practice. In this paper we review our research program building and piloting the Instructional Quality Assessment (IQA), a formal toolkit for rating instructional quality based primarily on classroom observation and student assignments. In the first part of the paper we review the need for, and some other efforts to provide, direct assessments of instructional practice. In the second part of this paper we briefly summarize the development of the IQA in reading comprehension and in mathematics at the elementary school level. In the third part of the paper we report on a large pilot study of the IQA, conducted in Spring 2003 in two moderately large urban school districts. We conclude with some ideas about future work and future directions for the IQA.

Table of Contents

Background

The Need to Measure Instructional Quality

Although its roots go much father back, standards-based educational reform and accountability came to the fore of American educational policy in the early 1980's, with the National Committee on Excellence in Education's (NCEE) open letter to the American people, *A Nation at Risk* (NCEE, 1983),which laid out a forceful set of recommendations including rigorous definitions of content standards in English, Mathematics, Science, Social Studies, and other basic subjects, as well as rigorous and measurable standards for student achievement and instructional quality. These ideas were codified ten years later into eight national goals in the Federal "Goals 2000: Educate America" Act of 1994 (National Education Goals Panel, 2000), which emphasized student achievement in both basic and challenging subjects, and specifically mentioned teacher education and professional development.

While student achievement is unquestionably the bottom line in state assessments (Doherty, 2003; Fuhrman, 1999)—and it is the focus of the recent the Federal "No Child Left Behind" Act of 2001 (NCLB)—it is essential to open up the educational process so that each major factor influencing student achievement can be examined; indeed as a proverb often quoted in industrial and software quality control goes, "That which cannot be measured, cannot be improved". In a climate of high-stakes achievement-oriented accountability, inasmuch as it is unacceptable to deny students the instruction they need to meet achievement standards, it is also unacceptable to deny educators the tools they need, to measure, reflect upon, and improve their own practices, to help students reach those standards. Yet, insufficient emphasis has been placed on monitoring instructional quality, arguably the most important *school* factor influencing student achievement (Darling-Hammond, 2000).

Quality of instruction has not been directly measured in many accountability systems because few assessment tools exist that have the potential to directly measure the quality of classroom practice on a large-scale basis. Two common data sources—teacher and student self-reports—are relatively cheap and do contain some signal about instructional quality, but if they are not done carefully, self-image biases (Meyer, 1999a; Spillane & Zuelli, 1999), recall biases (Lohr, 1999, p. 8), and competency and self-interest issues (e.g. related issues in Sproule, 2000) undermine the validity of the results. Two other relevant data sources—rating teachers' assignments/student work, and rating based on direct classroom observation—offer more scope for valid measurement of classroom instruction (Clare and Aschbacher, 2001; Clare Matsumura, 2000, 2003), but they are also much more expensive. They are more expensive for two reasons: first, raters must usually be compensated for their time (ranging from a bare minimum of 10-15 minutes per assignment portfolio, to hours or days per classroom observation); and second, there are expenses involved in training raters to draw appropriate inferences from the raw data for rating.

This last point brings up a more subtle difficulty: it is really only possible for a rater to distill and draw inferences from complex, real-time data, in the context of a specific theory of instructional practice. Without an observation protocol to help the rater to determine what is important in the observation and what can be filtered out, the rater will be overwhelmed and no useful rating can occur. Different theories of instructional practice—Direct Instruction (e.g. Kameenui & Carnine, 1998) vs. Constructivist teaching (e.g. Fosnot, 1996), to name two extremes—suggest rather different observation protocols, as well as different ways of judging the evidence obtained. Our own view is that the recent National Research Council reports *How People Learn: Brain, mind, experience, and school* (NRC, 1999a) and *How people learn: Bridging research and practice* (NRC, 1999b) provide an important outline of current scientific understanding of learning and its implications for teaching. For our purposes, the essential message of these reports is that powerful teaching and learning environments are:

- *Learner-Centered*: Teachers must be able to recognize predictable misconceptions of students that make the mastery of particular subject matter challenging; draw out preconceptions that aren't predictable; and work with preconceptions so that students can build on them, challenge them, and, when appropriate, replace them. Teachers must also consciously incorporate metacognitive instruction into curricula, so that students can learn to self-monitor and make decisions about their own learning.
- *Knowledge-Centered*: Teachers must teach some subject matter in depth, providing enough examples in which the same concept is at work, so that students can grasp the core concepts in an area, deepen their understanding, and engage in complex problem solving. Teachers must combine in-depth knowledge and organization of the subject area with pedagogical tools that include an understanding of how students' thinking about concepts in the subject develop.
- *Assessment-Centered*: Teachers must help students develop a clear understanding of what they should know and be able to do, setting learning goals and monitoring progress together. Students should produce quality work, showing evidence of understanding, not just recall; and assessment tasks should allow students to exhibit high-order thinking.
- *Community-Centered*: Teachers must arrange classroom activities and help students organize their work in ways that promote the kind of intellectual camaraderie and attitudes toward learning that build academic community. The community norms established in the classroom have strong effects on students' achievement.

Some Other Efforts to Measure Instructional Quality

Although much has been written about the quality of measures of instructional practice—going back to at least Hiller, Fisher, & Kaess (1969)—until recently efforts to develop broad-based, technically sound measurement instruments have been somewhat limited. As Meyer (1999b) puts it, this is due to the fact that, historically, education reforms have tinkered at the edges of the educational process (Marshall, Furmann, & O'Day, 1994); and even during the extensive reform efforts of the 1970's and 1980's policymakers focused on improving schools by adjusting resource allocations (racial balance, financial equity and the like) and focusing on outcome goals (e.g. minimum competency testing). The push for routine collection of instructional practice data came only in the late 1980's (e.g. Porter, 1991; Shavelson, McDonnell, Oakes, Carey, & Picus, 1987).

In the 1990's, the 1994-95 School and Staff Survey Teacher Follow-up Study involved survey data from 3,844 teachers (Henke, Chen and Goldman, 1999), but the reliability of the 22-item self-report instrument measuring instructional practice was problematic (e.g., Meyer, 1999b) since the items did not come from a single, coherent theoretical/pedagogical framework. As part of RAND's evaluation of the Federal Systemic Initiatives program of the 1990's, Klein, Hamilton, McCaffrey, Stecher, Robyn, & Burroughs (2000) studied instructional practice and student achievement with 627 teachers distributed over three elementary/middle grade levels and six sites. They found substantial variation in educational practice within schools, and, after controlling for background variables, a generally weak but positive relationship between frequency of "reform" teaching behaviors and student achievement; the relationship was somewhat stronger when achievement was measured with open-response tests than with multiple-choice tests. In a state-level effort, Fouts, Brown & Thieman (2002) found positive correlations between constructivist teaching behaviors and achievement, in 669 classrooms distributed among 34 elementary, middle, high and technical schools, and negative correlations between each of these variables and family income. After controlling for family income, constructivist behaviors still accounted for a small but significant portion of the variation in student achievement. The Study of Instructional Improvement program (SII, Regents of the University of Michigan, 2001) has developed a carefully constructed instrument to measure teachers' pedagogical content knowledge (Rowan, Schilling, Ball & Miller, 2001), scaled using the Rasch (1980) model from Item Response Theory. SII has also developed a set of "Instructional Logs" (Ball, Camburn, Correnti, Phelps & Wallace, 1999; SII, 2001) which are more highly formalized and detailed teacher self-report forms, to be completed two times a day over a period of 120 teaching days by each teacher, and is using these instruments in its intensive study of instructional practices in three leading school improvement programs.

There is, however, still a need for a technically sound tool that can be used for a variety of research and monitoring purposes more modest that SII's wholesale program evaluation, but still essential to maintaining and improving instructional practice. Such a tool might provide snapshots of instructional practice—instructional quality—at baseline measure before engaging in new professional development, as a post-measure to evaluate the effects of professional development, and at other regular intervals to help monitor and focus professional development efforts.

Development of the Instructional Quality Assessment

Overview

The *Instructional Quality Assessment (IQA)* has been under active development at the Learning Research and Development Center (LRDC) at the University of Pittsburgh since the beginning of 2002. The IQA consists of approximately 20 rubrics or rating items, organized into three clusters, together with training materials and observation protocols for raters who administer it. Separate versions of the IQA are needed for different subject areas and grade-bands. So far, pilot versions of the IQA have been developed for Mathematics and Reading Comprehension at the primary and upper elementary levels.

The IQA was conceptualized around a specific set of guidelines for instructional practice which integrates strong pedagogical knowledge with deeply rigorous subject matter knowledge, the *Principles of Learning* (Resnick & Hall, 2001; Institute for Learning, 2002). This set of statements about highly effective, effort-based learning and instructional practices is used to guide consulting and collaboration with client districts within LRDC's Institute for Learning (IFL) The Principles of Learning grew out of the study of teaching and learning in the High Performance Learning Communities (HPLC) project at LRDC, and its study of District Two in New York City in the 1990's (Resnick, Glennan & Lesgold, 2001). More broadly, the Principles of Learning arise from the same larger body of research that underlies the *How People Learn* (NRC 1999a,b) reports. Four of the Principles of Learning capture characteristics of expert instruction that can be observed in the classroom and by looking at the way an instructor constructs assignments for his or her students:

- *Academic Rigor* insists that lessons be built around specific important concepts in the subject area, and that students regularly engage in active reasoning about challenging content and core concepts in that subject; this necessarily entails subject matter expertise on the part of the teacher. Indeed, the principle of Academic Rigor encompasses the notions of active inquiry and in-depth learning of important content that is at the heart of the NRC's (1999a,b) notion of *Knowledge-Centered* teaching.
- *Clear Expectations* guides teachers to make standards-based expectations for performance clear to students, and encourages teachers and students to set learning goals and monitor progress together. This addresses basic opportunity-to-learn and alignment issues between instruction and assessment, and helps students (and teachers) benefit most

from the kind of ongoing formative assessment environment that is the key idea of *Assessment-Centered* teaching (NRC, 1999a,b).

- *Self-Management of Learning* emphasizes the importance of incorporating metacognitive skills into instruction, in all areas, so that students can develop self-monitoring and self-management strategies to regulate their own learning. These skills that are necessary for effective active-inquiry based learning, and encompass an important part of *Learner-Centered* teaching (NRC, 1999a,b).
- *Accountable Talk* identifies the characteristics of classroom discussions that support coherent, sustained social interaction in which students—and the teacher—build on each others' ideas and hold each other accountable to accurate knowledge and to rigorous thinking. In this way, social norms are created in the classroom that support active inquiry, deep learning, and clear expectations, and build a collegial academic atmosphere in the classroom as outlined in the NRC's (1999a,b) notion of *Community-Centered* teaching.

Development

Initial development of the IQA was strongly influenced by Clare Matsumura's (2000) efforts to assess collections of assignments and student work as indicators of overall instructional quality at the classroom level, and by Newmann, Lopez, & Bryk's (1998) work with the Consortium on Chicago School Research to measure the efficacy of reform efforts at the school-level. The TIMSS Videotape Classroom Study (Stigler, et. al, 1999) and the QUASAR Project (Silver, 1996; Stein, et al. 2000) influenced our development of observation protocols as well. Four sources of evidence were chosen for the IQA, because they are available through relatively brief classroom visits, short structured interviews, and limited post-visitation analysis of classroom artifacts, in order to keep the assessment as efficient and affordable as possible. They are:

- *Classroom observation*. Each rater observes one full lesson, approximately 45 minutes, per classroom/teacher. The teacher briefly describes the goals and activities of the lesson in a pre-visit questionnaire which raters review before observing the lesson. The rater dedicates much of the visit to recording and scoring evidence of Accountable Talk since this is the only source of evidence for that Principle of Learning. The rater also scores evidence of Academic Rigor by examining the texts, tasks, and implementation of tasks, that are part of the lesson.
- *Student interviews*. During the lesson, the rater also scores evidence of Clear Expectations and Self Management of Learning by conducting 3-5 minute structured interviews with students (asking questions such as "If I were a new student in this class, how would I know what to do to do a good job on this assignment?" and "Will you have a chance to make your work better after you hand it in?").

- *Teacher interviews*. Following the lesson, the rater conducts a brief, structured interview with the teacher. This interview is intended to gain any additional contextual information to help interpret the teacher's goals for the lesson, and how the observed lesson fits into instruction over time.
- *Teacher-generated assignments*. Each teacher provides a small portfolio of written assignments he or she has generated for the class (typically four assignments have been used, two of which are considered to be "especially challenging" and two of which students have very recently completed). For each assignment, the teacher completes a cover sheet describing the goals and grading criteria for the task, providing contextual information about how the task fits into instruction over time, and describing what kinds of scoring guides and self-evaluation opportunities students were given while working on the assignment. The teacher also provides examples of student work exemplifying low, medium and high performance levels on the assignment; student work is not rated directly, but instead is used to help interpret the assignment as "enacted" for students. These assignments are rated for aspects of Academic Rigor, Clear Expectations, and Self-Management of Learning.

To develop rubrics for the IQA, the Principles of Learning above were analyzed for specific themes or conceptual dimensions that capture the essential features of each Principle. For example, in the case of Clear Expectations and Self-Management of Learning, six such themes were identified: students' understanding of expectations; students' use of criteria to judge work; students' use of criteria to revise work; clarity of assignment directions; clarity and detail of grading criteria; and alignment of grading criteria and task. For each theme, one or more rubrics was created, that clearly defines manifestation of that theme in classroom instruction and artifacts, on a four point scale (1=poor and 4=excellent).

Rating with IQA rubrics is designed to be as low-inference as possible. Each scale point of each rubric is given a precise and explicit descriptor of instructional performance at that level, and the rating process is scaffolded by having the rater record focused field notes and checklist items during observation, before scoring rubrics. This effort to create relatively low-inference rating stands in contrast to many existing instruments designed to measure teaching effectiveness, which often contain general descriptors and tend to rely on extensive rater training coupled with the inferential capabilities and background knowledge of the raters. For example, the Queensland School Reform Longitudinal Study's (2002) instructional quality instrument, which built directly off of Newmann et. al.'s work (1998), contains rubrics that require extensive rater expertise. A specific illustration of this is seen in the Queensland School Reform Longitudinal Study's rubric that requires raters to make a judgment about whether the observed lesson was based on content that is considered "central to learning." Because the tool does not precisely define how to judge what qualifies as "central to learning" in different content areas and at various grade levels, high levels of rater expertise and training would be necessary to achieve reliability and to use the tool for teaching about quality instruction. As a second example, since the Horizon Protocol provides descriptors for only the first and last points on a five-point scale, rater training is required to learn what kind of evidence would lead to ratings along the different intermediate points. The IQA differs from these studies in that it provides two scaffolds for the rating process: focused field notes and/or checklists to help distill the raw observational data, and explicit descriptors of levels of quality instruction for each rubric.

The advantages of mapping out the attributes of quality instruction, scale point by scale point on each rubric, are two-fold. First, it makes the IQA more feasible to administer, and more reliable in practice. Raters' backgrounds do not have to be as extensive as with less-elaborated rating instruments, and rater training can be more efficient. Using relatively low-inference rubrics, raters who have been through rater training can achieve higher reliability of ratings, which is essential for maximizing information from each rated sample of classrooms.

Second, the IQA is a kind of "performance assessment" for teachers, just as achievement tests are assessments for students. Inasmuch as teachers—as well as students—are learners who need to develop clear expectations about their performance in order to manage their own learning, it is essential to specify what is being assessed by the IQA. Detailed and precise descriptors of IQA rubrics do this. Eventually we expect the IQA to function not only as an "external" summative tool but also as the basis of an "internal" *learning* instrument (Sheppard, 2000). District professionals and teachers who use IQA rubrics and materials based on them for self-study will learn specifically what we mean by "instructional quality," and can use the rubric descriptors to bolster instructional practice. General high-inference descriptors, on the other hand, would require a trained, "expert" rater to interpret and unpack the meaning of the descriptors.

Work to Date

Our work so far has focused on developing IQA rubrics, observation protocols and scoring guidelines, for mathematics and reading comprehension in the primary and upper elementary grades. We have found that some variation in the IQA is needed for different subject areas and grade levels. Broadly speaking, rubrics for Clear Expectations/Self-Management of Learning and Accountable Talk seem to work quite well across subject areas and grade levels, with relatively minor modification of observation protocols and benchmark examples of various scale points required[4]. On the other hand, Academic Rigor cannot be separated from subject area, and somewhat different rubrics and descriptors, as well as variations in observation protocols and rater training materials, are needed[5]. In the content area of mathematics, we have borrowed, with

[4] As an example, since younger students are not able to talk abstractly about "what good work is", interview scripts for younger students are somewhat different from interview scripts for older students.

[5] For example, we have found the same level of specificity in assignment directions needed to maintain Academic Rigor in reading comprehension, perhaps by limiting students' scope to substitute superficially relevant personal

few modifications, from the QUASAR framework (Smith & Stein, 1997; Stein, et al., 2000). In English Language Arts, we have expanded on the previous rubrics designed by Clare Matsumura (2000).

The IQA is administered in three phases. First, after classrooms are selected for study, teachers in selected classrooms fill out the pre-visit questionnaire for classroom observation and assemble assignment portfolios. Second, raters visit classrooms. During classroom visits, raters record the presence or absence of specific kinds of evidence, take limited field notes, and conduct short student interviews, according to specific observation protocols. Teacher interviews are conducted as soon after the lesson as possible; then, IQA rubrics relating to Academic Rigor, Clear Expectations/Self Management of Learning, and Accountable Talk in the classroom are scored. In the third phase, teachers' assignment portfolios are rated, offsite, using IQA rubrics relating to Academic Rigor and Clear Expectations/Self Management of Learning.

An earlier version of the IQA for mathematics and reading comprehension was field tested in a feasibility study May 2002 (Crosson, Junker, Matsumura & Resnick, 2003), in three K-8 schools in a large urban school district in the northeastern United States. IQA developers acted as raters, visiting five mathematics lessons and seven reading comprehension lessons in pairs. Rubrics were scored twice by each rater, once before and once after a consensus-building session within each rater pair. Audio recordings of all classroom observations, student interviews, and consensus-building sessions were made to assist in post-hoc evaluation of the rubrics. Although the small sample precluded full analysis, exploratory analyses were conducted including both summary/graphical methods and variance components analyses (e.g. McCulloch & Searle, 2001) where possible, using individual raters' scores before consensus-building sessions. Composite scores (total scores) for Academic Rigor, Accountable Talk and Clear Expectations showed larger effects for teachers than for schools or raters. Inter-rater reliability was generally high and did not differ substantially for math and reading comprehension. Based on these quantitative analyses, on qualitative analysis of audiotapes and raters' notes, and on informal review by LRDC/IFL staff, the IQA was revised to streamline observation and interview protocols, to strengthen the link between the content of the rubrics to research and theory in math and literacy instruction, and to more systematically define how student work samples could be used as a window on instructional quality.

A Pilot Study of the IQA

A larger trial of the revised IQA was conducted in Spring 2003, using sixteen mathematics lessons and fourteen reading comprehension lessons from randomly-sampled elementary schools

history for substantial interpretation of a text, is likely to provide too much specificity to maintain Academic Rigor in mathematics, transforming the enacted task from mathematics problem solving to routine computation.

in two similar-sized urban school districts, in different eastern states. The districts, to which we shall refer as District C and District D, had 26-28 elementary schools each, were selected for this study to be similar demographically[6] but differ in professional development and related efforts. Both districts were IFL-affiliated districts, but District C had been involved for a longer period of time than District D in activities related to the Principles of Learning, and a major purpose of this study was to determine whether the IQA was sensitive to this difference. A second major purpose of the study was to design and evaluate a full rater training program; in particular we wanted to know if professionals not associated with the IQA could be trained to be reliable IQA raters. Finally we wanted to learn about individual behavior of the rubrics and relationships among them, and whether differences could be detected on a per-rubric basis.

Method

With the help of the principle IFL staff person working with each school district, a District administrator in each district was approached with a proposal for the study. After negotiation in each district, a study design was agreed to, envisioning six randomly-sampled elementary schools each District, targeting four lessons/classrooms per school (mathematics at grades 2 and 4; and reading comprehension at grades 2 and 4), for a total of 24 classrooms per District. Schools were randomly selected by IQA staff, reviewed by the District administrator, and their principals were invited by letter from the District administrator to participate in the study. All six schools selected in District C participated in the study. In District D, one of the six randomly selected schools was withdrawn by the District for administrative reasons and replaced with another randomly-sampled school. IQA staff visited each school to recruit teachers. Participating teachers received a $100 bookstore gift certificate; the two Districts and the participating schools received brief reports of preliminary results of the study.

In both Districts, teacher recruitment varied greatly from building to building. To achieve full sample size in District C, a seventh randomly selected school was added, and one teacher each in grades 3 and 5, in different schools, were added to the study; in the end, seven schools and 17 classrooms/teachers participated from District C. Initially a full sample of teachers/classrooms was obtained in District D, but some teachers and schools dropped out during data collection so that in the end four schools and 13 classrooms/teachers participated from District D. In these 30 classrooms, 16 mathematics lessons were observed and 14 reading comprehension lessons were observed. The complete design for classroom/lesson observation is shown in Tables A1 and A2 in Appendix A.

Six raters were recruited from graduate schools of education in universities near the two Districts, and underwent a 2.5 day training program designed and administered by IQA

[6] Students in the two districts were 26% African American, 6% Asian, 47% Latino, 15% white, 6% other; 20% of these students were identified as 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.

developers. Raters visited classrooms in pairs, accompanied by an IQA staff member, observing lessons and interviewing up to four students per rater per classroom, depending on availability. Student interviews were guided by a standard script. Visits were scheduled in consultation with the teacher, so that the lesson included a group discussion (reading comprehension) or group problem solving (mathematics), as well as individual or small-group tasks.

All three observers (two raters and the IQA staff member) produced independent ratings of academic rigor (AR), clear expectations (CE) and accountable talk (AT), based on lesson observation and a short, scripted teacher interview; after this, the three observers produced and recorded consensus ratings for the classroom just observed. All four sets of ratings were recorded for each classroom. Raters were assigned to classrooms so that, within each District, each rater saw approximately the same number of classrooms, and across districts, each rater rated approximately the same number of times with each of three other raters.

Each participating teacher was also asked to prepare an assignment portfolio, consisting of four assignments he or she had prepared for the class: two challenging assignments and two recent assignments. For each assignment, the teacher filled out a short questionnaire describing the content, grading criteria, and other aspects of the assignment, and supplied graded examples of student work at low, medium and high levels of performance. These portfolios were analyzed offline approximately three weeks after classroom rating. Two raters, recruited from the six classroom raters, examined all assignments and produced independent and consensus ratings for academic rigor (AR) and clear expectations (CE), for each assignment in each teacher's portfolio.

The raters were not told why the districts were selected for the study, and were unfamiliar with the IQA prior to training; the IQA staff members who accompanied the raters to the classrooms also did not discuss these issues with the raters, nor did they discuss raters' individual ratings until the consensus-building sessions for each classroom. A similar procedure was used for rating assignments.

All classroom and assignment ratings were recorded on 4-point Likert scale rubrics, in which, broadly speaking, "1" denotes *non-proficient* performance, "2" denotes *approaching proficient* performance, "3" denotes *proficient* performance and "4" denotes *exemplary* performance. Missing data (rater unable to observe, lesson did not contain activity relevant to this rubric, etc., was marked "NA" (not applicable) for each rubric. The rating forms that raters used also contained observation checklists, stylized forms for field notes, and descriptions and examples of behavior characterizing each point on each rubric, to scaffold the rating process. Raters filled out checklists and field notes in real time during lesson observation, and generated individual ratings immediately after lesson observation; consensus-building sessions followed thereafter. All lessons were also audiotaped and transcribed for post-mortem analysis of the rating process and for qualitative analyses the processes being rated (see, e.g., Wolf et al., 2004).

Except where noted, individual raters' ratings before consensus-building sessions within each rater pair were used in subsequent analyses. For AT, only "consensus" scores were recorded because the raters found they didn't have time to both interview students and observe classroom activity, so they split up these two activities and combined them to produce single consensus rarings after the lesson. In some analyses, classrooms at grades 2 and 3 are grouped together and called "primary" grades and grades 4 and 5 are grouped together and called "upper" elementary grades. Schools were also identified as low-, middle- or high-achieving, based on whether they were ranked in the lower, middle or upper third of all elementary schools in their District, by a recent fourth grade standardized achievement test scores in that District.

Results

Separate analyses were performed for lesson observation and assignment ratings, in each of the following areas. All analyses were conducted using SPSS 11.0 (2001).

- *Reliability:* Reliability was calculated between the two trained raters overall and by principles and by rubrics. Both exact agreement and one point agreement were calculated for overall lesson observation scores. We also explored which adjacent rating categories were most difficult for raters to distinguish, by comparing exact agreement on the 4-point rubrics with exact agreement on 3-point rubrics created by merging each pair of adjacent rating categories. Reliability over time was also investigated, to see if rater training effects wore off or if there were practice effects from experience.
- *Score Distribution:* Means, standard deviations and histograms were calculated for each score distribution, overall and by rubrics. Separate analyses of some score distributions were also performed by district, grade or subject.
- *Relationship:* Spearman correlations were calculated between all pairs of rubrics, between principles, and between rubrics measuring the same Principle of Learning from different sources. Logistic regression analyses were also completed on the overall Lesson Observation data.

For Academic Rigor it is also necessary to break out the analyses according to whether the lesson or assignment cover Reading Comprehension or Mathematics. This is because the wording of the rubrics, and the observational protocols, are somewhat different in these two subject areas.

Lesson Observation Ratings

Tables 1A and 1B gives percent agreement, percent agreement within one scale point, Kappa and Spearman correlations, and intraclass correlation, for various aggregations of the lesson observation rubrics. Table 1A gives aggregate reliability indices for all rubrics, for rubrics scored just in Reading Comprehension or just Mathematics lessons, and for rubrics scored within each District. Percent exact agreement hovers around 50%, which is not very high, but percent agreement within one scale point is quite good indeed., at 95% or better. The Kappa and Spearman correlations are both moderate. Despite these moderately low results, the overall intraclass correlation is moderate to good, suggesting that total scores may offer a reliable index of instructional quality.

Table 1B presents the same reliability indices, for all classrooms, within each Principle of Learning. Percent exact agreement and Kappa were similar across Principles, and similar to the values in Table 1A. Spearman's r also ranges over the same values as in Table 1A, but are somewhat more variable than percent agreement or Kappa. Within each Principle of Learning except for AR in Mathematics, the intraclass correlations again suggest that total scores may provide reliable indices of instructional quality within each Principle.

To examine the increase in percent agreement when we move from exact agreement to agreement within one point, we also explored changes in reliability of rating when each two adjacent rating categories were merged. Table 1C gives the result and suggests (a) that the greatest gain could be had by merging the "proficient" (category 3) and "exemplary" (category 4) score points; however apparently there is confusion about other categories as well since this operation only brings the exact agreement up into the range of 60% or so, rather than the 95% seen in Table 1A.

We also examined the stability of rater agreement over time (see Figure 1). The first four time points represent days that the raters spent in District C, the last four represent days in District D. Within each district, percent exact agreement increased moderately as the data collection continued. Although there was a drop, as might be expected, in the transition from District C to District D, overall the percent agreement increased across districts. This suggests both that the raters continue to learn as they do "live" rating, and suggests that perhaps the rater training program should be extended by having the raters rate some "live" but out-of-sample classrooms, before rating "live" classrooms that will contribute to an IQA score.

Table 1A. Inter-rater reliability of lesson observation ratings.

	% exact agreement	1-point agreement	Kappa	Spearman's *r*	Intraclass correlation
Overall	51.0	96.0	.33	.58	.74
Reading	53.9	95.8	.36	.64	.80
Math	47.6	95.2	.29	.51	.68
District C	50.0	95.7	.27	.47	.68
District D	51.4	95.2	.34	.58	.74

Table 1B. Inter-rater reliability of lesson observation ratings, by Principle of Learning.

Principle of Learning	% exact agreement	Kappa	Spearman's *r*	Intraclass correlation
AT	51.0	.31	.63	.79
CE/SML	51.7	.30	.39	.60
AR: RC	50.0	.31	.61	.76
AR: Math	48.1	.27	.43	.47

Table 1C. Inter-rater reliability of lesson observation ratings, after merging rubric categories.

	% of exact agreement	Kappa
4 point scale (1-4)	51.0	.33
3 point scale (1, 2, 3 &4)	68.3	.40
3 point scale (1, 2 & 3, 4)	62.4	.41
3 point scale (1 & 2, 3, 4)	59.4	.38

Figure 1: Percent exact agreement over time.

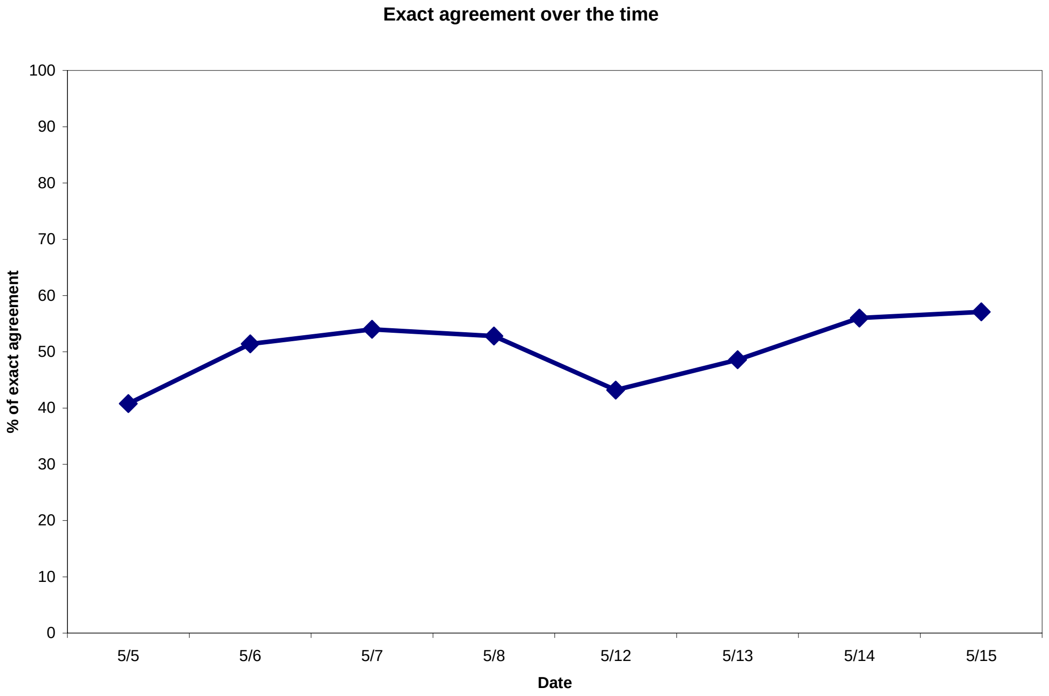


Table 2 gives variance components estimates in a variance components model for total scores of the AT rubrics, the CE rubrics, and separate totals for AR in Reading Comprehension and AR in Mathematics. The variance component for Rater for AT is zero because the raters divided the observational tasks to save time and produced only a single AT rating per rubrics per classroom. For each Principle of Learning, there is a sizable variance component for District, suggesting that total scores within Principle of Learning are sensitive to district differences. The Rubric variance component is also large, suggesting that the rubrics within each Principle do measure different aspects of that Principle. Finally, the Teacher variance component is much larger than the School variance component. This is consistent with other variance components and HLM analyses of school and teacher effects (and consistent with the earlier IQA pilot study of Crosson, et al., 2003): often, the teacher has a greater influence over the classroom environment than the school does.

Table 2. Estimates of variance components. Fixed effects were achievement rank (low, middle, high), grade (primary, upper), subject (reading, math); random effects were district, school, teacher, rater, item (rubrics)

Source of Variability	AT		CE		AR-Reading		AR-Math	
	Variance Components	% of Total Variance	Variance Components	% of Total Variance	Variance Components	% of Total Variance	Variance Components	% of Total Variance
District	0.199	13	0.328	29	0.319	19	0.090	13
School	0.082	5	0.026	2	0	0	0.050	7
Teacher	0.205	13	0.208	19	0.918	55	0.167	24
Rater	0[a]	0	0.054	5	0.052	3	0.064	9
Rubric	0.379	25	0.076	7	0.009	1	0.030	4
Residual	0.660	43	0.425	38	0.373	22	0.306	43

[a]. A negative variance component was set to zero.

Tables 3 and 4 explore differences between the districts on a rubric-by-rubric basis, for Accountable Talk (AT), Clear Expectations (CE) and Academic Rigor (AR) in Reading Comprehension and Mathematics. Table 3 gives means and standard deviations within each District for each rubric, and Table 4 reports the results of two-sample t-tests comparing the Districts. Histograms for the same data (for lesson observation ratings) are shown in Appendix B. In Tables 3 and 4, consensus ratings for each rubric, for n=17 classrooms in District C and n=s13 classrooms in District D, were used.

Referring to Table 4, all Accountable Talk rubrics, except for the participation and teacher linking rubrics, show significant difference in scores between the two districts, with district C scoring significantly higher than district D. Both of the Clear Expectations rubrics also showed significantly higher ratings, on average, for district C than district D. Only one of the Academic Rigor for Reading Comprehension rubrics, Lesson Activity, was scored significantly higher in District C than in District D. Rubrics reflecting the Academic Rigor of Classroom Discussion, and of Expecations (of student performance) were not significantly different. We believe this may be due to a confound with curriculum: anecdotal evidence suggests that District C’s reading comprehension curriculum was not sufficiently well-defined to support high-quality instruction.Two of the four Academic Rigor for Mathematics rubric differed significantly between the Districts (again favoring District C), the rubrics relating to Lesson Implementation and Expectations of student performance. Lesson Potential and Classroom Discussion rubrics did not show significant differences.

Table 3. Descriptive statistics for lesson observation rubrics, by District

AT Rubric	District C Mean (SD)		District D Mean (SD)	
AT1: Participation	3.53	(0.72)	3.15	(0.69)
AT2: Teacher's linking	2.53	(1.12)	1.92	(0.86)
AT3: Student's linking	2.47	(1.18)	1.23	(0.44)
AT4: Asking knowledge	3.53	(0.62)	2.69	(1.11)
AT5: Providing knowledge	3.59	(0.80)	2.46	(1.27)
AT6: Asking thinking	3.76	(0.44)	2.46	(1.13)
AT7: Providing thinking	3.65	(0.49)	2.38	(1.12)

CE Rubric	District C Mean (SD)		District D Mean (SD)	
CE1: Clarity and detail of expectations	3.00	(0.721)	1.85	(0.90)
CE2: Access to expectations	3.63	(0.51)	2.31	(1.18)

AR (Reading Comprehension) Rubric	District C Mean (SD)		District C Mean (SD)	
AR1: Discussion	3.11	(0.60)	2.60	(0.89)
AR2: Lesson activity	2.33	(0.71)	1.40	(0.55)
AR3: Expectations	2.44	(1.13)	1.40	(0.55)

AR (Mathematics) Rubric	District C Mean (SD)		District C Mean (SD)	
AR1: Potential	2.75	(.46)	2.50	(.76)
AR2: Implementation	2.63	(.52)	2.13	(.35)
AR3: Discussion	2.50	(.84)	1.80	(.84)
AR4: Expectations	2.88	(.64)	2.00	(.93)

Table 4. Between-District two-sample t-tests for lesson observation rubrics

	Mean Difference	T	df	Sig. (2-tailed)
AT1	0.38	1.445	28	.159
AT2	0.61	1.613	28	.118
AT3	1.24	3.594	28	.001
AT4	0.84	2.623	28	.014
AT5	1.13	2.987	28	.006
AT6	1.30	4.376	28	.000
AT7	1.26	4.164	28	.000

	Mean Difference	T	df	Sig. (2-tailed)
CE1	1.15	3.940	28	.000
CE2	1.22	3.829	28	.001

Reading Comprehension	Mean Difference	T	df	Sig. (2-tailed)
AR1	0.51	1.286	12	.223
AR2	0.93	2.542	12	.026
AR3	1.04	1.919	12	.079

Mathematics	Mean Difference	t	df	Sig. (2-tailed)
AR1	0.25	0.798	14	.438
AR2	0.50	2.256	14	.041
AR3	0.70	1.382	9	.200
AR4	0.88	2.198	14	.045

In Tables 5 and 6 we explore relationships among the rubrics by computing their correlations, across classrooms and districts. Tables 5A and 5B consider correlations of total scores within each Principle of Learning, separately for Reading Comprehension (Table 5A) and Mathematics (Table 5B). We can see that Academic Rigor correlates with both of the other two principles; this is consistent with the theoretical expectations underlying the Principles of Learning, in which Academic Rigor underlies all other aspects of instructional quality.

In Tables 6A and 6B we examine the same correlations, rubric-by-rubric. Here the results are less clear, but usually the strongest correlations occur between rubrics measuring the same Principle of Learning. This is suggestive that, with sufficient sample size, an appropriate factor analysis of the IQA rubrics would identify approximately simple structure breaking out according to each Principle of Learning. In Reading Comprehension, the Accountable Talk rubrics related to “atmosphere” in the classroom correlate more highly with Academic Rigor rubrics, and in Mathematics, Accountable Talk rubrics related to thinking and knowledge highly correlated with Academic Rigor. Clear Expectations rubrics tend to be correlated with Accountable Talk rubrics related to thinking and knowledge, and to many Academic Rigor rubrics.

Table 5A. Correlation of lesson scores in Reading Comprehension, by Principle of Learning

	AT	CE	AR
AT	-	.35	.68*
CE		-	.66*
AR			-

* p < .01

Table 5B. Correlation of lesson scores in Mathematics, by Principle of Learning

	AT	CE	AR
AT	-	.05	.64*
CE		-	.65*
AR			-

* p < .01

Table 6A. Correlation of lesson scores among Reading Comprehension rubrics

	AT1	AT2	AT3	AT4	AT5	AT6	AT7	CE1	CE2	AR1	AR2	AR3
AT1	-	.66*	.04	.47	.26	.23	.30	-.37	-.36	.38	.21	-.10
AT2		-	.38	.30	.53	-.07	.15	-.03	-.07	.56*	.38	.18
AT3			-	.13	.31	.17	.34	.56*	.38	.70**	.59*	.39
AT4				-	.64*	.29	.16	-.29	-.51	.38	-.09	-.25
AT5					-	.11	.24	.06	-.06	.59*	.07	.14
AT6						-	.89**	.02	-.08	.44	.00	-.06
AT7							-	.12	.07	.64*	.18	.15
CE1								-	.83**	.18	.59*	.68**
CE2									-	-.09	.44	.59*
AR1										-	.27	.13
AR2											-	.68**
AR3												-

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

Table 6B. Correlation of lesson scores among Mathematics rubrics

	AT1	AT2	AT3	AT4	AT5	AT6	AT7	CE1	CE2	AR1	AR2	AR3	AR4
AT1	-	.67**	.68**	.54*	.52*	.24	.17	.19	.25	.22	.29	.13	.34
AT2		-	.62**	.56*	.48	.46	.17	.12	-.05	.39	.15	.32	.35
AT3			-	.68**	.58*	.36	.32	.37	.29	.34	.17	.14	.54*
AT4				-	.93**	.38	.63**	.24	.20	.36	.23	.50	.57*
AT5					-	.45	.71	.28	.35	.43	.37	.63*	.62**
AT6						-	.68**	.74**	.46	.75**	.80**	.81**	.74**
AT7							-	.50	.38	.63**	.65**	.71*	.79**
CE1								-	.83**	.63**	.66**	.55	.73**
CE2									-	.41	.52*	.59	.55*
AR1										-	.71**	.82**	.89**
AR2											-	.68*	.68**
AR3												-	.82**
AR4													-

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

Assignment Ratings

Here we only briefly summarize some of the results for ratings of teachers' assignment portfolios; more complete analyses of the assignment portfolios for Reading Comprehension are provided by Clare Matsumura, et al. (2004), and more complete analyses of the assignment portfolios for Mathematics are given by Boston, et al. (2004).

Tables 7 and 8 give percent exact agreement, and Kappa and Spearman correlations, for individual rubrics used for scoring assignments. Overall, it appears that assignment rating is more reliable than lesson observation rating, although particular rubrics such as CE (clarity of expectations) may still suffer poor reliability.

Table 7. Inter-rater reliability of assignment ratings for the reading comprehension assignments (*N* = 52 assignments)

Rubric	% exact agreement	Kappa	Spearman's *r*
AR: Grist*	81.1	.66	.76
AR1: Potential	71.2	.59	.84
AR2: Implementation	69.2	.56	.84
AR3: Expectations	63.5	.51	.83
CE: Clarity of Expectations	44.2	.24	.56

*Note: *N* = 37 for this dimension. "Grist" is a measure of rigor inherent in the text being considered; however, it was not possible to rate the rigor of the text for every assignment (e.g. if the raters were unfamiliar with a text and could not locate it at the time of assignment ratings).

Table 8. Inter-rater reliability for AR: Math Rubrics for Lesson Observation and Assignment scores (N=54 assignments).

Rubric	% exact agreement	Kappa	Spearman's *r*
AR1: Potential	65.5	.51	.73
AR2: Implementation	60.0	.43	.72
AR3: Discussion	67.3	.53	.74
AR4: Expectations	62.7	.43	.68

Table 9 gives a brief variance components analysis for the assignment ratings, totaled within the AR and CE rubrics (AT depends on social interaction, and s othat is very similar to the variance components analysis of Table 2 above. It is interesting to note that Rater has a negligible variance component in this analysis. School again contributes almost nothing to the variance components model, when District, Teacher and Assignment are considered. District consistently contributes the largest variance component, again suggesting that the IQA rubrics can reliably make such distinctions. Finally, the variance component for Assignments is consistently second-largest, after District. This surprised us; past work by Clare Matsumura (2003) suggests that collecting four assignments ought to produce fairly stable estimates of the quality of assignment. In retrospect, however, we think that our instructions to teachers, to include two challenging and two recent assignments, acted to artificially increase this variance component (since Challenge and Academic Rigor, for example, are often related). Better might be to have the teachers simply include four recent assignments.

Table 9. Estimates of variance components, reading comprehension assignments. Fixed effects were achievement rank (low, middle, high), grade (primary, upper), subject (reading, math); random effects were district, school, teacher, rater, item (rubrics), and assignment.

	AR		CE	
Source of Variability	Variance Components	% of Total Variance	Variance Components	% of Total Variance
District	0.378	31	0.864	57
School	0[a]	0	0	0
Teacher	0.214	18	0.172	11
Rater	0	0	0	0
Assignment	0.351	29	0.171	11
Rubric	0.019	2	-*	-
Residual	0.246	20	0.318	21

[a]. A negative variance component was set to zero.

* CE only has one rubric.

Tables 10 and 11 provide between-district comparisons on a rubric-by-rubric basis, for assignment ratings, that is entirely analogous to Tables 3 and 4. Histograms for the same data are provided in Appendix C. Once again, there are significant differences between districts, favoring District C, except for two AR rubrics in Reading Comprehension classrooms. Once again we suspect a confound with curriculum here; with the curriculum in District C apparently unable to support fully academically rigorous work for students.

Table 10. Descriptive statistics for assignment ratings, by District

Rubrics (Reading Comprehension)	District C Mean (SD)		District D Mean (SD)	
AR1: potential	2.44	(1.05)	2.10	(.85)
AR2: implementation	1.91	(1.00)	1.60	(.68)
AR3: expectations	2.63	(1.07)	1.75	(.85)
CE: clarity of expectations	2.97	(.80)	1.50	(.61)
Rubrics (Mathematics)	District C Mean (SD)		District D Mean (SD)	
AR1: Potential	3.15	(.53)	1.93	(.72)
AR2: Implementation	2.63	(.79)	1.61	(.69)
AR3: Rigor in response	2.67	(.78)	1.50	(.79)
AR4: Expectations	3.07	(.39)	1.96	(.58)
CE: clarity of expectations	3.19	(.62)	1.71	(.76)

Table 11. Between-District two-sample t-tests for assignment ratings

Rubrics (Reading Comprehension)	Mean Difference	T	Df	Sig. (2-tailed)
AR1	.34	1.271	50	.210
AR2	.31	1.316	50	.194
AR3	.88	3.262	50	.002
CE1	1.37	6.728	50	.000
Rubrics (Mathematics)	Mean Difference	t	Df	Sig. (2-tailed)
AR1	1.22	7.176	53	.000
AR2	1.02	5.113	53	.000
AR3	1.17	5.483	53	.000
A44	1.15	8.430	53	.000
CE1	1.47	7.847	53	.000

Results on the relationships between rubrics are discussed in great detail by Clare Matsumura et al. (2004) and Boston et al. (2004), and so we will omit a detailed discussion here. Broadly speaking the results are similar to what is seen in Tables 6A and 6B above: rubrics tend to hang together better, within and between Principles of Learning, in Mathematics than they do in Reading Comprehension. Further refinement of the Reading Comprehension rubrics and observation protocols may be needed, for both lesson observation and assignment rating, to improve their reliability and cohesiveness.

Discussion

The Spring 2003 Pilot Study was designed to answer three major questions; we consider each in turn.

Rater reliability

Can naïve external raters be trained to reliably rate the IQA? If so, then the IQA can be developed as an “turnkey package” of rating materials and rater training materials, that can be shared with school districts and other organizations who wish to train and use their own raters. If not, then at least in the short term, the IQA should be rated only by a limited number of raters carefully trained and monitored by IQA staff to maintain high reliability. An IQA rater training program was developed and used in this study, to answer this question.

Exact agreement between trained raters was only moderate (47.6-51.0%), but agreement to within one scale point was quite good (95.2-96%); moreover exact agreement increased markedly over the time course of the study. If naïve raters are to be trained for rating the IQA, then a longer training period involving some “live”, out of sample, rating, seems to be required.

Variance components analysis using various total scores showed small effects for raters, and larger effects for teachers and rubrics than schools, within each district. These results tend to support the notion that the IQA could be scored by trained external raters, if only total scores per Principle of Learning, or total scores at some higher level of aggregation, are desired. For reliable scoring of individual rubrics, however, it seems likely that both the rubrics themselves, as well as the rater training program, will have to be further refined. Until then, a limited number of raters carefully trained and monitored by the IQA team should be used, when high reliability of rating individual rubrics is desired.

Differentiation between Districts

The study of Crosson et al. (2003) established that the IQA was sensitive to variation in teachers' practice. However, that study was not designed to establish a relationship between IQA score variation and degree of effort or success in implementing instructional practices consistent with the Principles of Learning. Two districts with differing levels of involvement in activites related to the Principles of Learning were compared in this study, to try to answer this question.

Variance components analysis in the present study displayed a strong variance component for District, suggesting that total IQA scores could well-differentiate between districts with different levels of instructional quality related to the Principles of Learning. Between-district comparisons of consensus scores show that most individual rubrics on the IQA are sensitive to these differences. Indeed, most rubrics showed significant differences between the districts, favoring district C, which had a longer involvement in efforts to implement instructional practices consistent with the Principles of Learning, with typical average raw score differences of one scale point or more per rubric between districts; even rubrics that did not show significant differences showed trend effects favoring district C. The least sensitive rubrics were for Academic Rigor in reading comprehension, but this may be due to a confound with curriculum: anecdotal evidence suggests that District C's reading comprehension curriculum was not sufficiently well-defined to support high-quality instruction.

Relationship among Rubrics

How are the various rubrics in the IQA related to one another? Do they appear to be functioning independently of one another? Are they so closely connected that some can be dropped without loss of reliability? Although full answers to this question require psychometric methods such as multidimensional item response theory and factor analysis, that demand much larger sample sizes than were available in the present study, we did explore this question with correlational analyses.

Here the findings were mixed, but somewhat encouraging. Broadly speaking, we found that AR rubrics tend to be more highly correlated with one another than rubrics for AT or CE.

Moreover, AR rubrics tend to be correlated with rubrics from AT and CE, supporting the notion embedded in the Principles of Learning that Academic Rigor underlies Accountable Talk and Clear Expectations (and indeed, all of the Principles), at least when these are observed according to the observational protocols taught in the IQA rater training program. We also found that rubrics used for observing and scoring Mathematics lessons and assignments tended to cohere as a scale more (i.e. higher intercorrelations) than when used for observing and scoring Reading Comprehension lessons and assignments, and rubrics for scoring lesson observations were somewhat more coherent than those used for assignment scoring. This latter phenomenon may be partly due to our request for "two challenging and two recent assignments" in each teacher's assignment portfolio, which induced more assignment variability (for example, as measured in a variance components analysis) than past research (e.g., Clare Matsumura, 2003) lead us to expect. Overall, it is encouraging that in some cases we do have fairly coherent sets of rubrics, and studying what makes these rubrics work well may help us to improve the others; it is also encouraging that in the case of assignment rating, a simpler request of teachers may significantly improve the rubric score data.

Conclusions and Future Work

Educators, policy-makers, and researchers need to be able to assess the efficacy of specific interventions in schools and school districts. Despite a plethora of reform programs in place across the nation to improve the quality of teaching, the success of these ventures generally has been assessed in one way—through student outcome scores on standardized tests of achievement. This approach has limitations: If a new professional development intervention is improving achievement scores, what feature of instruction has changed? If professional development is not improving student scores, is it because the intervention is not changing instructional practice, or because the changes in instructional practice did not affect achievement? A tool is needed to provide snapshots of instructional practice itself, before and after implementing new professional development or other interventions, and at other regular intervals to help monitor and focus efforts to improve instructional practice.

The *Instructional Quality Assessment (IQA)* has been developed to provide an essential tool in determining what works in professional development and instructional practice. In addition to the results shown here, Clare Matsumura et al. (2004) explore the use of the IQA to assess the quality of reading comprehension assignments, Boston et al. (2004) explores the IQA as an instrument to assess instructional quality in mathematics lessons, and Wolf et al. (2004) explores the use of the IQA as a measure of high quality classroom talk. Of course there is much more to do, but these results taken together suggest that the IQA is developing into a useful tool to assess many aspects of instructional practice.

Given the measure of success that we have had so far, it is now possible to consider several technical questions that must be answered before the IQA can be broadly used: (1) Is the IQA useful as a measure of instructional quality, as broadly defined in current scientific thinking (e.g., NRC, 1999a,b)? We are planning to conduct at least one study of the IQA involving a district that has been successful as measured by student achievement, and employs professional development that is consistent with NRC (1999a,b) guidelines but is not organized around the Principles of Learning, to try to answer this question. (2) Is the IQA equally predictive of achievement gains on all types of state achievement tests, or does the IQA measure instructional quality that is only relevant to certain kinds of student achievement tests (high cognitive demand, open ended response)? By comparing IQA scores (based on richer ratings of classroom practice and teachers' assignments) with student achievement in states using student assessments of varying degrees of richness, we hope to be able to address this question rigorously. In addition we will look at which rubrics and sources of evidence for the IQA are most closely tied to student learning across different content areas. Finally, (3) What are the sample size and data collection (sampling design) needs for reliable inference from the IQA in practice? We plan to develop sampling design guidelines for various IQA applications. We will also explore whether, through training or experience, raters can become reliable enough that only one rater per classroom (instead of pairs of raters as in all preliminary work to date) can be used.

In addition to developing the IQA as a rigorous "external" or summative evaluation of program interventions, it is important to leverage the development effort for IQA in two ways: to provide a system of feedback for schools and districts about professional development resources should be targeted most effectively; and to provide descriptors of good instructional practice—clear expectations for teachers' performance—that can themselves be the basis of professional development efforts. Determining just what faculty's learning needs are and where to focus professional development can be extremely challenging for instructional leaders, especially if they are in the initial stages of learning about expert instruction themselves. Likewise, teachers who are just beginning to develop a vision of highly effective instructional practices can hardly be expected to identify their own learning needs. Crosson et al. (2004) discuss some ways in which the IQA may be useful as the basis for a broader set of formative professional development tools. We anticipate work on this "formative" variant of the IQA to continue in parallel with our development of the IQA as a useful external assessment tool.

References

Abbott, M. L., & Fouts, J. T. (2003). *Constructivist Teaching and Student Achievement: The Results of a School-level Classroom Observation Study in Washington*. Lynnewood, Washington: Washington School Research Center, Seattle Pacific University. (Accessed December 2003 at http://www.spu.edu/wsrc/currentresearch.html).

Applebee, A.N. (1984). *Contexts for learning to write: Studies of secondary school instruction*. Norwood, NJ: Ablex.

Ball, D. L, Camburn, E., Correnti, R., Phelps, G. & Wallace, R. (1999). *New tools for research on instruction: A web-based teacher log.* Working paper, Center for Teaching Policy. Seattle: University of Washington.

Ball, D. L., & Cohen, D. K. (1999). Developing practice, developing practitioners: Toward a Practice-based theory of professional education. In L. Darling-Hammond & G. Sykes (Eds.), *Teaching as a learning profession* (pp. 3-31). San Francisco, CA: Jossey-Bass Publishers.

Beck, I.L. & McKeown, M.G. (2002). Text Talk: Capturing the Benefits of Read-Aloud Experiences for Young Children. *The Reading Teacher, 55(*1), 10-20.

Beck, I.L., McKeown, M.G, Hamilton, R.L., & Kucan, L. (1997). *Questioning the Author*. Newark, Delaware: International Reading Association.

Brown, A.L, Palinscar, AS & Armbruster, BB (1994). Instructing comprehension fostering activities in interactive learning situations. In Ruddell, R. B., Ruddell M.R., & Singer, H. (1994*). Theoretical Models and processes of reading (Fourth Edition), Chapter 36*. Newark: Delaware: International Reading Association.

Boston, M. and Wolf, M.-K. (2004). *Using the Instructional Quality Assessment (IQA) Toolkit to Assess Academic Rigor in Mathematics Lessons and Assignments.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego CA, April 2004.

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

Clare Matsumura, L. (2000). *Using teacher's assignments as an indicator of classroom practice*. (Center for the Study of Evaluation Technical Report #532). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Clare Matsumura, L. (2003). Teachers' assignments and student work as indicators of instructional quality. *Educational Researcher*.

Clare Matsumura, L., Wolf, M.-K., Crosson A., Levison, A. Peterson, M., Resnick, L., and Junker, B. (2004). *Assessing the Quality of Reading Comprehension Assignments and Student Work.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego CA, April 2004.

Carr, M S., Braunger, J., & Hagans, R. (1998). *A Regional Depiction: Standards-Based Curriculum Reform in the Northwest.* Portland OR: Northwest Regional Edcuation Laboratory. (Accessed December 2003 at http://www.nwrel.org/scpd/ci/regdepiction/).

Cohen, D.K., & Ball, D.L. (1994). Relations between policy and practice: A commentary. *Educational Evaluation and Policy Analysis, 12*, 249-256.

Cohen, D.K, McLaughlin, M.W., & Talbert, J.E. (Eds). (1993). *Teaching for understanding:Challenges for policy and practice.* San Francisco: Jossey-Bass.

Crosson, A., Junker, B. W., Clare Matsumura, L., & 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.

Crosson, A., *et al.* (2004). *The Instructional Quality Assessment as a professional development tool.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego CA, April 2004.

Cuban, L. (1990). Reforming again, and again, and again. *Educational Researcher*, *19*, 3-13.

Darling-Hammond, L. (2000, January). Teacher quality and student achievement: A review of state policy evidence. *Educational Policy Analysis Archives*, *8*(1). Retrieved February 28, 2001 from http://olam.ed.asu.edu/epaa/v8n1

Doherty, K.M. (2003). Education Issues: Standards. Education Week on the Web. (Accessed December 2003 at http://www.edweek.org/context/topics/issuespage.cfm?id=55)

Fink, E., & Resnick, L. B. (2001). Developing principals as instructional leaders. *Phi Delta Kappan*, 82 (8), p598-606.

Finn, J. D., & Achilles, C. M. (1999). Tennessee's class size study: Findings, implications, and misconceptions. *Educational Evaluation and Policy Analysis, 20*(2), 97-109.

Fosnot, C. T. (1996). *Constructivism theory, perspectives, and practice.* New York: Teachers College Press.

Fouts, J. T., Brown, C. & Thieman, G. (2002). *Classroom Instruction in Gates Grantee Schools: A Baseline Report Prepared for the Bill & Melinda Gates Foundation*. Seattle, WA: The Bill & Melinda Gates Foundation. (Accessed December 2003 at http://www.gatesfoundation.org/Education/ResearchandEvaluation/).

Fuhrman, S.H. (1993). *Designing coherent education policy: Improving the system*. San Francisco: Jossey-Bass.

Fuhrman, S.H. (1999). The New Accountability, *Consortium for Policy Research in Education*, Policy Briefs. (Accessed December 2003 at http://www.cpre.org/Publications/rb27.pdf).

Henke, R.R., Chen, X., and Goldman, G. (1999). *What Happens in Classrooms? Instructional Practices in Elementary and Secondary Schools: 1994-95 (NCES 1999-348)*. U.S. Department of Education. Washington, DC: U.S. Government Printing Office. (Accessed December 2003 at http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=1999348).

Hiller J.H., Fisher G.A., & Kaess W (1969). Computer investigation of verbal characteristics of effective classroom lecturing. *American Educational Research Journal, 6(4)*: 661-675.

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

Kameenui, E.J., & Carnine, D.W. (1998). *Effective teaching strategies that accomodate diverse learners.* Upper Saddle River, NJ: Merrill.

Klein, S., Hamilton, L., McCaffrey, D., Stecher, B., Robyn, A., & Burroughs, D. (2000). *Teaching Practices and Student Achievement: Report of the First-Year Findings form the "Mosaic" Study of Systematic Initiatives in Mathematics and Science*. Santa Monica, CA: RAND.

Lieberman, A. (1994). Teacher development: Commitment and challenge. In P. Grimmett & J. Neufield (Eds.), *Teacher development and the struggle for authenticity* (pp. 15-30). New York: Teachers College Press.

Linn, R.L., & Baker, E.L. (1998, Fall). School quality: Some missing pieces. *CRESST Line*. Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing.

Lohr, S.L. (1999). *Sampling: Design and Analysis.* Pacific Grove CA: Brooks-Cole.

Marshall, S., Fuhrman, S., and O'Day, J. (1994). National Curriculum Standards: Are They Desirable and Feasible? In R. Elmore and S. Fuhrman (Eds.), *The Governance of Curriculum: 1994 Yearbook of the Association for Supervision and Curriculum Development.* Alexandria, VA: Association for Supervision and Curriculum Development.

Mayer, D.P. (1999a). Measuring instructional practice: Can policymakers trust survey data? *Educational Evaluation and Policy Analysis, 21,* 29-45.

Mayer, D.P. (1999b). Invited commentary: moving toward better instructional practice data. *Education Statistics Quarterly [online]*. Washington, DC: National Center for Educational Statistics (Accessed December 2003 at http://nces.ed.gov/pubs99/quarterly/summer/2-feature/2-esq12-c.html).

McCulloch, C.E. & Searle, S.R. (2001). *Generalized, linear, and mixed models.*New York: John Wiley and Sons.

National Center for Educational Progress (2003). *NAEP -- Background Questionnaires*. (Accessed December 2003 at http://nces.ed.gov/nationsreportcard/bgquest.asp).

National Center on Education and the Economy and the University of Pittsburgh (1999). *Reading and Writing Grade-by-Grade: Primary Literacy Standards for Kindergarten through Third Grade* Washington DC.

National Commission on Excellence in Education (1983). *A Nation at Risk: The Imperative for Educational Reform.* Washngton, DC: Author. Currently available at http://www.ed.gov/pubs/NatAtRisk/ (Accessed December 2003).

National Commission on Teaching and America's Future (1996). *What matters most: Report of the National Commission on Teaching and America's Future*. Kutztown, PA: Author.

National Education Goals Panel (2000). *History, 1989-Present.* Washington, DC: Author. (Accessed December 2003, at http://www.negp.gov/page1-7.htm).

National Reading Panel (2000). *Teaching Children to Read: Reports of the Subgroups*. Washington, DC: National Institute of Health and Human Development.

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

Pearson, P.D. & Fielding,L.(1991). Comprehension Instruction. In R. Barr, M.L. Kamil, P. Mosenthal, P.D. Pearson (Eds.), *Handbook of Reading Research (Volume II).* New York: Longman.

Porter, A. (1991). Creating a System of School Process Indicators. *Educational Evaluation and Policy Analysis 13(1)*: 13-29.

Pressley, M. (1998). *Reading Instruction that Works: The Case for Balanced Teaching*. NY: Guilford Publications.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests (expanded ed.).* Chicago: The University of Chicago Press. (Originial work published 1960).

Regents of the University of Michigan (2001). Study of Instructional Improvement [Homepage]. (Accessed December 2003 at http://www.sii.soe.umich.edu/).

Resnick, L. B. (1995). From aptitude to effort: A new foundation for our schools. *Daedalus, 124,* 55-62.

Resnick, L. B. (1999, June 16). Making America smarter. *Education Week,* pp. 38-40.

Resnick, L. B., Glennan, T. K., and Lesgold, A. (2001). *Taking Powerful Teaching to Scale.* Analytic Essay Prepared for the Spencer Foundation.

Resnick, L. B., & Hall, M. W. (1998). Learning organizations for sustainable education reform. *Daedalus, 127*, 89-118.

Resnick, L. B., & Hall, M. W. (2001). The Principles of Learning: Study tools for educators. [CD-ROM, version 2.0]. Available: www.instituteforlearning.org

Resnick, L. B., & Harwell, M. (2000, August*). Instructional variation and student achievement in a standards-based education district* (CSE Tech. Report No. 522). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Resnick, L. B., & Nelson-Le Gall, S. (1997). Socializing intelligence. In L. Smith, J. Dockrell, & P. Tomlinson (Eds.), *Piaget, Vygotsky and beyond* (pp. 145-158). London/New York: Routledge.

Rowan, B., Schilling, S. G., Ball, D. L., & Miller, R.. (2001) *Measuring Teachers' Pedagogical Content Knowledge in Surveys: An Exploratory Study*. Consortium for Policy Research in Education, Study of Instructional Improvement, Research Note S-2. Ann Arbor: University of Michigan.

Saunders, W., Goldenberg, C., & Hamann, J. (1992). Instructional conversations beget instructional conversations. *Teaching and teacher education, 8,* 199-218.

Shavelson, R., McDonnell, L., Oakes, J., Carey, N., & Picus, L. (1987). *Indicator Systems for Monitoring Mathematics and Science Education*. Santa Monica, CA: The RAND Corporation.

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

Silver, E. A. & Stein, M.K. (1996). The QUASAR Project: The "Revolution of the Possible" in Mathematics Instructional Reform in Middle Schools. *Urban Education, 30(4)*, 478-521.

Snow, C.E., Burns, & Griffin, P. (1998). *Preventing Reading Difficulties in Young Children*. Washington D.C.: National Academy Press.

Snow, C.E. (2002). Reading for Understanding: Toward an R&D Program in Reading Comprehension. Washington, DC: OERI.

Spillane, J.P., & Zeuli, J.S. (1999). Reform and teaching: Exploring patterns of practice in the context of national and state mathematics reforms. *Educational Evaluation and Policy Analysis, 21*(1), 1-27.

Sproule, R. (2000). Student Evaluation of Teaching: A Methodological Critique of Conventional Practices. *Education Policy Analysis Archives, 8(2).* (Accessed December 2003 at http://epaa.asu.edu/epaa/v8n50.html).

SPSS Incorporated (2001). *SPSS 11.0.* Chicago, IL: SPSS, Inc.

Stein, M.K., Smith, M.S., Henningson, M.A., & Silver, E.A. (2000*). Implementing Standards-Based Mathematics Instruction.* NY: Teachers College Press.

Stigler, J.W., Gonzales, P., Kawanaka, T., Knoll, S., & Serrano, A. (1999). *The TIMSS Videotape Classroom Study.*

Study of Instructional Improvement (2001). *Overview of SII Survey Instruments. Consortium for Policy Research in Education*, Study of Instructional Improvement. Ann Arbor: University of Michigan.

Tharp, R. G., & Gallimore, R. (1988). Rousing minds to life: Teaching, learning, and schooling in social context. New York: Cambridge University Press.

Tyack, D., & Tobin, W. (1994). The “grammar” of schooling: Why has it been so hard to change? *American Educational Research Journal*, *31*(3), 453-479.

United States Department of Education (2002). No Child Left Behind: A Desktop Reference. http://www.ed.gov/admins/lead/account/nclbreference/index.html (accessed December 2003).

Wolf, M.-K., Crosson, A. and Resnick, L. (2004). *Classroom talk for rigorous reading comprehension instruction.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego CA, April 2004.

Wright, S. P., Horn, S.P. & Sanders, W.L. (1997). Teacher and Classroom Context Effects on Student Achievement: Implications for Teacher Evaluation. . In *Journal of Personnel Evaluation in Education 11*. 57-67.

Appendix A: Summary of the Spring 2003 IQA Pilot Study Design

Table A1: Assignment of raters for classroom observation, District C. Trained external raters labeled R1-R6; IQA staff raters labeled S1-S3.

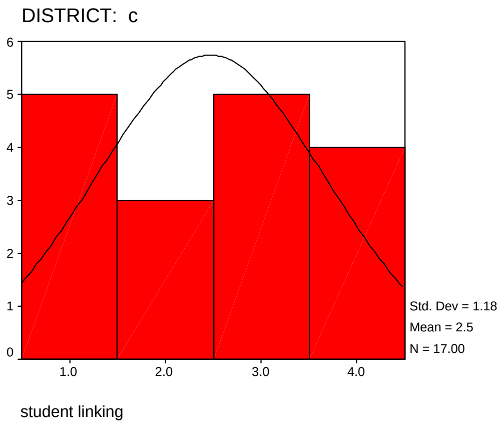
	External		IQA Staff					Achievement
	Rater Pair		Member	Grade	Content	Teacher	School	Rank
1	R2	R5	S3	4	Math	B	V	Low
2	R3	R4	S2	2	Reading	B	U	Middle
3	R2	R6	S3	3	Reading	A	T	Middle
4	R3	R5	S2	3	Math	B	T	Middle
5	R3	R5	S2	2	Math	A	U	Middle
6	R1	R5	S1	2	Math	C	U	Middle
7	R1	R5	S1	4	Math	D	U	Middle
8	R2	R4	S3	4	Reading	E	U	Middle
9	R2	R4	S1	2	Math	A	V	Low
10	R3	R6	S2	4	Math	A	W	Low
11	R2	R4	S3	2	Reading	A	X	High
12	R1	R6	S1	2	Reading	A	Y	High
13	R3	R5	S2	2	Reading	B	Y	High
14	R1	R6	S1	4	Math	C	Y	High
15	R3	R5	S2	5	Reading	D	Y	High
16	R1	R6	S1	4	Reading	A	Z	Middle
17	R2	R4	S3	4	Reading	B	Z	Middle

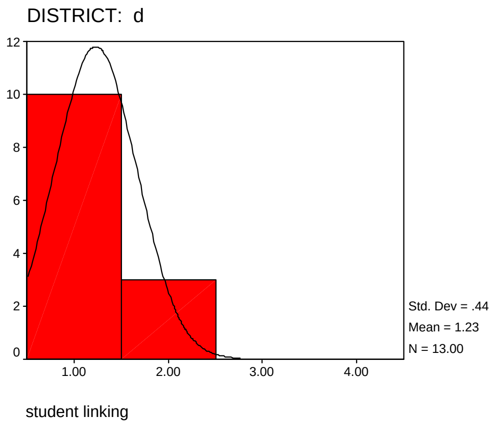
Table A2: Assignment of raters for classroom observation, District D. Trained external raters labeled R1-R6; IQA staff raters labeled S1-S3.

	External		IQA Staff					Achievement
	Rater Pair		Member	Grade	Content	Teacher	School	Rank
18	R1	R4	S1	2	Reading	A	A	High
19	R1	R4	S1	2	Math	B	A	High
20	R2	R4	S3	2	Math	C	A	High
21	R3	R6	S2	2	Reading	A	B	High
22	R3	R6	S2	4	Reading	B	B	High
23	R2	R5	S3	4	Math	D	B	High
24	R1	R4	S1	2	Math	E	B	High
25	R3	R5	S2	4	Math	F	B	High
26	R3	R4	S2	2	Reading	A	D	Low
27	R2	R6	S3	2	Math	B	D	Low
28	R2	R6	S3	4	Reading	C	D	Low
29	R1	R6	S1	4	Math	D	D	Low
30	R6	--	S3	4	Math	A	E	Middle

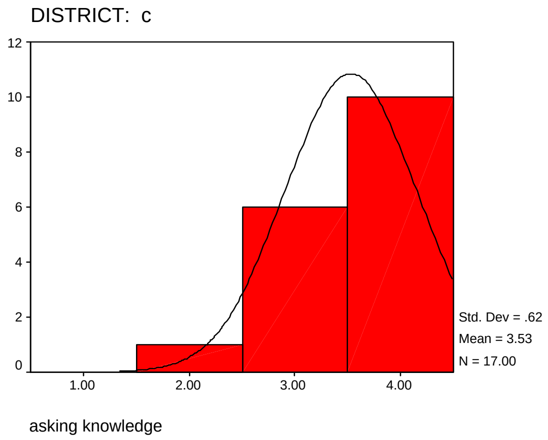
Appendix B: Histograms for Lesson Observation Ratings

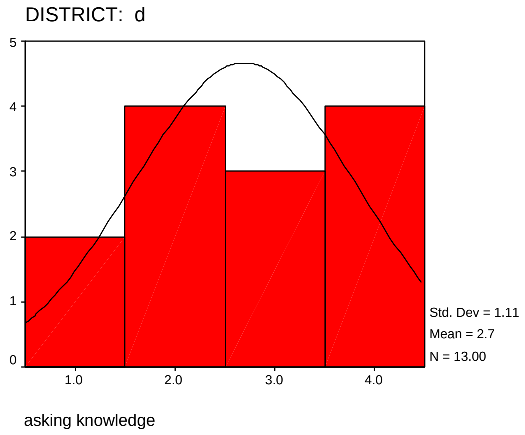
AT3: Student's linking (Lesson Observation)



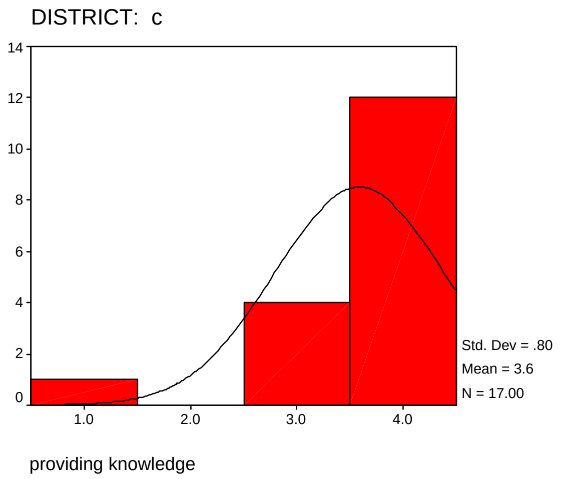


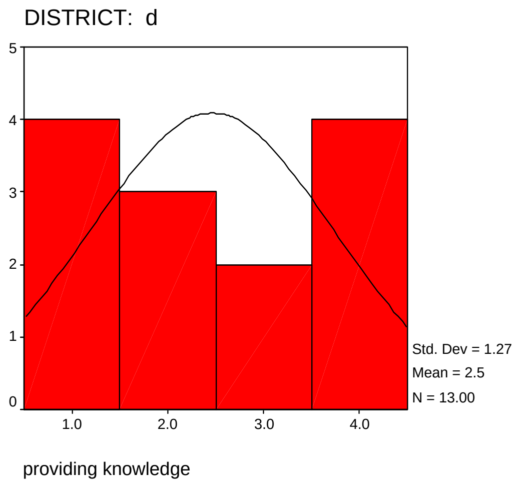
AT4: Asking Knowledge (Lesson Observation)



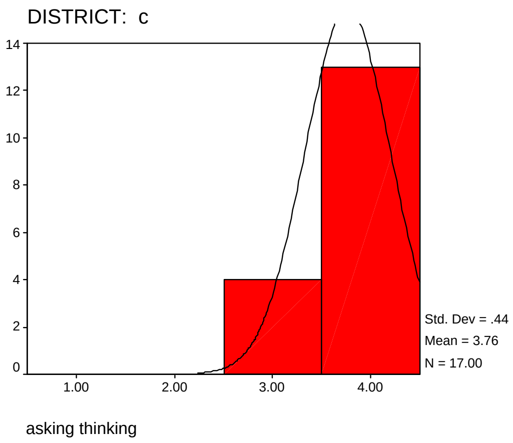


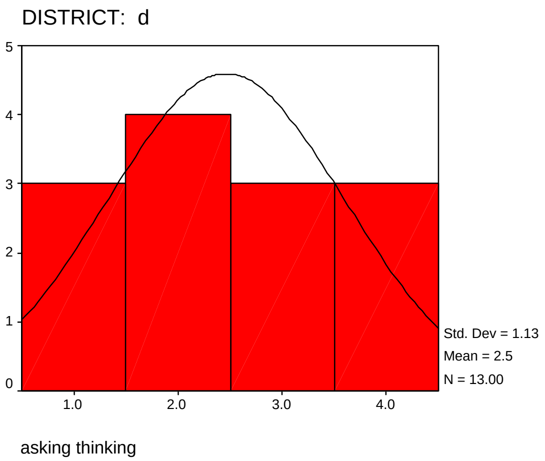
AT5: Providing Knowledge (Lesson Observation)



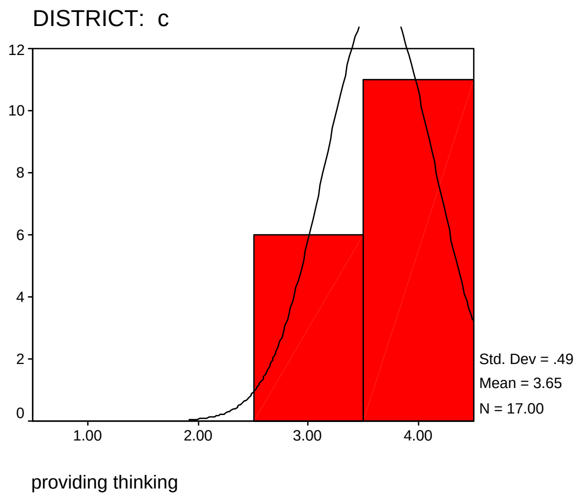


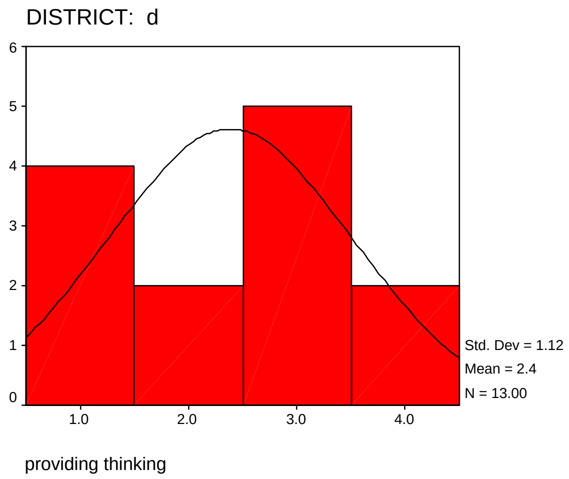
AT6: Asking Thinking (Lesson Observation)



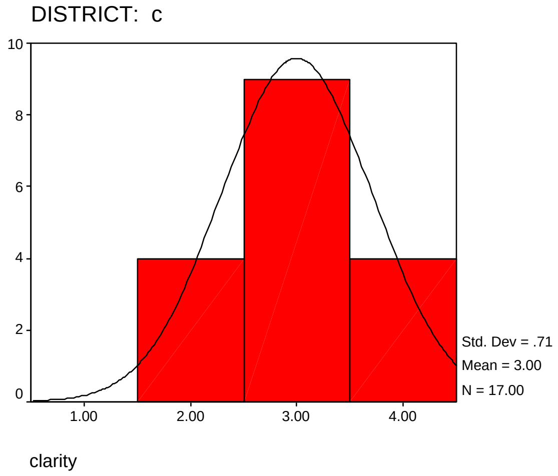


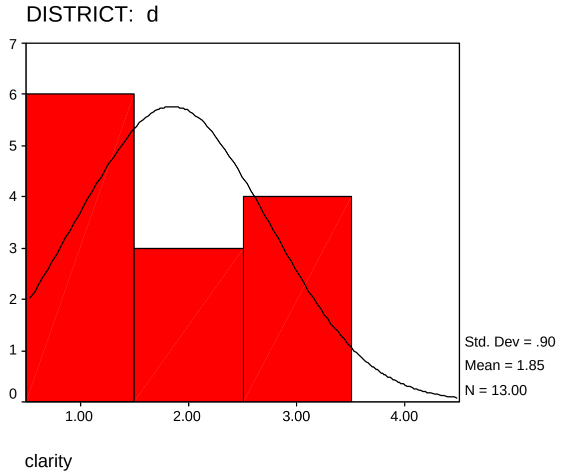
AT7: Providing Thinking (Lesson Observation)


DISTRICT: c

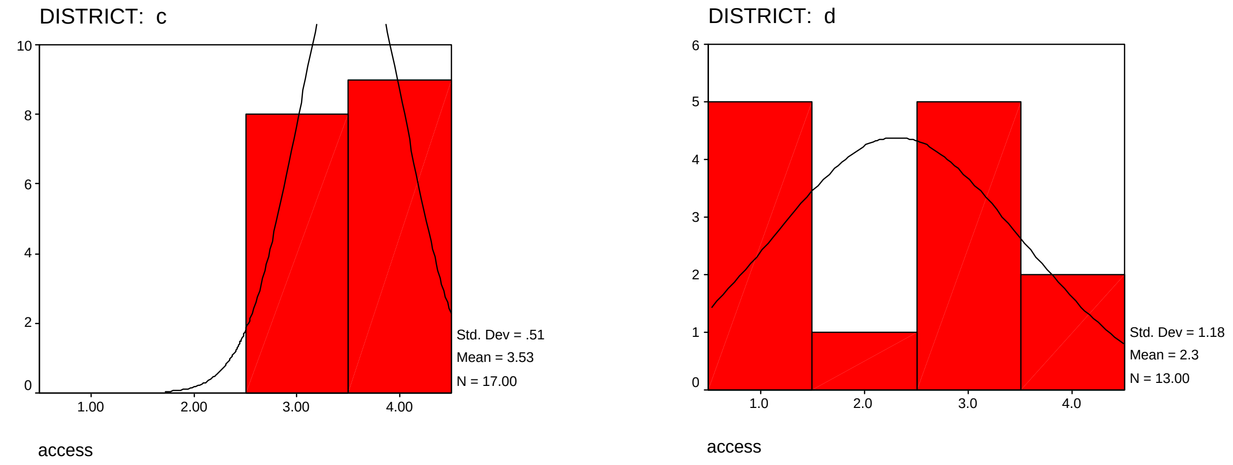

DISTRICT: d

CE1: Clarity and Detail of Expectations (Lesson Observation)

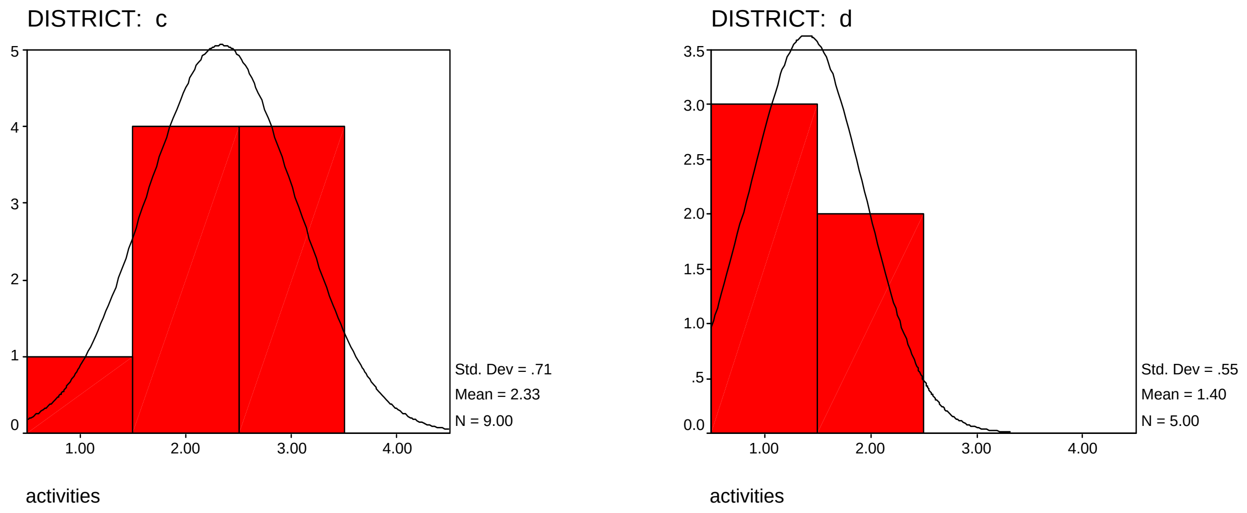

DISTRICT: c


DISTRICT: d

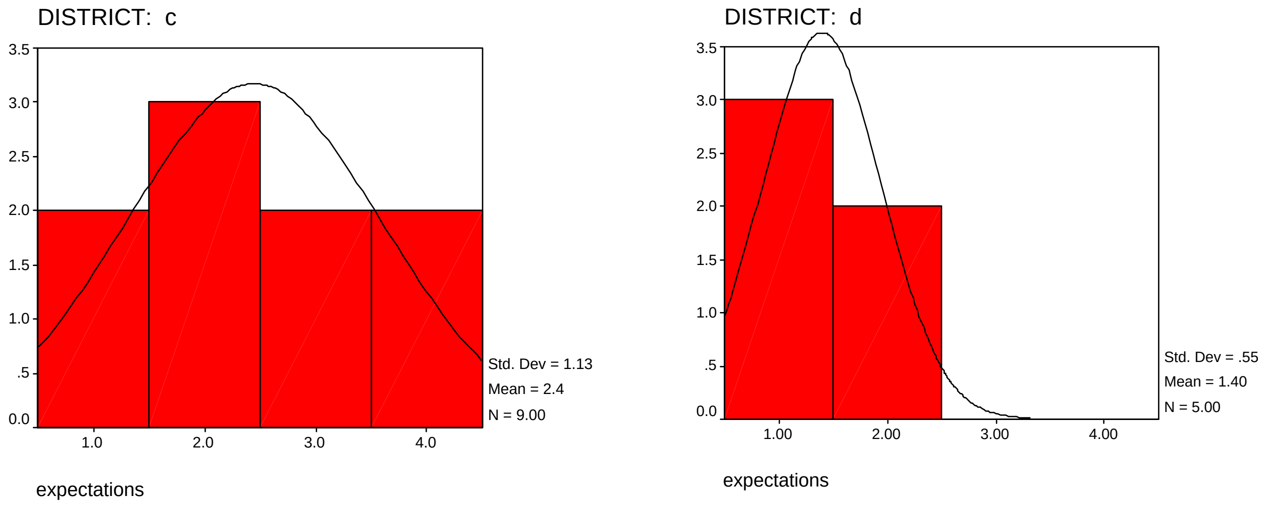
CE2: Access to Expectations (Lesson Observation)



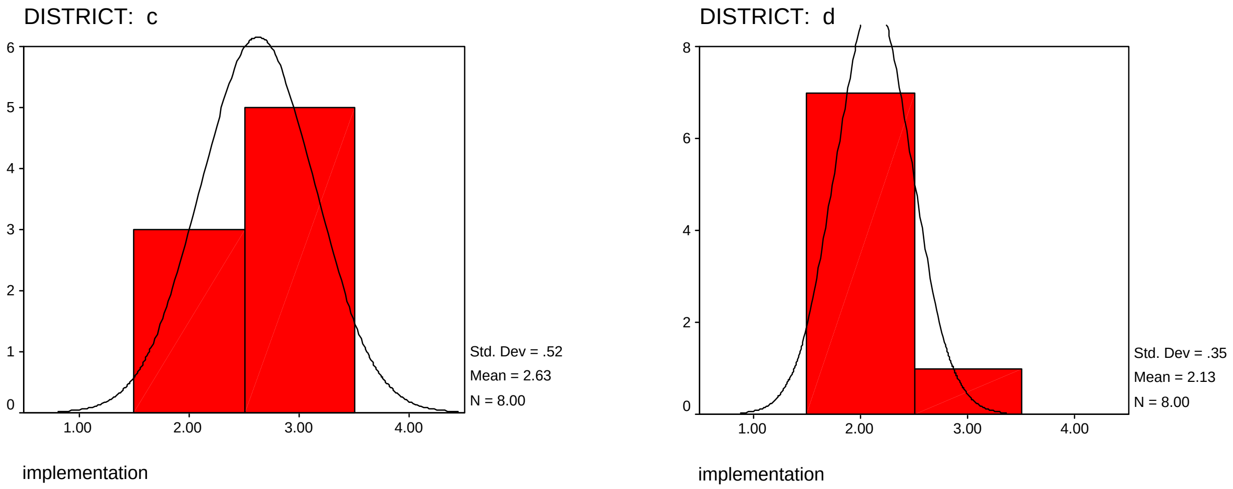
AR1: Rigor of Activities, Reading Comprehension (Lesson Observation)



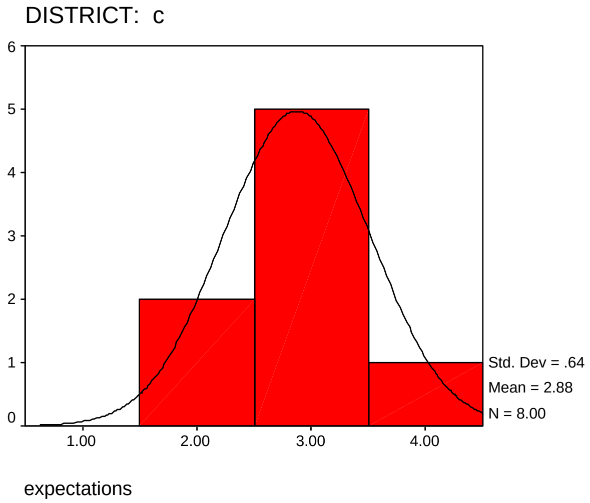
AR2: Rigor of Expectations, Reading Comprehension (Lesson Observation)

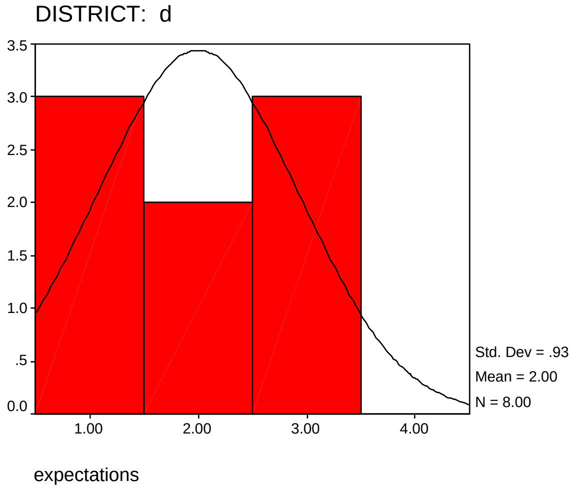


AR2: Rigor of Implementation, Mathematics (Lesson Observation)



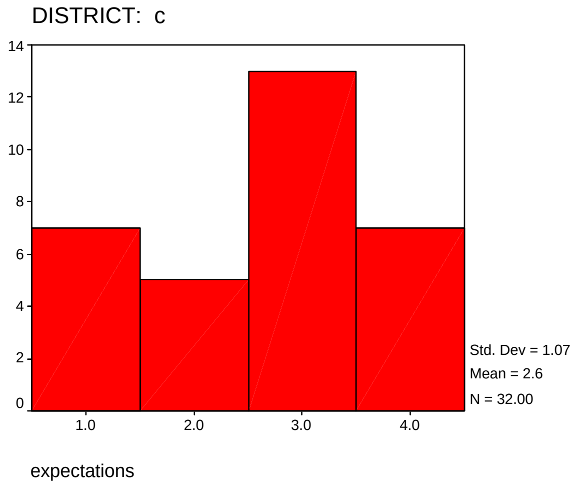
AR4: Rigor of Expectations, Mathematics (Lesson Observation)

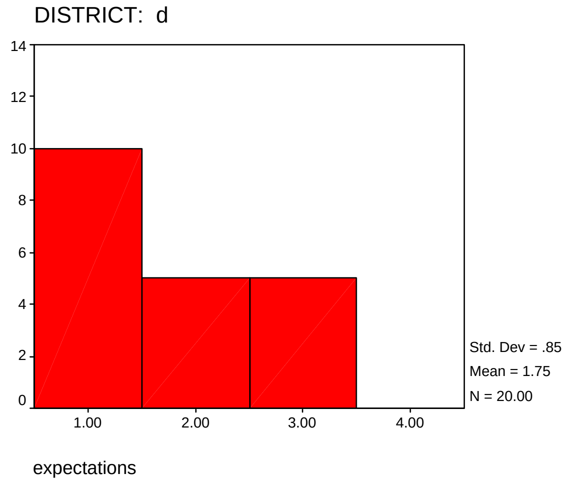




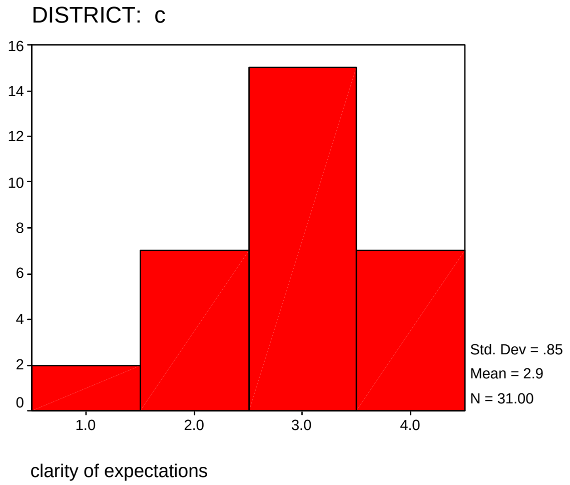
Appendix C: Histograms for Assignment Ratings

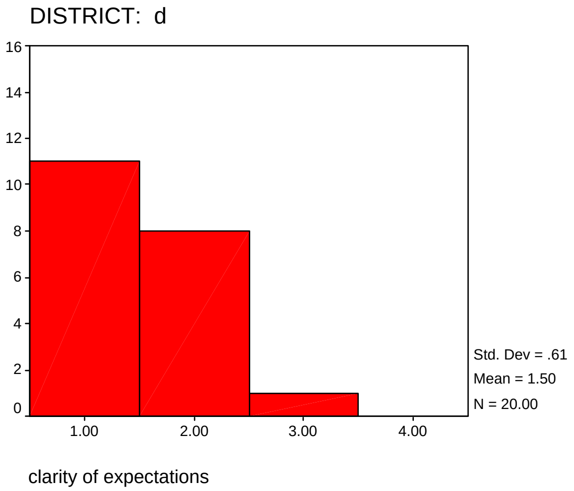
AR3: Rigor of Expectations, Reading Comprehension (Assignments)



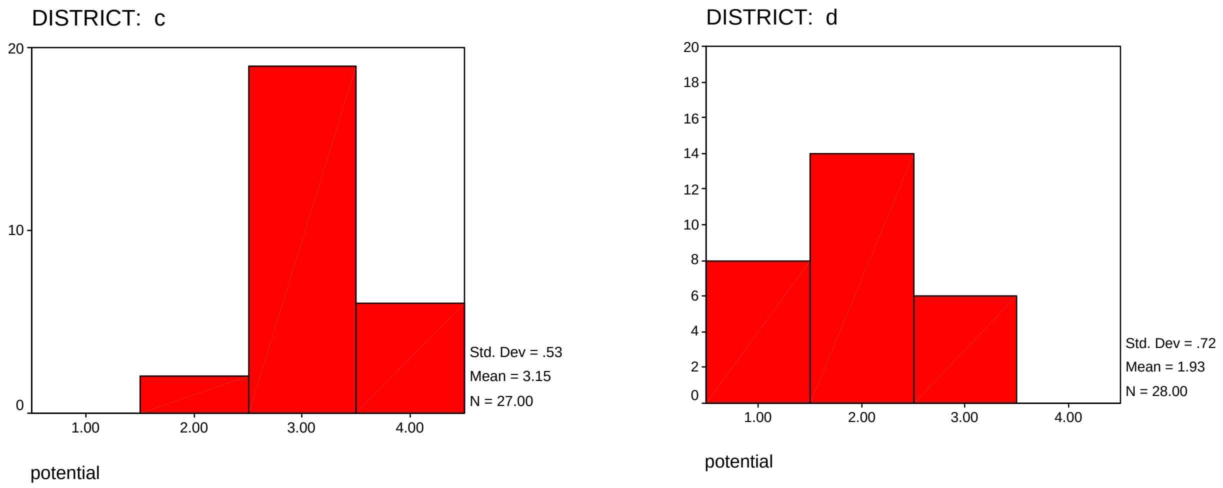


CE: Clarity of Expectations, Reading Comprehension (Assignments)

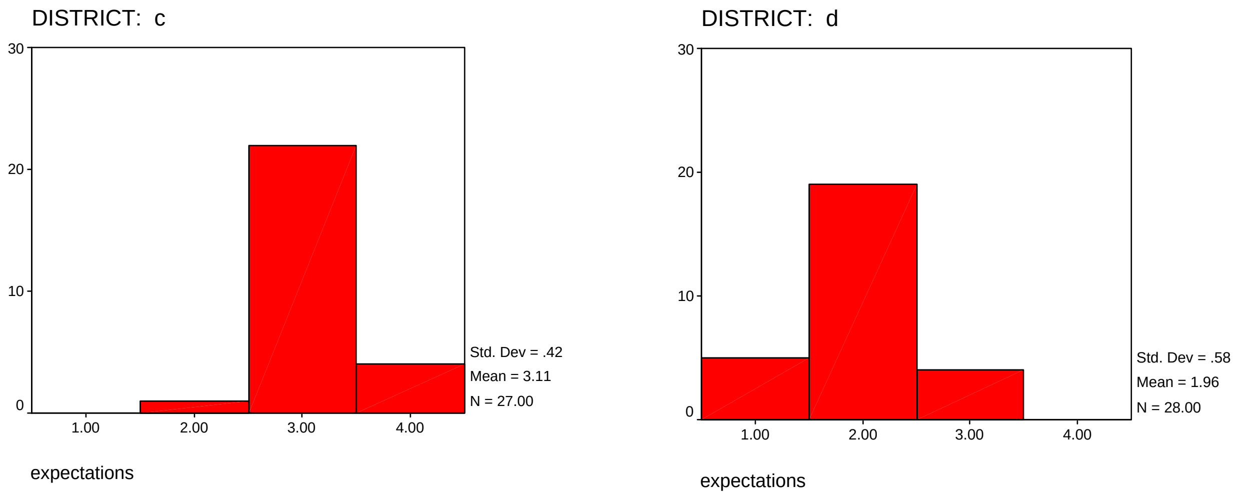




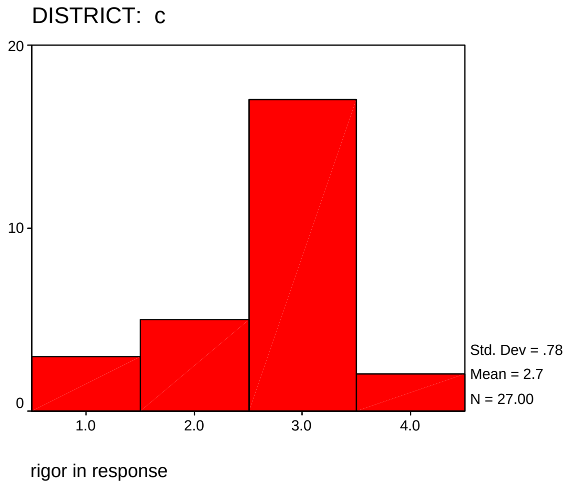
AR1: Rigor of Assignment Potential, Mathematics (Assignments)

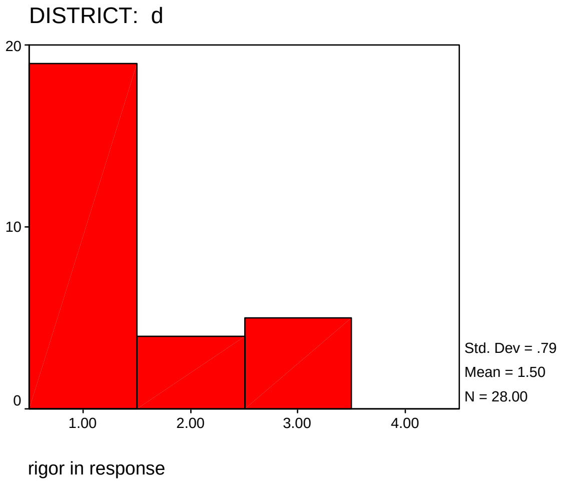


AR2: Rigor of Assignment Implementation, Mathematics (Assignments)

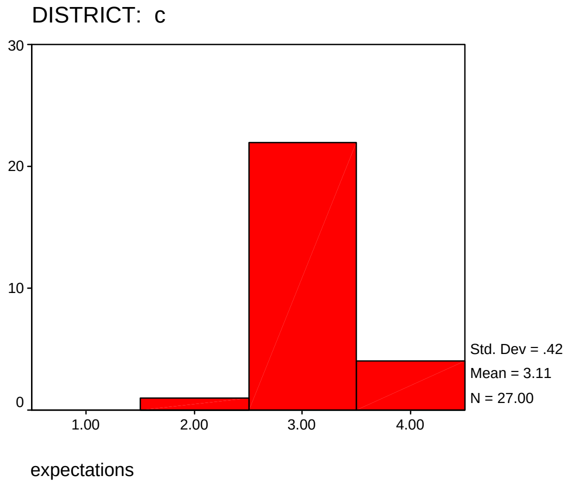


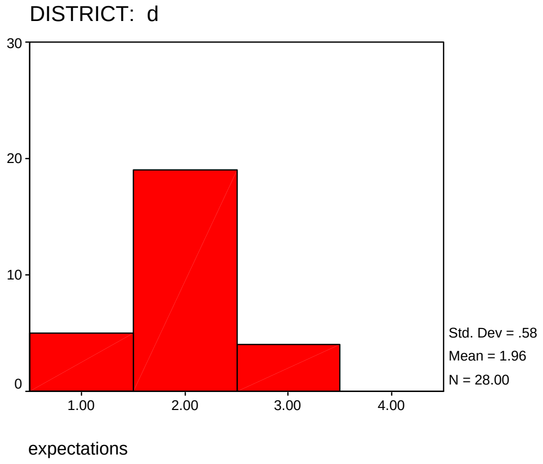
AR3: Rigor in Response, Mathematics (Assignments)



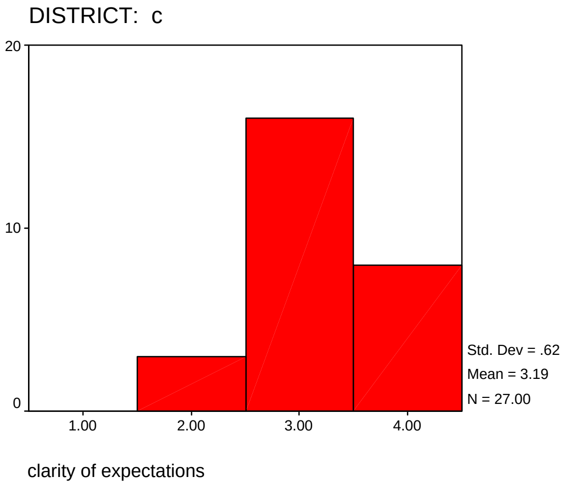


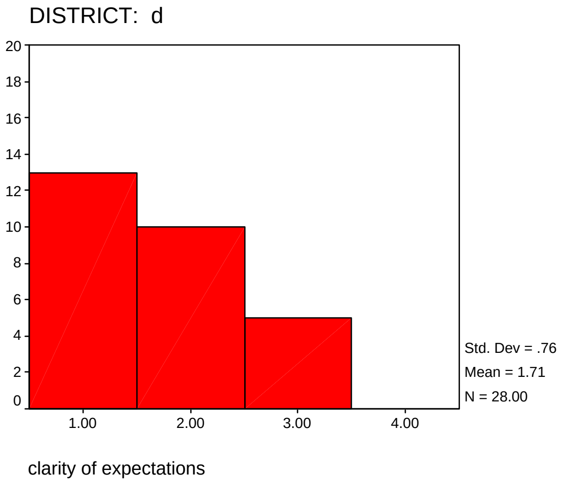
AR4: Rigor of Expectations, Mathematics (Assignments)





CE: Clarity of Expectations, Mathematics (Assignments)


DISTRICT: c
Std. Dev = .62
Mean = 3.19
N = 27.00
clarity of expectations


DISTRICT: d
Std. Dev = .76
Mean = 1.71
N = 28.00
clarity of expectations