Documentation of the 1996-2002 Chicago Annenberg Research Project Strand on Authentic Intellectual Demand Exhibited in Assignments and Student Work

A Technical Process Manual

August 2002

Stacy Wenzel
Jenny Nagaoka
Loretta Morris
Sabrina Billings
Carol Fendt

Consortium on Chicago School Research 1313 East 60th Street Chicago, IL 60637
773-702-3364 773-702-2010 - fax
www.consortium-chicago.org

Documentation of the 1996-2002 Chicago Annenberg Research Project Strand on Authentic Intellectual Demand Exhibited in Assignments and Student Work: A Technical Process Manual

August 15, 2002

Contents

Written by

Staff of the Chicago Annenberg Research Project:
Sabrina Billings, Carol Fendt, Loretta Morris, Jenny Nagaoka and Stacy Wenzel.

While this document reflects a collaborative effort and the work of many colleagues, specific individuals wrote each section. Billings and Morris wrote the data management subsections of section B. Fendt wrote section C. Nagaoka wrote sections D and F. Wenzel wrote sections A, E, F and the data collection subsection of section B. All added documents to the Appendix. Fendt and Wenzel edited the manual.

For additional information on this manual, please contact: Stacy Wenzel, swenzel@uic.edu , 312-413-1872. For additional information on the Chicago Annenberg Research Project and the Consortium on Chicago School Research, see www.consortium-chicago.org .

Overview

Project History

In March 1996, the Consortium on Chicago School Research (CCSR) proposed to evaluate the Chicago Annenberg Challenge's (CAC) newly funded reform efforts in the Chicago public schools. The evaluation aimed to understand how schools involved in CAC efforts developed as organizations and, in turn, how these developments benefited students. The study aimed to look not just at student outcomes based on standardized test or survey data but to deepen knowledge on student learning by examining collected samples of assignments given and the work students produced in response. By fall 1996, the Chicago Annenberg Research Project (CARP) at CCSR was funded and collection of assignments and student work had begun. The initial funding would allow data collection through 1999-2000. However in 2000, the Chicago Annenberg Challenge provided an additional grant to CARP that allowed expanded research with an additional enlarged sample of assignments and student work collected in 2000-01.

The project completed data collection and analysis for samples of assignments and work from academic years: 1996-97, 1997-1998, 1998-99, and 2000-01. The first, or baseline, data collection took place in the 1996-97 or 1997-98 school years, depending on when schools joined the CAC reform initiative. The second major data collection point was in 1998-99 when assignments but not student work was sampled. The third and last major data collection point took place in the 2000-01 school year. For the description of the research process that follows, we refer to the 1996-97 school year as Year 1, 1997-98 as Year 2, 1998-99 as Year 3, 1999-2000 as Year 4, and 2000-01 as Year 5.

In this manual, we describe the process by which we collected, scored, and analyzed the thousands of teacher assignments and pieces of student work collected from our sample of Chicago public schools participating in Chicago Annenberg Challenge-funded networks. After providing a brief background of this work, we highlight three main aspects of the process: (a) collection and management of the data, (b) assignment of scores to the data, and (c) statistical analyses of the data. In addition, we offer some insights on the overall management of this large project and make observations on lessons we have learned that may prove useful the next time researchers undertake this type of work.

Intellectual Foundation

The intellectual foundation for this strand of research drew on previous work by Fred Newmann and Gary Wehlage at the University of Wisconsin at Madison's Center on the Organization and Restructuring of Schools.[1] As a member of the CARP leadership team, Newmann actively shaped this strand of research. Yet new ideas from the CARP team of researchers along with the broad scope and longitudinal nature of the research design

[1] Newmann, Fred M. and Gary G. Wehlage. (1995). *Successful School Restructuring: A Report to the Public and Educators.* Madison, WI: Center on Organization and Restructuring of Schools and Newmann, Fred M. (1996). *Authentic Achievement: Restructuring Schools for Intellectual Quality*. San Francisco: Jossey-Bass.

allowed for new findings on teaching and learning and provided added insight into methodological issues.

In this study, we were particularly concerned with student academic learning that includes and goes beyond the acquisition of basic knowledge and skills to include deep understanding of subject matter and the ability of students to produce "authentic" intellectual work. This involves the development of cognitive capacities that allow students to work with existing knowledge and then create new knowledge that allows them to analyze and solve real-world problems. This also includes students' abilities to communicate and explain their ideas to others in elaborated ways. This kind of learning is the type that students will need as adults who can manage their personal affairs, and become economically productive and responsible members of society.[2]

Findings

The findings based on this strand of research are available in numerous CARP reports. Here we note briefly a selection of key findings from these reports. As we discuss in detail in this manual, the sample of schools from which we base our findings was selected purposively and cannot be used to generalize to all Chicago public schools. Still, these findings have important implications for Chicago and the broader education community. In the sample of schools in our study:

- In 1997, the typical classroom assignment in both writing and mathematics made only modest academic demands on students.[3]
- Students showed more authentic intellectual work in writing than in mathematics.[4]
- Students whose assignments were more authentic produced more authentic intellectual work in both writing and mathematics.[5]
- The quality of classroom assignments improved between 1997 and 1999.[6]
- In 1999, the level of challenge in mathematics assignments still remained quite low. For example, more than 80 percent of sixth and eighth grade math assignments provided only minimal or no challenge.[7]
- Students who received assignments requiring more challenging intellectual work also achieved greater than average gains on the Iowa Test of Basic Skills (ITBS) reading

---

[2] Newmann, Fred M., Anthony S. Bryk, and Gudelia Lopez. (1998). *The Quality of Intellectual Work in Chicago Schools: A Baseline Report.* Chicago: Consortium on Chicago School Research.
[3] Newmann et al. (1998).
[4] Newmann et al. (1998).
[5] Newmann et al. (1998).
[6] Bryk, Anthony S., Jenny K. Nagaoka, and Fred M. Newmann. (2000). *Chicago classroom demands for authentic intellectual work: Trends from 1997–1999.* Data brief of the Chicago Annenberg Research Project. Chicago: Consortium on Chicago School Research.
[7] Bryk et al. (2000).

and math and demonstrated higher performance in reading, math, and writing on Illinois Goals Assessment Program (IGAP).[8]

- Both students with high and low prior achievement levels benefit from being exposed to high quality assignments.[9]

Data Collection and Management

Significant effort and resources were needed to obtain the assignments and student work documents and then to prepare them so that they could be analyzed systematically. Researchers worked with teachers who provided a strategic sample of documents. The researchers then carefully prepared the documents by adding a unique identification code and cleaning off all identifying information. Researchers turned documents over to CARP data managers who electronically and physically archived the data.

Collection from Schools

Sampling networks and schools. Sample selection began with the networks. In 1996 and 1997, more than 40 networks of schools and external partners were awarded multi-year implementation grants by the Chicago Annenberg Challenge. These networks included a maximum of 211 schools a year, approximately 90 percent of which were elementary schools. From these networks and schools, we selected an initial sample of 11 networks. We selected networks with diverse organizational foci, networks with both newly formed and well-established relationships with schools, and networks with different types of external partners (universities, school educational organizations, community organizations, and cultural institutions).[10] Next we selected sample schools from the networks. We selected two or three schools as research sites from each of these networks. One to two schools were chosen because of their promise for working well with their external partners and succeeding in their efforts to develop. An additional school was chosen because of indications that it might struggle to succeed. Our intention was to create a purposive sample of schools that would allow us to understand reasons for more or less successful development. Consortium survey data and assessments of the network external partners informed our school sample section.

We selected our sample of networks and schools in two stages. A first group was selected in the fall of 1996 from the networks and schools that received the first round of Annenberg funding. A second group was selected in the fall of 1997 from those receiving

---

[8] Newmann, Fred M., Anthony S. Bryk, A., and Jenny K. Nagaoka. (2001). *Authentic Intellectual Work and Standardized Tests: Conflict or Coexistence?* Chicago: Consortium on Chicago School Research.

[9] Newmann et al. (2001).

[10] See the following for more on the characteristics of Chicago Annenberg Challenge implementation networks: Newmann, Fred M., and Karin Sconzert. (June 2000*). School Improvement with External Partners*. Chicago: Consortium on Chicago School Research.

funding in the second round. In all, our original sample included 18 elementary and middle schools and five high schools.

From the original sample of schools, we experienced some attrition over the years. In 1996-97 we collected assignments and work from 12 elementary schools and by 1997-98 we had added schools so we collected from 18 elementary schools and 5 high schools. This changed for 1998-99 when several schools opted out of our research and we added a few replacement schools leaving us with 17 elementary schools and 5 high schools. In 2000-01 we chose to focus on 14 elementary schools and did no collection in high schools.

Although we did not intend to select a group of schools that was demographically representative of all Annenberg schools, the 14 elementary schools that make up our longitudinal field research sample are quite typical of schools across CAC and the Chicago public system as a whole. Of these 14 elementary schools, six enrolled primarily African-American students, three enrolled primarily Latino students, three enrolled a combination of both African-American and Latino students (at least 85 percent of the total enrollment), and two enrolled a more mixed group that included between 15 and 30 percent white students. On average, 32 percent of students in our field research schools scored at or above the national average in reading on the 1999 Iowa Tests of Basic Skills (ITBS), and 37 percent scored at this level in math. Our field research schools ranged from 17 to 60 percent of students at or above the national norms on the ITBS in reading and 16 to 78 percent of students at or above national norms in math. Average student enrollment for the schools was 900, ranging from 600 to 1,600 students. Table 1 in the Appendix shows the characteristics of these groups of schools.

Sampling classrooms and teachers. From each school in the study, we aimed to collect data from the following numbers of classrooms. In elementary schools, we chose 2 math classrooms and 2 language arts classrooms from each of three grade-levels (3, 6, and 8). We chose these grade levels because in Chicago public schools, they were grades targeted by high profile policies to end social promotion through high stakes testing. In high schools, we chose 3 math classrooms and 3 English classrooms from each of two grade-levels (9 and 10). The classrooms eligible for selection had to be instructed primarily in English so that data collected did not require translation.

In elementary school primary and sometimes secondary grades, the same teachers instruct students in both math and language arts. Thus, for example, while we had four "classrooms" to study in grade 3 at one school, we may have only worked with two teachers—both in self contained classrooms where they taught both math and language arts. In some schools, we were not able to work with our specified number of classrooms if it was a small school or if specific teachers opted out of the research. In other schools, we may have chosen to work with more than our specified number of teachers given teacher interest and our goal to collect a larger sample.

Which teachers we collected data from within a grade level varied from school to school. In some schools, we worked with all teachers with math and language arts classes. In other schools, the principal or contact person for the school's CAC network suggested teachers who we invited to participate in the study. In other schools, all teachers in eligible grades met with our researchers who explained the project. Those interested then volunteered.

Sampling of assignments and work. During the years of this study, the amount of data collected from each classroom varied. There were also changes in the times of year when samples were collected. Further there were two types of assignments collected—(a) typical or time–sampled and (b) challenging. These were collected in both math and language arts/English. In Years 1 and 2, reading assignments and student work were collected but not included in this research.

Researchers were instructed to ask consistently for typical assignments explaining that we wanted assignments teachers had given to students during the last week. The assignments needed to require students to produce some type of written work. It could be something students did in class, at home, or both. It could be an assignment that students worked on alone or in a group. See Document 2 in Appendix; this is the face sheet that was given to teachers to accompany the typical assignments they turned in.

The challenging assignment needed to be one that the teacher considered among the most challenging and important given to their students—a written assignment that gave them the best sense of how well their students were learning and understanding a subject or skill at their highest level. Similar to the typical/time sampled assignment, the assignment asked students to produce some type of written work and could be something students did in class, at home, or both. However, the challenging assignment had to be one that students worked on alone.

Student work that corresponded to this challenging assignment was collected. This work had to be in written form and able to be photocopied. It had to be individual work—not group work. See Document 3 in the Appendix; this is the face sheet that was given to teachers to accompany the challenging assignments and student work they turned in.

In Year 1, two typical and two challenging assignments and two student work samples were collected from each classroom. In Years 2 and 3, four typical and two challenging assignments and two student work samples were collected from each classroom. In order to increase the reliability of our analyses, in Year 5 we increased the sample size to six typical and two challenging assignments and two student work samples. The annual schedules for these collections are shown in Table 4 in the Appendix

The challenge of the collection process. While assignments and student work were created as regular classroom educational activities, a great deal of additional effort by teachers and researchers was required to bring these documents into the possession of our research project. How researchers introduced teachers to the work and then

maintained rapport with them was no trivial matter—it was the foundation on which this whole research enterprise rests.

The initial contact with a school typically was initiated by the researcher through a CAC network contact person or the school principal. How to access teachers was negotiated and then the researcher met teachers as a group to explain:

- Overall project goals
- Type of data teachers contribute
- Voluntary nature of participation
- Confidentiality
- Timetable for data collection

When teachers agreed to participate, researchers spoke with them individually to determine when the most convenient times would be to stop at the school and pick up assignments and student work. Some teachers preferred researchers to meet them before or after school while others where able to meet during prep periods.

In principle this process sounds straightforward. In reality it was not. A few researchers reported that collecting from some teachers was easy. They visited the school at the time agreed upon at the last meeting, collected documents, thanked the teacher and left. This was rare. More likely in order to walk away with a set of data documents, a researcher used reminder phone call, faxes, and notes to let the teacher know they would be by to pick up data. Even with these reminders, to collect the data from one teacher for one time point during the year, it was not unusual for the researcher to have to visit the school two or three times. With the recognition that a researcher worked with a school of three to a dozen participating teachers, it is easy to see that many trips needed to be made to the school.

When at the school collecting data, the researchers found themselves spending additional time. When collecting student work, researchers needed to photocopy the work and return originals to the teacher. Some researchers were able to photocopy at their school's main office while others did the photocopying at the CARP office. At times, researchers needed to write out the verbal instructions that teachers gave to their students in assignments if the teacher had no written document with this information. At other times, teachers needed more time to write something out themselves or to find the documents, so researchers needed to wait for them.

In Years 1, 2, and 3, teachers gave us their assignments and student work with no promised reward for their participation. At the end of each of these years we did express our gratitude with small gifts (i.e., coffee mug, pen and pencil set) to the teachers and to the school (i.e., flower arrangements, donuts). At the start of Year 5, teachers were offered an honorarium of $100 for full participation in the research—that is they must contribute a complete sample of assignments and work, plus allow us to interview, survey and observe them.

All of this work to collect data took place within a context that many teachers sometimes found annoying or even threatening. The time needed to pull together

assignments and student work was just one more demand on teachers who had very little time to do anything during the workday other than attend to the immediate needs of their students. In some schools, teachers were also not happy to work with researchers who they characterized as always coming to their school, asking for things and never contributing anything in return. Still in other cases, teachers were wary of sharing any evidence of what they were doing in their classroom, fearing that this data would be used by their principals to evaluate them.

Protection of human subjects and informed consent. All procedures we used in this research was pre-approved by the University of Chicago's Institutional Review Board which is charged with protecting human subjects from research risk. To minimize the risks posed to participants in this study, all data collected was held in confidence by researchers. This meant that we did not present research findings that could be linked to an identifiable school, network, or individual. Therefore, we made sure our research could not be used by the Chicago Annenberg Challenge to influence funding of schools nor could it be used by school administrators for accountability purposes.

We also worked to assure that all participants understood the voluntary nature of contributing to the study and did not feel coerced to participate. In some cases, principals recommended to us which teachers they wanted in the study. However when we spoke to these teachers individually, we made it clear that participation was their individual decision. If they opted out of the study, we did not inform the principal of this.

Participants were informed of what they were agreeing to do. With participating teachers sharing assignments and work, we received their verbal consent for participation informally. The IRB did not require us to collect formal written consent from adult participants. However, in order to collect student work, student assent and parent permission was required in written form. A sample form is included as Document 5 in the Appendix. Researchers gave a stack of consent forms to teachers, who distributed them to students and parents. The forms are "passive consent forms," that is, students and parents only returned them to the teacher (who gave them to the researcher) if they did not agree to contribute data. The consent form was written in English on one side and translated into Spanish on the other side of the paper.

Scope of work. To understand the size of the data collection undertaken in this study, we present below some quantities of the documents, teachers and schools involved.

	Yr 1	Yr 2	Yr 3	Yr 4	Yr 5
• Total assignments collected	349	953	715	0	896
• Teachers contributing assignments and/or work	74	116	87	0	71
• Schools contributing assignments	12	23	22	0	14

Management and Archiving of Data

With the huge quantity and steady influx of assignments and student work, it was necessary to have in place a streamlined system of data management and archiving, in which a strong attention to detail and accuracy was critical. With clear procedures in place, we were able to stay on top of the data that our fieldworkers brought in on an almost daily basis. At the end of the year, our well-maintained records and files would prove to be invaluable in organizing the assignments and student work for scoring.

Coding documents. In order to aid in the management of the data, as well as to ensure the confidentiality of the participants, we developed a coding system in which every assignment and every piece of student work received a unique identification code. For assignments, the code consisted of six parts: a school ID, an indication of whether the assignment was typical or challenging, a teacher ID, the subject, the collection number, and the year of collection. In this way, the code 2C03EM2Z represents a challenging (C) math (M) assignment from teacher 03E in school 2. It is the second such collection in the project's fifth year (2Z). To the extent that it was possible, school and teacher IDs remained consistent from year to year, so that, for example, school number 2 was always school number 2, even if number 1 dropped out of the study. Likewise, teacher 03E remained 03E, unless she changed to another grade. Once that teacher stopped being 03E, that ID was retired for her school. This consistency facilitated cross-year comparison of teacher participation.

The codes for student work were comprised of an assignment code as described above, plus the student's official Chicago Public Schools-issued ID. Before coding student work, however, a couple of preliminary steps were required. First, we had to check our Student ID Database (see Document 6 in the Appendix) in which we kept a record of the students whose parents denied permission for them to participate in the study. If assignments from these students were included in the sample, we pulled them out. Among the remaining pieces of student work, we randomly selected ten pieces to be used for the scoring process—selecting them using a random number list (See Document 7 in the Appendix). This procedure guaranteed that if a teacher had submitted the pieces of work in any particular order (last name, performance, etc.), these biases would be eliminated from our study. Occasionally, a teacher turned in fewer than ten pieces of student work for a given assignment, in which case all submitted pieces were used.

The assignment codes and student work codes were used in several databases, which allowed us to track the schools, teachers, subject matter, and students participating in the project, within and across years. The Teacher and School Code Database (see Document 8 in the Appendix) was updated at the beginning of the school year and throughout the year based on information received from our fieldworkers. Fieldworkers were also responsible for bringing in lists of student IDs for the classes from which they were collecting, which served as the basis of our Student ID Database.

Each fieldworker was given a list of the relevant school, teacher, and student IDs, based on the information they had provided us. The fieldworkers were thus able to code their own assignments and student work, saving the office staff considerable time. Codes

were written or typed on small labels and attached to the upper right side of each page of each piece of data collected (see a sample of this as Document 9 in the Appendix). Over the course of the project, we found that consistent placement of these labels on the documents was crucial. When dealing with such large quantities of data, minor details such as this, when not managed systematically, can become a major disruption.

Cleaning documents. We cleaned all incoming data both to ensure participant confidentiality. Because the data would also be reviewed and assigned a score (see next section on "Assigning quantitative scores to data") we also did this to avoid providing any background information that might influence the scorer when evaluating a piece of data. Cleaning the data involved removing from the student work, handouts, and assignments any words that could identify the student, school personnel, the school or its neighborhood, and any other identifying information. All teacher evaluations of students' work were also removed (See Document 10 in the Appendix for an example of marks to be cleaned off).

Through a certain amount of trial and error, we discovered a good method for document cleaning. We used black marker to cover these identifiers and applied a correction stick over the top of this. Alone, neither material did a sufficient job. Often the marker only partially hid the information, and the correction stick sometimes flaked and could be scratched off. We were also careful to check the flip side of the document before using the black marker. If there was writing on the opposite side, then only correction stick was used to avoid marker bleed-through.

At the beginning of the school year, fieldworkers were given guidelines for cleaning the data that they collected (Document 11 in Appendix). In addition, we individually trained fieldworkers upon their first data delivery to our office. As assignments and student work came in from the fieldworkers, we immediately checked to make sure the data were coded and cleaned correctly, occasionally requesting researchers to improve on what they had turned in. With the huge volume of data passing through our doors, it was critical that fieldworkers took personal responsibility for being careful and consistent in cleaning their documents.

Managing assignments and student work. Once assignments were cleaned and coded, our next step was to check them into the Assignments Collected Inventory Database (Document 12 in the Appendix). This database was very helpful over the course of the year for tracking many aspects of the collection process, including the progress of individual fieldworkers as well as that of particular teachers and schools. Often, fieldworkers double-checked with us for verification of which assignments they had turned in and which remained outstanding. This database allowed us to find out such information quickly. After entering the assignment, we initialed them to avoid duplicate data entry.

Following check-in, assignments were then electronically archived, which involved transferring the teacher instructions from the collection sheet into a Word document (Document 13 in the Appendix). We did not electronically archive the

accompanying worksheets, as this would have required an enormous investment of time for scanning. However, this might be a useful practice for future projects, in that it would create a more complete electronic archive of collected assignments. The electronic archive was organized according to subject, with nested subcategories of school, grade, type of assignment (challenging or typical), and collection number, allowing for easy retrieval.

Once checked in and electronically archived, assignments were ready to be filed. Because of the frequent need to access this data, our paper archive had to be kept in meticulous order. Math and writing/language arts assignments were filed separately, and within each subject, the challenging and typical were also separate. Within each of these groups, assignments were filed by school and then by collection number. Furthermore, since the typed assignment, rather than the teacher-written collection sheet, would be used during the scoring process, it proved useful to prepare the assignment for scoring as we went along. This meant that in addition to the collection sheets and accompanying attachments, we included in the paper file two copies of the typed assignment with photocopied attachments. When it came time to prepare for scoring, we just removed one copy of the assignment, leaving the original and an extra copy on file.

Student work was handled very much like assignments. Each piece of student work that was actually going to be used (that was not eliminated by random selection or by lack of parental consent) was entered into our Student Work Database (Document 14 in the Appendix). We did not keep electronic archives of student work; again, this would have involved a significant input of time for scanning, and our project did not require electronic access to and storage of student work. Since student work was only collected in conjunction with challenging assignments, we filed all student work with the accompanying challenging assignment. As with the assignments, we prepared student work for scoring as we went along, by attaching to each piece the corresponding, typed assignment and attachments. All extra pieces of student work were clipped to the back of the assignment and selected student work and kept in case a problem arose later with one of the assignments making it necessary to provide additional samples. When it came time for scoring, we simply pulled out the already prepared pieces of student work from the file.

Assigning Quantitative Scores to Data

Measurement rubrics were designed to score these materials according to the level of authentic intellectual work they demanded of students or the level of authentic intellectual work that students actually produced. These rubrics measured three standards each for both mathematics and writing (See rubrics that are Documents 15 and 16 in Appendix). Groups of teachers from non-Annenberg Chicago schools were recruited and trained to use these rubrics. The goal was that they would score with a high degree of reliability. These teachers scored both assignments and student work. Scoring of assignments and work was done in the summer following the academic year in which they were collected.

The Quantity of Data Scored

As we mentioned in the Introduction section, the aim of this work was to gain a deeper understanding of student "authentic intellectual work." Writing and mathematics experts, respectively, created the rubrics (Documents 15 and 16 in the Appendix) to measure key aspects of authentic intellectual work. Three standards are defined for each subject.

These differ slightly for student work in comparison to teacher assignments. However, in brief, the three standards are "construction of knowledge," "elaborated written communication," and "connection to student lives." For each standard, scores from 1 to 3 or 1 to 4 were defined. Therefore, for a given piece of data, three scores needed to be assigned—one for each standard.

Here we look at scoring of assignments done after our Year 5 collections, using grade 3 math as an example to illustrate the quantity of scoring to be done. We collected a total of 183 third grade math assignments in Year 5 that needed to be scored. Recall that these come from 14 schools and that we collected 6 samples (4 typical and 2 challenging) of student work from each third grade teacher of math in the study. In addition, along with the newly collected Year 5 data we also re-scored at the same time data that was collected in Years 1, 2, and 3. A random sample of 50 grade 3 assignments from each of these years brings the total to be scored to 333 (183 + 50 + 50+ 50 = 333). (See the next section "Statistical analysis of data" for more complete discussion of why we re-scored data and how these data were selected for re-scoring.) Document _17_ in the Appendix shows an estimated tally of the amount of assignments and work handled in Year 5.

Now, each of the 333 unique teacher assignments needed to be scored according to each of the three standards. In other words, 999 scores need to be assigned to fully quantify the level of authentic intellectual work shown in this sample.

As we will detail in the next section, several scorers were hired to score this grade 3 math data. Given multiple scorers, we recognized that there would be some differences in how they applied the standards—regardless of how carefully they were trained to minimize inconsistencies. Therefore in order to account for differences among scorers and later control in the analyses for "scorer severity," we had to have some of the data scored by two different scorers on the same standards. In order to calculate the inter-scorer differences half of the data documents needed to be double scored. That means in addition to the 999 scores already needed, another 499 scores needed to assigned for a total of 1,498.

In addition to math assignments, math scorers were responsible for scoring student work. Student work was only collected twice, once for each semester. About ten pieces of work were collected per class per collection. In the case of third grade math in 2001, we collected 445 pieces of work. Another 100 pieces were pulled from Year 1 and Year 2 (50 each) for re-scoring. Again each of these pieces of student work needed to be scored according to each of the three standards for student work in math, totaling 1,635

scorings. Half as many were then double scored to check inter-scorer reliability, for a total of 2,452 scores needed for third grade student work. Therefore, our third grade math scorers in 2001 were responsible for a grand total of 3,950 scores.

Scheduling and Staffing the Scoring

How long does it take to assign all these scores? How many scorers were needed? After years of working on the scoring, we were able to estimate the length of time scorers needed to read a piece of student work or assignment and arrive at a score. This allowed us to determine reasonable schedules and hire staff accordingly.

On average we estimated it took 1.3 minutes for a scorer to assign a score to an item. From this estimation and the estimation of work to be scored, for the post-Year 5 scoring we made a schedule based on a maximum scoring time of 85 minutes per standard for assignments and 150 minutes per standard for student work (see Document 18 in Appendix). The schedule included 90 minutes of training per standard with the first assignment and student work standards receiving 110 minutes. Refresher scoring time consisted of 20 minutes per standard. In each year of scoring, we planned a four-day schedule beginning each day at 8:30 A.M. and ending at 5:00 P.M. with morning and afternoon breaks of 15 minutes and a lunch break of 45 minutes. Depending on the flow of the work, some days ended earlier than others.

Given the volume of data to score and our choice to keep the scoring to 4 days, we typically needed between 6-9 scorers per grade per subject. We found it advantageous to hire an alternate scorer for each grade in each subject to guard against having a shortage of scorers due to absentees. Early in the project, the decision was made to hire Chicago public school teachers to do the scoring. These teacher-scorers would be peers of the teachers who contributed the data—they would be credible judges with perspectives close to those teachers who provided the data. The project saw these scorers as doing a job to contribute reliable scoring for analysis. They were seen as doing a critical job. However, the project recognized that for some of the scorers, they would gain some knew insights into the nature of assignments and student learning that could benefit them professionally.

A part-time staff person with strong connections with Chicago teachers was hired to recruit teacher-scorers. Hiring the needed scorers was not easy given that many teachers teach summer school in Chicago and many others have other professional or personal obligations. Over the years, we developed some guidelines that helped us evaluate the strength of prospective teacher scorers. Strong candidates needed to:

- Currently be teaching in one of the grades from which we were scoring work, or be familiar with teaching children at these grades
- Be able to carefully follow the scoring rubric and scores according to its criteria,
- Be open to the authentic pedagogy framework
- Willing to learn
- Understand that this was “work” and be willing commit to the schedule
- Not be a teacher in a school from which we collected assignments and student work

- Come from a school where other teachers have also agreed to participate in the scoring process
- Represent the diverse characteristics of teachers along racial and gender lines

In the later years of the research project, there were a number of scorers from previous years who returned to score again. In these cases, their previous work as scorers was evaluated so to assure that only reliable scorers were rehired. Despite our attention to these criteria for good scorers, our hires were not uniformly successful in their scoring. As we explain in the following, the scoring and then analysis process were then designed to minimize problems caused by an inconsistent scoring.

Organizing Data for Scoring: The Matrix

Once we knew how much data was to be scored and how many scorers would do the work, we needed to organize how the vast amount of data was to be distributed for scoring. We used a series of matrices to organize this distribution. A unique matrix was needed for each grade by subject and by type of work (assignment or student work). For example, we needed separate matrices for (a) third grade math assignment, (b) third grade math student work, (c) third grade writing assignments, and (d) third grade writing student work. Each matrix specifies who scores what document on what standard.

Central to our organization was that a number of pieces of data were bundled together to form a packet—usually 9-11 pieces per packet. These packets were the central physical artifacts that were used in the scoring process. How they were organized and passed from person to person was a complex and essential challenge to plan and implement.

To produce a matrix, we needed the following information: a final count of what data would be scored, the ID numbers for all data to be scored, a count of scorers hired for each grade in each subject, and the number of pieces of data to be placed into each packet. If these variables change, the configuration of the matrix would also change. Therefore, in the process of doing this research it was very important to have a carefully followed timeline to hire scorers, finalize the inventory of assignments/work, etc. One challenge we faced was that we sometimes did not have the number of scorers confirmed on schedule. In these cases, we had to create multiple matrices and wait later than we had hoped to finalize our scoring plan. See Document _19_ for the spring/summer schedule used to plan the preparations for the Year 5 scoring. It shows the complexity of planning required.

Built into our design of matrices are a number of key guidelines. For example:

- All assignments/work are scored at least one on each standard
- Half of the assignments/work are double scored
- Each possible pair of scorers works together at least once for each standard
- It is best if each packet is scored only once by a scorer. However, due to the number of scorers scoring, some scorers will most likely have to score a packet twice. If this is the case, try to put distance between the times that they see the same packet. For example if Scorer A sees a packet for Standard 1, Scorer A should not see that packet

for Standard 2. But if needed could possibly see it for re-scoring of Standard 2, or better for Standard 3.

Creating these matrices was a time consuming and tedious task. In the Appendix we include Document 20 “Year 5 scoring: The how to manual” for a thorough record of how to use Excel to make a matrix.

With a matrix created, packets were prepared for scoring. To do this, we looked at the matrix to see which assignments/work belonged in what packets and which scorers were to score which packets. Then it was a matter of pulling these items out of the file folders where they were stored. Assignments/work were then photocopied (if they had not already been so) and double-checked to see that they were readable and that no identifiers remained. The assignments/work were placed inside a manila folder in the order dictated by the matrix.

Using an Excel to Word merge, we produced score sheets for each packet for each of the standards (again see the appendix for greater detail in how to do this in Document 20 and for a sample score sheet Document 21).

Training Scorers for Reliability

The quality of the research rests on the reliability of the scoring. We worked hard to hire teacher-scorers who were capable of and committed to doing the work of scoring and, then, worked to make sure their training was thorough and ongoing throughout the scoring process. The detailed process of training the scorers was also practiced or “piloted” in the months prior to the official scoring sessions. The consistency of the training of scorers was strengthened by these pilot sessions and also by the projects attempt to use the same training leadership across the five years of the study.

The training structure embedded within the scoring sessions was the same for both math and writing. For each scoring session, one trainer was assigned to each of the three grades (3, 6, and 8). Each trainer was responsible for training his/her team of teacher-scorers on all three standards. Using the “Manual for scoring tasks and student work” in the appropriate subject (see Documents 15 and 16 in the Appendix), trainers introduced the teacher-scorers to the standards and scoring criteria.

Using these manuals, the trainers first reviewed the general rules for scoring by having participants read them aloud and put them into their own words. This was followed by a discussion of any questions participants may have had. Next the specific scoring rules for the standard were read out loud and similarly discussed, with emphasis placed on identifying and understanding the key words that distinguished adjacent scores (e.g. a score of ‘2’ versus ‘3’) from one another. Then the trainer asked teachers to score one or two “training papers.” Training papers were sample assignments or work selected by trainers. Scorers were asked what scores they assigned and then the trainer lead a discussion until consensus on the correct scores was reached. During these discussions, the trainer emphasized that everyone needed to be convinced that there was one correct

score. Scorers needed to understand that scores were to be assigned according to the specific rules of the manual. The rationale for assigning a score needed to follow the precise language in the manual and refer to specific parts of the assignment/work scored. All scorers needed to be present and active in the discussion.

During the training, each trainer conducted thorough discussions of at least four training papers. During the scoring period, each teacher scored a packet of about ten assignments/work. Each packet was double scored independently by another teacher. During the refresher period, the trainer asked teachers if they had any questions or concerns that needed clarification. After discussing these issues, the trainer proceeded to have teachers score the refresher papers and discuss the scores as in the earlier training. Before returning to the scoring, the trainer or a group member summarized the key points that emerged from the refresher period.

Statistical Analysis of Data

Once the scoring of the student work and teacher assignments had been completed, we needed to develop measures that would be useful for analysis and the reporting of results. First, we wanted to produce a measure of authentic intellectual work for each assignment and piece of work based on the scores on the three standards. After the measures had been produced, we needed to develop a means of reporting the results in a manner that would be easily understood and interpreted. For the purpose of analysis, we also needed to develop measures that were aggregated to the student level for student work and to the classroom level for teacher assignments. Another major analytical piece came after we had collected multiple years of data, as we had to figure out a means of equating the scores from different years. The equated scores on teacher assignments and student work were then analyzed to generate longitudinal trends.

The different stages of analysis for the project involved the use of multiple software packages. This was further complicated by being in multiple operating systems, Windows, Dos Linux, and Unix, requiring transferring of data to different locations and being put into various formats. The development of the matrices and other databases was done in Excel and Word, the assignment and student work measurement development was conducted in Facets, a DOS program, while the primary data analysis was done in SAS for Unix and the measures at the student and teacher-classroom level were created in HLM (hierarchical linear models version 5.0) for Unix. For the most part, the analysis and measure development of student work and teacher assignments paralleled each other. Where differences are important they are noted and then explained in detail.

Database Creation

After the teacher assignments and student work had been scored, the scores were recorded in matrices similar to those described in the previous section. These matrices formed the foundation for the analysis work and the creation of the assignment and student work measures. As before, a separate matrix was created for teacher assignments

and student work, by grade and subject, for a total of twelve separate files for each year. Each piece of student work (or assignment) had a separate line in the database. The matrix recorded the two scores for each of the three standards for which a teacher-scorer had given a score (since not all standards were scored twice for a given piece, some cells had missing data). Besides the scores and scorer names, the matrix included background information such as the assignment ids, student ids (for student work), school, whether a assignment was typical or challenging (for assignments), teacher ids, and room number.

After the data had been recorded in the Excel matrices, they were transferred to our Unix server as space delimited text files and then read into SAS data format. Almost all statistical analysis at the Consortium is done in SAS, and all of our data is stored in SAS format, making it necessary to convert the student work and assignment scores to this format to allow for future analysis using data from the Consortium's archive. Data that had been collected on the teachers and students that had been stored in other Excel databases were also read into SAS data format and merged with the assignment and student work data sets. The project had been granted permission to use the CPS student ids, room numbers and unit numbers as identifiers for the assignments and student work. This allowed the project to link the data collected on assignments and student work to the Consortium's archive of data on students, classrooms, and schools, including student test scores, demographics, and student survey responses as well as teacher survey responses aggregated to the school or grade-level. Being able to link the data from this project to existing data greatly expanded our analytic possibilities.

Measure Development

One of the primary concerns in developing measures for the teacher assignments and student work was the likelihood of inconsistencies in the scoring of the standards among the teacher-scorers and variation in the difficulty of the three standards despite training on scoring the standards. We were also concerned about differences in the difficulty of going from one category on a standard to the next category, that is, going from category one to category two may be more difficult than going from category two to category three. In order to statistically adjust for these inconsistencies, the student work and teacher assignment measures were developed using Many-Facet Rasch Analysis with statistical software specifically designed for this model, FACETS.

Besides adjusting for differences in rater severity, standards, and categories within standards, another advantage of the Rasch model is that it allows for missing data. While more data allow for better estimates, the Rasch model is able to make precise estimates of rater severity, standard difficulty, and extent of intellectual challenge without requiring all teacher-raters to score each assignment and each piece of student work. The FACETS program also allows for detecting bias among the teacher-scorers use of scores and standards. Another advantage of using the Rasch model and the FACETS program is that standard errors for each measure of the assignments (or student work), teacher-scorer, and standard are produced in the output.

In the Rasch analysis, we constructed overall measures for the intellectual quality of all teacher assignments and student work. We made separate scales for each grade and subject, as the standards, by design, were applied differently by grade and subject. This analysis statistically adjusted the recorded ratings for differences in the severity of the teacher-raters and the difficulty associated with achieving each rating on each standard. Each piece of student work or teacher assignment has information on three different facets: the scores given by the teacher-raters on the three standards, the standards, and the teacher-raters. From the different pieces of information, a quantitative measure is estimated for each element of each facet, that is, each teacher-rater has a measure, each standard has a measure, and each piece of student work or assignment has a separate measure. All of the measures are placed on one common interval scale so that each element can be compared to the others. The Many-Facet Rasch model used here was as follows:

$$\log(P_{nijk} / P_{nijk-1}) = B_n - D_i - C_j - F_k$$

where

P_{nijk} is the probability of assignment n being given a rating of k on standard i by judge j;
P_{nijk-1} is the probability of assignment n being given a rating of k-1 on standard i by judge j;
B_n is the intellectual quality of assignment (or piece of student work) n;
D_i is the difficulty of standard I;
C_j is the severity of judge j; and
F_k is the difficulty of receiving a rating of k compared to a rating of k-1.

Thus, the final measure of the assignment quality, B_n, aggregates the score across all three standards, while adjusting for the difficulty of each standard and the relative severity of the scorers.

In order to do the FACETS analysis, control files had to be constructed using the scores that had been inputted into SAS. The FACETS program control files used text files with a series of commands followed by assignment (or student work) ids and then the data, arranged so that each score had a separate line. For example, if an assignment had been scored twice on each of the three standards, it would have six different lines in the data section. See Document 22 in the Appendix for a sample FACETS control file. We used a model where all three facets were positive and non-centered, with a separate rating scale created for each standard, but not for the other two facets. When generating ids for the student work analysis, the assignment id number was appended to the student id so that the item could be linked to its assignment in future analysis. After the files were constructed, they were transferred from our UNIX machine to a desktop computer where the FACETS analysis was run. We constructed the files using a SAS program that generated the control files. Alternately the files could have been made from the Excel database of score using a Windows program such as Word or Word Pad, and saved as text files.

The FACETS program produces a large number of tables in its output files. In our study, we focused on the tables produced that provided information on the scales for the three standards, their step difficulties, the severity of the teacher-raters, and the assignments or student work. During measure development, we checked the reliability, bias statistics, and fit statistics to make sure that the measures were working well and to detect aberrant scores, or raters and to detect bias in the interactions between raters and standards.

The product of this stage of analysis came from the item measures (teacher assignment and student work) that were read from table 7.3.1 in the FACETS output. To read the data into SAS, the table was edited to only include the output from this table and transferred to our UNIX machine. From the table, the values for the item measure, the standard error and the infit (information weighted mean-square fit statistic with an expectation of 1) and outfit (unweighted mean-square, sensitive to outliers) statistics as well as the assignment (or student work) ids were read into SAS. As in other item response theory applications, the final measure of assignment authentic intellectual challenge (or student work) exists in a logit metric and has a standard error attached to each measure. For most analysis, we used an inflated standard error that had been adjusted by the square root of the maximum of three values, the infit, the outfit statistic or 1. For reporting purposes, we converted the scores to a 0 to 10 point scale by using the following formula:

Assignment (or student work) measure = 10*(logit measure – minimum)/range
where
logit measure is the score in the logit metric produced by the FACETS output;
minimum is the minimum score;
range is the maximum score – minimum score.

Then standard errors were also put on the same scale by multiplying the standards errors by 10/range.

Equating Scores Across Years

Initial measure development, while a difficult and complex learning process, used data from only one session of scoring which was conducted under the same training conditions and with the same teacher-raters. When we examined two years of scored data, it became clear that while the standards and training were intended to be consistent, teacher-raters had significant differences in how they assigned scores in the two years. After extensive analysis of the data from the two years, we determined that equating the scores from the two years would be necessary if scores from different years were to be compared.

We first suspected that we had a problem with bias when the assignment scores increased dramatically between Year 2 and Year 1. To determine whether we had a problem with bias, that is teachers were assigning scores or using the standards differently in the two years, we had two experts (Jolliffe in writing and Gutstein in math)

who had trained the teachers in the use of the standards re-score a subset of assignments for one grade in each subject. We then we compared the two sets of scoring and we found that the Year 2 scores on all of the standards were higher than the scores given by the two experts while the Year 1 scores were about the same or slightly lower. Although we were not able to determine the exact cause of the drift in the use of the standards, we were realized that comparing the assignments and student work collected and scored in Year 1 against the Year 2 assignments would require equating the two scoring sessions.

We put off developing measures for the teacher assignments until the following year when a sub-sample of items from the first two years could be re-scored and then equated so that data from all three years could be placed on the same scale. We selected thirty assignments from each year-grade-subject matter combination, taking care to include assignments with a wide range of scores. Equating student work scores was not done until the next data collection and scoring of student work in Year 5.

Having re-scored assignments allowed us to equate our measures across years, adjusting for possible differences over time in the relative severity of how scorers applied the standards. Specifically, using a Linux statistical package called R (equivalent to S in Windows or Unix), we calculated a Tukey's bi-weighted mean for the difference between the original scores and re-scores for each year, subject and grade. These "adjustment effects" were then added to the original scores recorded for the Year 1 and Year 2 assignments to place them on the same scale as the Year 3 assignments.

In Year 5, we again re-scored a subset of assignments and, for the first time, re-scored student work from previous years in order to be able to equate the Year 5 scores with those from other years. In this round of scoring, we re-scored the assignments that had been re-scored in Year 3 and an additional 20 assignments per grade-subject combination to improve the reliability of our re-scores. We re-scored 50 pieces of student work per grade-subject per each of the previous years. Because we had these additional scores, we were able to make additional comparisons to check the reliability of our re-scoring process and to make more reliable adjustment effects. Again, we calculated a Tukey's bi-weighted mean for the difference between the various scorings (1999 – 1997, 1999 – 1998, 2001 – 1997, 2001 – 1998, 2001 – 1999 for assignments and 2001 – 1997 and 2001 – 1998 for student work) for each subject-matter grade combination.[11]

We then had two sets of comparisons which could be used to develop our adjustment effects, the comparison between the original score and the re-score in 2001 and the comparison of the original score and the re-score in 1999, adjusted for the difference between the 1999 and 2001 scoring of the same assignment. We decided to use both sets of comparisons and weigh them by their standard errors in order to develop an adjustment effect that would place all assignments and student work on the 2001 scale. The formula used to calculate the new adjustment coefficient for 1997 and 1998 assignments is:

[11] Throughout this manual we refer to the 1996-97 school year as Year 1, 1997-98 as Year 2, 1998-99 as Year 3, 1999-2000 as Year 4, and 2000-01 as Year 5. When we discuss statistical analyses, 1997 is Year 1; 1998 is Year 2; 1999 is Year 3; 2000 is Year 4; and 2001 is Year 5.

Adjustment =$[(\lambda_1 * \text{Estimate 1}) + (\lambda_2 * \text{Estimate 2})]/ (\lambda_1 + \lambda_2)$

where (example shown for 1997 (Year 1) scores)
2001= biweighted mean of 2001 (Year 5) scoring of re-scored items;
1999= biweighted mean of 1999 (Year 3) scoring of re-scored items;
1997= biweighted mean of 1997 (Year 1) scoring of re-scored items;
Estimate 1 = (2001-1999) + (1999-1997);
Estimate 2 = 2001–1997;
$SE^2_{est1} = (SE^2_{2001-1999} + SE^2_{1999-1997})$;
$SE^2_{est2} = (SE^2_{1999 - 1997})$;
$\lambda_1 = (SE^2_{est1} / \sum SE^2_i)^{-1}$;
$\lambda_2 = (SE^2_{est2} / \sum SE^2_i)^{-1}$.

The adjustment for 1999 assignments and 1997 and 1998 student work simply uses the difference between the 2001 score and the original score, estimate 2 above.

In the 2001 analysis, we also explored another means of equating the scores. Initially we attempted to run the FACETS analysis with the assignments scores and standards anchored on their 1999 values to put all assignments on the 1999 scale. However, we found that the standard errors on the assignment scores produced by this methodology to be unacceptably large relative to the scale of the scores, particularly for scores near the minimum of the scale. We also found that anchoring the assignments on their 1999 logit scores was causing some scores on standards to have large z-scores. (The z-scores are bias estimates that have been standardized to have a mean of zero and standard deviation of 1.)

Creating Categories of Scores

To provide a more substantive standards-based interpretation for these data, we also divided the distribution of adjusted scores into four categories: extensive, moderate, minimal, and none to describe the extent of challenge of the assignments and the extent of authentic intellectual work. As with all of our analysis, separate cut-points to define the categories were developed for each grade and subject matter combination. Cut-points are based on the relative location of the three standards on the logit scale on which the teacher-scorers and assignments (or student work) are placed.

The process of developing cut-points is largely subjective, but we use several guidelines. We usually use transition points on the different standards as cut-points, although if two standards have nearly equal transition points, we may use the midpoint between the two points. The categories created by the cut-points are selected to be as internally consistent as possible so that the categories have a clear meaning. For example, a piece of student work belonging to the lowest category, no challenge, would be expected to have received a score of one on each of the standards from all teacher-raters.

The primary table used in making cut-points is table 6.0 from the FACETS output that graphs the scale with the different relative locations of the teacher-scorers, items (assignments or student work), and the standards. See Document 23 in the Appendix for a sample table 6.0. Each standard is also mapped using the scale with the location of the categories from the scoring rubrics, but their location is not based on the scale used in the main figure. The standard scales must be re-centered using the measure values from tables 8.1-8.3 (Category Statistics). See Document 24 for sample tables. The measure values are indicated by a "*" on table 6.0 and each standard map must be realigned using its measure value from the table (with the sign reversed) on the overall scale before cut-points can be made.

Aggregating Teacher Assignments and Student Work

The main objective of our study was to relate the intellectual demands in classroom instruction, as manifest in teachers' assignments and student work. For teacher assignments this meant aggregating the scores to the teacher level. For student assignments this meant aggregating the measures to the student level. For this purpose, we needed to combine the multiple individual assignment measures collected from each teacher (or student) over the course of the project into an overall measure of intellectual demand. While the idea here is straightforward, the actual process for doing this was complicated by two factors. First, a varying number of teachers and students participated in this study over the course of project (three years for student work and four for assignments). Second, although we had a fixed data collection design, teachers actually provided us with a varying number of challenging and typical assignments each year. In some classrooms in some years, we collected only 2 or 3 assignments instead of the original design of 6 assignments. For the student work, we only scored about ten items per classroom so that for most students we had only one assignment, although for others we had two. Both of these factors introduced considerable noise in the data and some potential for bias as well. Most of the HLM analysis was initially developed using assignments, as we did not have equated student work scores until 2001.

In order to control for these extraneous sources of variability in the assignments, we developed a formal model to measure intellectual demand. Our original intent was to measure this separately for each classroom for each year. However, we found in a preliminary HLM analysis, where we nested classrooms within teachers, that we could not reliably estimate a variance component for this latter factor. Thus, we eventually chose to aggregate, separately for math and writing, all of the assignments obtained from each teacher, across the one, two, three or four years that the teacher might have participated in the study. In essence, we developed two teacher-level measures for the intellectual demands of assignments, one in writing and a second one for math. In subsequent analyses where assignment quality is used as a predictor variable, these teacher-level measures are linked to all of the classes that each teacher taught.

Formally, a three-level HLM was used to develop our measure of classroom intellectual quality. The outcome variable at level 1 consisted of the individual assignment measures generated from the Rasch analysis. Two level-1 dummy variables

were used as predictors to distinguish between the math and writing assignments. All of the variables in the level-1 model were weighted by the inverse of the standard errors of measurement associated with each individual assignment measure. (These were produced by the Rasch analysis.) The major function of this level-1 measurement model is to take into account the unreliability of the individual assignment scores. Formally, the two coefficients produced here, π_{1jk} and π_{2jk}, can be thought of as latent "true scores" for assignments j from teacher k.

Level 1:

$Y_{ijk} = \pi_{1jk}$ (Math) + π_{2jk} (Writing) + ε_{ijk}

where Y_{ijk} k = assignment quality score, and ε_{ijk} is now assumed N(0,1) given the reweighting by the standard errors of measurement.

At level 2, both of these coefficients become outcome variables where we have multiple assignments per teacher. A level-2 dummy variable, which distinguishes challenging from typical assignments, was grand-mean centered and its effect fixed.

Level 2:

$\pi_{1jk} = \beta_{10k} + \beta_{20k}$ (Challenging) + r_{10k}
$\pi_{2jk} = \beta_{20k} + \beta_{21k}$ (Challenging) + r_{20k}

As a result, the intercept terms, β_{10k} and β_{20k} are overall measures of teachers' assignment quality in mathematics and writing respectively, adjusted for differences among teachers in the number and types of assignments they submitted.

Finally, at level 3 (i.e. the teacher level), indicator variables for grade 6 and grade 8 were included in order to adjust for grade-specific effects in the assignment rating system.

Level 3:

$\beta_{10k} = \gamma_{100} + \gamma_{101}$ (Grade 6) + γ_{102} (Grade 8) + u_{1k}
$\beta_{11\kappa} = \gamma_{110}$
$\beta_{20k} = \gamma_{200} + \gamma_{201}$ (Grade 6) + γ_{202} (Grade 8) + u_{2k}
$\beta_{21\kappa} = \gamma_{210}$

After running the HLM, the residual files were read into SAS and the empirical Bayes estimates from this model for β_{10k} and β_{20k} were used in subsequent analyses as an overall measure of the assignment quality that students were exposed to in mathematics and writing respectively.

The model used for aggregating the student work measures to the student level was almost identical to the one used for aggregating the assignments. Because it was not

necessary to control for type of assignments since all student work items were based on challenging assignments, unlike the assignment model, level 1 and level 2 were based on different groupings. In the student work model, level 1 was student work, level 2 was students, and level 3 was the teacher-classroom.

Level 1:

$Y_{ijk} = \pi_{1jk}$ (Math) + π_{2jk} (Writing) + ε_{ijk}

where Y_{ijk} = student work score, and ε_{ijk} is now assumed N(0,1) given the reweighting by the standard errors of measurement.

At level 2, both of these coefficients become outcome variables where we have multiple assignments per student. Unlike the assignment trend model, no predictors are used at level 2.

Level 2:

$\pi_{1jk} = \beta_{10k} + r_{10k}$
$\pi_{2jk} = \beta_{20k} + r_{20k}$

The intercept terms, β_{10k} and β_{20k} are overall measures of students level of intellectual work in mathematics and writing respectively, adjusted for differences among students in the number and pieces of work that were included in the analysis.

As with the assignments, at level 3 (i.e. the teacher-classroom level), indicator variables for grade 6 and grade 8 were included in order to adjust for grade-specific effects in the use of the standards in scoring the student work.

Level 3:

$\beta_{10k} = \gamma_{100} + \gamma_{101}$ (Grade 6) + γ_{102} (Grade 8) + u_{1k}
$\beta_{20k} = \gamma_{200} + \gamma_{201}$ (Grade 6) + γ_{202} (Grade 8) + u_{2k}

Score Trends

Another stage of the project was examining the trends in student work and teacher assignments over time. We estimated the mean trends in the intellectual demands of assignments and student work in writing and math using separate analysis with hierarchical linear models (HLM). The actual analytic model used was as follows. Level 1 was a measurement model with an intercept and for assignments, three effects coded dummy variables for the years 1997, 1999, and 2001 with 1998 as the excluded category. The outcome variable consisted of the assignment measures generated from many-facet Rasch measurement. All of the elements in the level-1 model were weighted by the

inverse of the standard error of the assignment measures. (These are produced as a by-product of the Rasch analysis.) The major function of this level 1 measurement model is to take into account the unreliability of the assignment scores. Formally, the four coefficients produced here, π_{1jk} , π_{2jk} , π_{3jk} , and π_{4jk}, can be thought of as latent "true scores" for the assignments, with the intercept term, π_{1jk}, representing the overall adjusted mean for all three years. Each of these becomes an outcome variable in the level 2 model where we have multiple assignments per classroom.

Level 1:

$$Y_{ijk} = \pi_{1jk} + \pi_{2jk}\,(1997) + \pi_{3jk}\,(1999) + \pi_{4jk}\,(2001) + \varepsilon_{ijk}$$

where Y_{ijk} = assignment quality score for mathematics or writing, and
ε_{ijk} is now assumed N(0,1) given the re-weighting by the standard errors of measurement.

Level 2:

$$\pi_{1jk} = \beta_{10k}\,(\text{Typical}) + \beta_{11k}\,(\text{Challenging}) + r_{10k}$$
$$\pi_{2jk} = \beta_{20k}\,(\text{Typical}) + \beta_{21k}\,(\text{Challenging}) + r_{20k}$$
$$\pi_{3jk} = \beta_{30k}(\text{Typical}) + \beta_{31k}\,(\text{Challenging}) + r_{30k}$$
$$\pi_{4jk} = \beta_{40k}\,(\text{Typical}) + \beta_{41k}\,(\text{Challenging}) + r_{40k}$$

Level 3:

$$\beta_{10k} = \gamma_{100} + \gamma_{101}\,(\text{Grade 6}) + \gamma_{102}\,(\text{Grade 8}) + u_{10k}$$
$$\beta_{11\kappa} = \gamma_{110} + \gamma_{111}\,(\text{Grade 6}) + \gamma_{112}\,(\text{Grade 8})$$
$$\beta_{20k} = \gamma_{200} + \gamma_{201}\,(\text{Grade 6}) + \gamma_{202}\,(\text{Grade 8})$$
$$\beta_{21\kappa} = \gamma_{210} + \gamma_{211}\,(\text{Grade 6}) + \gamma_{212}\,(\text{Grade 8})$$
$$\beta_{30k} = \gamma_{300} + \gamma_{301}\,(\text{Grade 6}) + \gamma_{302}\,(\text{Grade 8})$$
$$\beta_{31\kappa} = \gamma_{310} + \gamma_{311}\,(\text{Grade 6}) + \gamma_{312}\,(\text{Grade 8})$$
$$\beta_{40k} = \gamma_{400} + \gamma_{401}\,(\text{Grade 6}) + \gamma_{402}\,(\text{Grade 8})$$
$$\beta_{41k} = \gamma_{410} + \gamma_{411}\,(\text{Grade 6}) + \gamma_{412}\,(\text{Grade 8})$$

At level 2, dummy variables were entered for challenging and typical assignments. As a result, the intercept terms β_{10k} is the overall classroom mean score for typical assignments and β_{11k} is the overall classroom mean for challenging assignments, adjusted for differences among teachers in the number and types of assignments they submitted. γ_{200} is the adjusted year effects for 1997, γ_{300} is the adjusted year effects for 1999, and γ_{400} is the adjusted year effects for 2001. Finally, at level 3 (i.e. the classroom level), effects-coded indicator variables for grade 6 and grade 8 were included in order to estimate the grade-specific effects.

The model used for student work was similar to that used for assignments except that level 1 was pieces of student work, level 2 was students, and level 3 was classrooms and no predictors at level 2 were included since only student work for challenging assignments were collected. Also because student work was only collected in 1997, 1998, and 2001, only dummy variables for 1997 and 2001 were used, with 1998 as the excluded category.

Level 1:

$Y_{ijk} = \pi_{1jk} + \pi_{2jk}\,(1997) + \pi_{3k}\,(2001) + \varepsilon_{ijk}$

where Y_{ijk} = student work score on authentic intellectual work for mathematics or writing, and
ε_{ijk} is now assumed N(0,1) given the re-weighting by the standard errors of measurement.

Level 2:

$\pi_{1jk} = \beta_{10k} + r_{10k}$
$\pi_{2jk} = \beta_{20k} + r_{20k}$
$\pi_{3jk} = \beta_{30k} + r_{30k}$

Level 3:

$\beta_{10k} = \gamma_{100} + \gamma_{101}\,(\text{Grade 6}) + \gamma_{102}\,(\text{Grade 8}) + u_{10k}$
$\beta_{20k} = \gamma_{200} + \gamma_{201}\,(\text{Grade 6}) + \gamma_{202}\,(\text{Grade 8})$
$\beta_{30k} = \gamma_{300} + \gamma_{301}\,(\text{Grade 6}) + \gamma_{302}\,(\text{Grade 8})$

To calculate the trends, we used the coefficients read from the output of the HLM analysis. The dummy variables for grade and assignment type were effects coded so that the intercept, γ_{100}, is the grand mean for all assignments (student work) for all years in the study for all grades for typical assignments and γ_{110} is the grand mean for challenging assignments. Means for specific grade and year combinations can be calculated by adding the coefficients to the grand mean for that type of assignment (or for student work). So that, for example, the mean for typical assignments for grade 6 in 1999 would be:

$\gamma_{100} + \gamma_{101} + \gamma_{200} + \gamma_{201}$.

Project Management

This project utilized large amounts of resources. Central to the success of the project was the hiring and management of many staff members with diverse types of expertise. We briefly describe the staffing of the project and a rough outline of how funding was allocated for this project.

Staffing and Hiring

Under the direction of CCSR co-director Mark Smylie, CARP was conducted collaboratively by dozens of researchers from more than eight universities. Just within the assignments and student work research strand, there were several types of staff roles: field researchers collecting data, core project staff managing data and project administration, a team scoring data, statistical data analysts, and a lead team for research design. Below we describe only those aspects of their work related to this specific strand of the larger Project.

The field researchers were responsible for collecting assignment and student work as well as other data needed to document the development over time of a school. For each school, there was one lead researcher and one research assistant. In almost all cases, the research assistant was the primary collector of assignments and student work. Most of the research-assistants were graduate students at local universities: University of Chicago, UIC, Loyola, and Roosevelt. Recruitment was through word-of-mouth and by posted notices to relevant departments at the above universities. See Document 25 in the Appendix for an advertisement describing the research assistantships. Turnover of this group of staff was such that just one RA from Year 1 was still involved with CARP by Year 5. Retaining RAs was difficult given the nature of graduate student life changes and our inability to employ RA s continuously given our research design that included no data collection—and thus, no jobs—in Year 4.

The core workers on the project shared a central office suite, facilitating the work of other groups on the project. Their core work responsibilities related to the assignment and student work strand of the Project included support of research assistants, data management, administrative support, hiring, communication with the many individuals on different aspects of the project, etc. At the beginning of Year 3 there was significant turnover in this staff. Associate Director BetsAnn Smith, one of the founding leaders of the Project, took a faculty position at Michigan State University. Fieldwork Manager Karen DeMoss relocated to another office within the Consortium. Data Manager Gudelia Lopez joined the Chicago Public Schools as a researcher. New to the Project were Stacy Wenzel, Director of Fieldwork, and Tamara Perry, Fieldwork Manager and Qualitative Analyst. Loretta Morris took on a new challenge as the Project Fieldwork and Data Coordinator. Nicolas Leon continued to work as Research Assistant in addition to providing general administrative support. Other key part-time staff continued including Verity Elston and Sabrina Billings. Verity assisted the Director of Fieldwork and participated in the report analysis and preparation. Sabrina provided general administrative support and aided in data check-in, entry, management, and retrieval for

analysis. By the start of Year 5, further turnover found Tamara, Nicolas and Verity gone and the addition of new part-time key staff Carol Fendt and Loreen Miller.

Fred Newmann (U Wisconsin) led the research group that scored the assignments and the student work. Two professors (David Jolliffe and Eric Gutstein, both of DePaul University) worked with Fred and with teams of other subject matter experts to develop scoring rubrics. The trainers working with the teams were, in math, Judy Merlau (University of Illinois) and Jean Biddulph (DePaul University), and, in writing, Carmen Manning and Kendra Sisserson (both of University of Chicago), and Annie Knepler (University of Illinois at Chicago).

University of Chicago staff Jenny Nagaoka and Gudelia Lopez with the direction of Tony Bryk and advice of the Consortium's Data Group performed statistical analysis for this aspect of the project.

The dozen or so senior research designers on the lead team for this project met as a group since its inception in order to plan the process and the written products that came-out of the project. Several of these researchers were professors at research universities–several from locales across the country. Several were recognized experts and well published in educational issues related to our project. The design group members also worked on the field research, data management, data scoring or analytic aspects of the work. This group included professors (Mark Smylie (UIC), Tony Bryk (U Chicago), Fred Newmann (U Wisconsin), BetsAnn Smith (MSU), Valerie Lee (U Michigan), Julie Smith (U Michigan)); Consortium post doctoral research professionals (Tamara Perry, Stacy Wenzel); Consortium professional staff (Tania Gutierrez, Jenny Nagaoka); and graduate students (Becky Greenberg (UIC), Rodney Harris (UIC), Karin Sconzert (U Chicago), Sara Hallman (U Chicago), Carol Fendt (UIC)).

Budget

The Chicago Annenberg Research Project received from CAC an initial grant of $3 million and an additional grant of $360,000. These funds covered not only the assignment and student work strand of research we highlight in this manual but also research based on citywide surveys of teachers, students and principals; longitudinal field work on school development; interviews with CAC external partners; and written records and fieldwork documenting the Challenge as an organization. Estimation of the cost of the assignment and student work of CARP is difficult based on the multiple roles played by staff across various research strands. However the following estimates offer some perspective on our expenses in conducting this work.

Rough estimate of salaries and consulting fees only, *in Dollars*

	Yr 1	Yr 2	Yr 3	Yr 4	Yr 5*	Total
Core staff whose majority responsibility was the assignment and student work strand	*55,000*	*55,000*	*105,000*	*85,000*	*65,000*	*365 K*
Research assistants collecting assignments and student work	*40,000*	*80,000*	*130,000*	*0*	*60,000*	*310 K*
Math and language arts teams designing rubrics and training teachers	*0*	*19,000*	*27,000*	*0*	*22,000*	*343 K*
Teachers hired to score assignments and student work	*500*	*25,000*	*32,000*	*0*	*27,000*	
Research design team member leading assignment and student work strand	*36,000*	*32,000*	*33,000*	*34,000*	*56,000*	
Statistical analysis budget --total	*10,000*	*20,000*	*38,000*	*47,000*	*115,00*	*230 K*

* Year 5 includes some funds allocated and spent in Year 6 to complete analysis of data collected in Year 5.

The cost of this work might be estimated in terms of the total of the figures above or about $1.2 million. Roughly one third of this cost was in the salary of core project members and another third in costs for scoring the data. Another quarter was needed to fund the collection of data done by project research assistants. In addition, around 20% of these funds were for statistical analysis. In perspective, it is fair to say that the physical logical handling of data documents at the schools and in the project archives was the most resource-intensive part of this work.

Lessons Learned

After five years of experience working with teacher assignments and student work, we have learned a number of lessons that may be helpful to others considering undertaking a similar project. Key among these lessons are issues of (1) assigning scores to data in a manner that allows comparison across years, (2) minimizing error in the process of electronic data management, (3) the importance of staff recruitment and retention, and (4) consideration of the scope of work.

In this longitudinal project, being able to compare data from different years was critical. We learned the hard way that despite our attempts to be consistent in training across years, the teacher-scorers tended to drift apart and develop their own definition of the standards that varied across the years. Our system of double scoring, re-scoring and careful statistical controls allowed us to make sense of the data despite this drift. However, future projects may be better able to carry out their scoring so to avoid some of the drift. One option would be to score the data for all years of the project in one session so that the highest level of consistency could be assured. However, we recognize as was our case, that results from a longitudinal project are desired at intermediate points as soon after data collection as possible. This makes conducting different scoring sessions and

then equating scores necessary. Given multiple scoring sessions, it is critical (a) that the same staff run these multiple scoring sessions, (b) that the wording of the standard remain exactly the same and (c) that a set of assignments and student work be used as benchmarks for defining scores on each standard. In our project, we met criteria (a) but not criteria (b) and (c).

Handling such a vast amount of data both physically and electronically offers many opportunities for data loss and error. Great care in the initial planning for this data management and manipulation is needed. For this project we used various versions of software that we were familiar with and that we knew would perform the functions that we needed. The downside of this was that it required transferring the data between operating systems and formats, adding the possibility of error and requiring additional time to do the transfer and to assure that the data were correct. Very careful manual data entry was used in this project. Our system required stringent double checking and a carefully coordinated data management staff. Future projects may be able to better plan their data bases and analysis software for greater compatibility. Further use of new technology may also improve the data management process. For example, field researchers could use digital cameras to record the image of student work and hand held computers to enter teacher assignment or other data. This technology could cut down data entry time and lessen errors.

Having the right staff and retaining them throughout a longitudinal project is critical to its success. This holds true for all aspects of the project. Finding the proper expertise for the various roles on the project requires considerable work. With our project, being situated in an organization with access to strong research faculty and staff and graduate students from several universities was greatly advantageous. Despite staff turnover, there was a rich pool from which to find new hires. Yet despite a good pool of hires, retaining staff was important. For example, collecting assignments and student work from schools proceeds best if a researcher has developed rapport with teachers and has learned how to best foster their full participation. In our project, we saw long term relationships with schools pay off with a higher yield of data and we also saw staff turnover cause us to lose some school data. Our project also benefited from having a data manager, lead research designers, and several staff who remained through most the life of the project.

In designing this project, there were many decisions made where the ideal situation had to be weighed with the logistics of carrying out a project at such a large scale. As with any research using statistical analysis, having more data and more information is critical in developing reliable measures. This is always counterbalanced with the limitations of time and resources in collecting data. The critical decision points in our study were many but included some of the following:

- how many pieces of student work or assignments to collect from each teacher
- how many teachers and schools to collect data from
- how many standards to use to define authentic intellectual work
- how many categories in each standard
- how many scorers to use

- how many times to have each teacher assignment or student work scored
- how many items to have re-scored.

We may have altered the course of this project by making different decisions on any of these points. An experienced leadership team devised its priorities and made decisions accordingly on our project. Future projects will need to likewise determine their goals (i.e., research questions; level of analysis whether system-, school-, teacher- or student-level; key theoretical frameworks) and then balance them with resources (i.e., numbers of teachers available; how much time individual teachers willing/able to spend; a few expert scorers or many well-trained lay scorers).

Table 1:

Field Research Schools are Comparable to All Annenberg Elementary Schools and to Elementary Schools Citywide
1998-1999

	Field Research Schools
Average student enrollment	**Range 600-1,600**
Racial breakdown of students	
African-American	50%
Latino	41%
White	7%
Asian/Pacific Islander	<1%
Native American	
Students with free or reduced-price lunch	**89%**
Students enrolled in bilingual education	**21%**
Of the eighth-grade students of 1993	
Graduated from a CPS high school	39%
Dropped out	37%
Left CPS	22%
1999 ITBS—students at or above national norms—in grades three through eight	
Reading	32%
Math	37%

Document 2: Typical Assignment, Face Sheet

Annenberg Research Project
Time Sampled Mathematics Assignment Form
Collection 1
6/00

For this assignment we **DO NOT NEED** student work.

This assignment will be collected during November 6, 2000

Please help us by providing the following information.

School: _____ **Grade**: _____ **Room Number**: _____

Teacher: Dr. Mrs. Mr. Ms. Last Name: _____ First Name: _____

Students Completed This Assignment: _____ In Class _____ At Home _____ Both (class & home)

Number of students in this class: _____

ASSIGNMENT TITLE: _____

ASSIGNMENT INSTRUCTIONS: ***Please attach a blank copy of any handouts students received for this assignment. If they worked out of a book, we will need a copy of the relevant pages. If you gave any instructions orally, please write down your instructions*** **as you would relay them to a student who was home sick**. If students were asked to use any special materials, please let us know what they were.

PLEASE PRINT

_____
_____
_____
_____
_____
_____
_____
_____
_____
_____

(*Please Use the Back if You Need More Space)*

Document 3: Challenging Assignment, Face Sheet, both sides completed

Annenberg Research Project
Challenging Writing Assignment Form
Collection 1 10/00

Your Annenberg researcher will work with you to arrange the best collection method for you.

Assignment will be collected on: _____/_____/_____

Please help us by providing the following information.

Assignment Due Date: _____/_____/_____ **Collection period**: **June 8, 2001**

School: Lincoln Elementary_____ **Grade**: __8_ **Room Number**: __433_____

Teacher: Dr. Mr. Mrs. Ms. Last Name: _____Boop_____ FirstName:__Betty_____

Students Completed This Assignment: __X__ In Class _____ At Home ____ Both (class & home)

Number of students in this class: __27_

ASSIGNMENT TITLE: _____

ASSIGNMENT INSTRUCTIONS: ***Please attach a blank copy of any handouts students received for this assignment. If they worked out of a book, we will need a copy of the relevant pages. If you gave any instructions orally, please write down your instructions*** **as you would relay them to a student who was home sick**. If students were asked to use any special materials, please let us know what they were. **"Student work submitted must be individual assignments, not group work." PLEASE PRINT.**

We are going to write a narrative essay describing a time when you were surprised. Remember, a narrative describes a personal experience or something you have seen. When you write your paper describe feelings. Use words that will help the reader to imagine how you felt when you were surprised. Also, don't forget to use transitional words to move from one point the next.

(*Please Use the Back if You Need More Space*)

Document 3 continued

Annenberg Researcher Use Only:

Total number of student assignments submitted __25_____
Reason for excluded assignments: Poor copy ____ No name ____ Missing __1__
Not Participating _1___
Other: _____

Table 4: Annual Schedule of Data Collection

	Year 1	Year 2	Year 3	Year 5
Typical Assignment	--	Fall	December	September
Typical Assignment	--	Fall	January	October
Typical Assignment	--	--	--	November
Challenging Assignment	--	Fall	January	January
Student Work from Challenging Assignment	--	Fall	January	January
Typical Assignment	Spring	Spring	March	January
Typical Assignment	Spring	Spring	May	February
Typical Assignment	--	--	--	March
Challenging Assignment	Spring	Spring	May	May
Student Work from Challenging Assignment	Spring	Spring	May	May
Challenging Assignment	Spring	--	--	--
Student Work from Challenging Assignment	Spring	--	--	--

Document 5

Passive Permission Form for Submission of Student Work for the Chicago Annenberg Research Project

September 2000

Dear Parent or Guardian:

Your son or daughter's school is participating in a study to help us learn about how student class work relates to student achievement. We are writing to ask your permission for your son or daughter to be a part of this effort.

Teachers in your child's school will be sharing the kinds of assignments that they give to students. We will also be collecting students' work on some assignments in order to track how students respond to different kinds of tasks. All student work will be kept strictly confidential. The results of our findings will only be reported for groups of students, such as "80% of sixth graders who were given complex mathematical work performed at high academic levels."

If you DO NOT want your son or daughter to participate, fill in the information below, and ask your child to return this sheet to his or her Writing or Mathematics teacher.

Thank you for your cooperation.

Sincerely,

Mark Smylie
Director, Annenberg Research Project

I DO NOT WANT my child, _____, to take part in the Annenberg Research Project.

_____ _____

Signature of Parent or Guardian Date

The Consortium is an affiliation of universities, educational organizations, and the Chicago Public Schools' research department, whose purpose is to conduct studies of school reform and school improvement and share the results widely.

The Consortium on Chicago School Research
1313 East 60th Street, Chicago, IL 60637

Document 6: Sample Student ID Database, Grade 3

STUDENT IDS GRADE 3

Room	Grade	Studentl	Studentf	Initial	CPS id	Passive Form	Comments	Chmath1	Chwrtg1
312	3	Bennett	Laura	A	36043206			1C03AM1Z	1C03AW1Z
312	3	Bradley	Karen		36036920			1C03AM1Z	
312	3	Cooper	Marcia	P	36072296			1C03AM1Z	1C03AW1Z
312	3	Harper	James		33305606			1C03AM1Z	1C03AW1Z
312	3	Jackson	Carl		38680587	1			
312	3	Smith	Leonard	J	40032967			1C03AM1Z	
312	3	Thompson	Sasha		40897046	1			

1 is used to identify students not participating.

Document 7: Annenberg Research Project, Random Number List (1 – 45)

29	22	20	13	9	18	39	14	31
34	8	40	43	38	32	30	44	12
33	6	39	20	38	40	26	14	20
26	10	31	0	28	3	15	26	39
20	29	44	22	12	13	36	32	0
4	3	29	28	23	32	25	31	27
37	43	19	18	2	4	9	23	10
25	31	1	41	30	6	15	39	25
24	26	25	27	42	21	43	22	5
35	5	9	22	19	24	20	39	25
39	1	33	14	42	29	8	6	15
42	22	37	3	5	2	19	6	31
30	34	15	11	18	9	14	29	7
40	41	17	24	18	27	34	13	18
12	18	25	16	4	13	4	32	15
11	29	21	8	39	25	12	15	41
11	27	5	30	20	23	9	45	19
39	28	11	9	22	41	38	41	6
27	21	41	5	36	28	19	23	6
7	34	14	20	11	13	16	5	26
20	2	26	44	22	41	29	34	41
39	22	36	16	15	31	27	28	41
14	37	13	4	44	17	42	40	43
28	44	11	38	38	8	9	22	33
17	9	5	2	2	2	19	2	35
31	37	42	27	3	1	39	26	24
13	17	44	27	16	40	10	35	43
18	10	19	32	33	2	1	42	11
22	28	21	10	45	18	21	35	11
41	39	19	24	25	40	5	14	0
10	25	15	15	23	18	12	35	14
42	34	24	22	27	34	5	22	14
8	25	12	34	42	2	44	34	27
15	31	32	44	29	30	12	6	27
16	22	7	4	8	30	5	8	27
38	13	8	8	17	27	45	30	18
1	20	26	5	6	37	21	30	45
9	33	13	13	21	30	9	29	15
11	40	14	6	32	19	10	24	31
42	26	13	0	45	37	39	34	42
19	39	38	43	16	22	33	39	33
45	2	5	8	41	43	21	9	11
36	2	44	28	28	22	11	19	33
43	26	7	27	36	1	9	33	16
20	9	27	1	10	27	2	37	18
6	40	15	9	37	13	26	24	29
42	12	10	19	44	33	22	13	24
1	25	32	2	23	21	37	42	27
33	2	8	26	7	22	18	1	5

Document 8: Teacher and School Codes Database

Description:	This database is designed to record and track teachers and schools participating in the fieldwork portion of the research project.
Archived:	This database is archived by project year.
Variables:	Schid, School, Tchrd, Room, Teacherf, Teacherl, Grade, Subject, Yr1, Yr2, Comments
Codes:	Apply to project year - ex. Yr1. Teachers participating in project year 1 receive a 1, if not participating a 0.
Linkable:	This database is linkable to the database called Teacher Codes and is linkable by TCHRID (teacher id), Assignments Collected Inventory, Observation Summaries

Document 8: Teacher and School Code Database continued

Schid	School	Tchrd	Room	Teacherf	Teacherl	Grade	Subject	Yr 1	Yr2	Comments
1	Cadillac	03A	312	John	Jones	3	M	1	1	
1	Cadillac	03B	311	Sharon	Johnson	3	M-W	1	1	
1	Cadillac	03C	314	Bob	Frost	3	M-W	1	0	
1	Cadillac	06A	318	Blake	Errington	6	W	0	1	
1	Cadillac	06B	320	Jim	Phelps	6	M-W	0	1	
1	Cadillac	06C	316	Susan	Stone	6	W	1	2	Tch now Gr 3
2	Lincoln	03A	313	Barbara	Jackson	3	M-W	0	1	
2	Lincoln	03B	332	Lynn	Johnson	3	M-W	0	1	
2	Lincoln	06A	412	Mary	Doe	6	W	0	1	
2	Lincoln	08A	433	Betty	Boop	8	W	0	1	
2	Lincoln	08B	434	Marcus	Garvey	8	M	0	1	Obsrvd only
3	Cougar	02A	101	Marcey	Finder	2	M-W	1	0	Obsrvd only
3	Cougar	03A	105	Dana	Smith	3	M-R-W	0	0	
3	Cougar	03B	108	William	Smith	3	M-W	0	1	
3	Cougar	03C	105	Kathleen	Guest	3	M-W	0	1	
4	Mustang	06A	303	Alma	Jackson	6	M	1	2	Withdrew MidYr
4	Mustang	06B	304	Mary	Dell	6	W	0	0	Withdrew
4	Mustang	08A	311	Teresita	Forest	8	M	0	1	
4	Mustang	08B	310	Kelly	Cane	8	W	0	1	

ey: Schid = School ID
Each school is given an ID number. The school ID is used when coding: assignments and observations.

School Name of schools that are being followed in the study.

Tchrd = Teacher ID
Each teacher in the study is given an ID that consist of their grade level. Attached to the grade level is a letter. If there are multiple teachers in a school at the same grade level a letter is added at the end of their code. The first teacher in the same grade (same school) is letter A, the 2nd is letter B and the 3rd is letter C.

Room
The room number that corresponds to the teachers/students that are being followed in the study.

TeacherF Teacher's First Name

TeacherL Teacher's Last Name

Grade
Grade of the teachers/students that is being followed in the study.

Subject
Subject of the teachers/students that is being followed in the study. M= Math R= Reading W=Writing

YrX
Study Year of the project in which the teachers participated.
0 and shading = participated in previous year of study, but not in the current year.
1 = active participation in the current project year. 2 = was participating but no longer is.

Document 9:

Annenberg Research Project
Coding Assignments

We place important identifying codes *(names are not used when coding assignments, interviews, observations, etc. for confidentiality purposes)* on all the data we collect. The following codes will help us place these ID codes on our assignments and observations.

Time sampled assignment codes

Type of assignment	Type of collection	Colored labels (to label work collected)
Writing = W	Time sampled = T	3rd grade = yellow
Math = M	Challenging = C	6th grade = blue
		8th grade = red

Collection period (*varies depending on project*)

Time Sampled Tasks	Challenging Tasks (student work)
Collection 1 = 1	Collection 1 First Semester = 1
Collection 2 = 2	Collection 2 Second Semester = 2
Collection 3 = 3	
Collection 4 = 4	
Collection 5 = 5	
Collection 6 = 6	

Teacher Ids	School Name	School ID
Betty Boop = 08A	Cadillac	1
Marcus Garvey = 08B	Cougar	3
	Lincoln	2
	Mustang	4

ASSIGNMENT IDS

The assignment ID includes the following identifying pieces of information: School ID, Type of collection, Teacher ID, Type of assignment, Collection period, and Project Year

Example: I collect a writing assignment from Betty Boop, an 8th grade teacher at Lincoln school during the first timed sample collection.

From the list of codes that follow we figure out that....

School ID = 2
Type of collection = T
Teacher ID = 08A
Type of assignment = W
Collection period = 1
Collection year = A (*use alpha coding. For example a project consist of 4 years of research: Year 1 = A, Year 2 = B, Year 3 =C, Year 4 = D)*
Color of label = red.

Thus, the Assignment ID = 2T08AW1A (on a red label)

Document 10: Sample Student Work with Identifiers Removed

2C08BW2B 35624555

James Harper Room 434

2/19/01

Mr. Garvey

A Surprise

I was surprised with a trophy for soccer. We won the championship game. We had 9 points and they had 0. My friend and classmate Francisco made 5 goals and I made 4 goals. The coach took the team to eat at Little Caesers on 93rd. We played video games and 3D games. While we were at Little Caesers I got a trophy. I did not know that I was getting a trophy. I thought Francisco was going to get it. I was surprised that I got the trophy. Mr. Garvey is our soccer coach and my gym teacher at Lincoln Elementary School. It was a great day for me.

Good work James, but you forget to include

Document 11: Cleaning and Submitting Tasks and Student Work
Annenberg Research Project
(*For Research Assistants*)

Assignments (Tasks)

1) Make sure the correct form was used and all blanks have been filled.
2) Make sure that the instructions match the assignment and that all necessary attachments are included. (Remove anything that will identify the school)
3) Edit "teacher's instructions" to remove unnecessary information, such as a lesson plan for the week or previous assignments. Sometimes it does help to have a quick review of the previous assignments leading up to this one, but it must be brief.
4) Label the assignment form using the same color coding scheme as the time-sampled assignments. If the assignment includes attachments, staple all the attachments and place an identical assignment label on the first page of the attachments as you did on the assignment form.

Student Work

1) Make sure all the **student work matches the assignment**. If it does not match write on the top right corner **"wrong assn"** and move to the back of the stack.
2) Make sure each piece of student work has the **student's name** on it. If it does not have a name on it write on the top right corner **"no name"** and move to the back of the stack.
3) Separate the **poorly copied pieces** of student work from the rest. Place the poorly copied pieces at the back of that assignment stack and write ***"bad copy"*** on each on the top right hand corner. **Note: if the work is illegible due to poor penmanship we still include it.**
4) Also move to the back any piece of student work for those students whose parents **returned a passive permission form**. You'll find this information on excel sheets labeled Non-participants. On the top right hand corner of these students' work write **"NP"** to the back of the stack.
5) Count the number of student work that are **good copies. Number each piece of student work 1-?? On the top left hand corner with a red pen.**
6) Fill in all blanks in the *Annenberg Researcher Use Only* box on the back of the assignment form.
7) Match all the good pieces of student work with the student's id and place label on the top right hand corner of the page.
8) Select 10 pieces of student work from each assignment stack. Number them with a red pen in the upper right hand corner of the work. We'll do this using a **random numbers list**.
9) Remove all student and school identifying information from the 10 selected pieces only. This includes student name and school name from the heading and also from the content of the student work. Also remove neighborhood names, specific street addresses, and names of relatives if both first and last names were provided (remove first or last name-no need to remove both). We'll do this with a correction stick and black markers.

Document 12: Assignments Collected Inventory Database

Name: **Assignments Collected Inventory**

Description: This database is designed to track assignments and student work collected from teachers participating in the fieldwork data collection portion of the research project.

Archived: A project varies in duration. This database should be archived by project year as school and teacher participation varies from year to year.

Variables:

SCHOOL - School Name

SCHID - ID assigned to school

TCHRID - ID assigned to teacher

RM# - Room number that assignments and student work are being collected from

GRADE - For teachers and students participating in collection

WRTG1- 1st Writing assignment collected for study year
Subject and number of collections varies from project to project.

CHWRTG1 - 1st Challenging assignment with corresponding student work collected for study year

Codes:

1 - Assignment submitted by teacher

X - Assignment not submitted by teacher *(teacher doesn't teach the subject, no longer is participating in the study, assignment not submitted)*

Linkable: This database is linkable to the database called Teacher Codes and is linkable by TCHRID(teacher id).

Document 12: Assignments Collected Inventory Database continued

	SCHID	TCHRID	RM#	GRADE	WRTG1	WRTG2	CHWRTG1	CHWRTG2	MATH1	MATH2	CHMATH1	CHMATH2
Cadillac	1	03A	312	3	1	1	1	1	1	1	1	1
Cadillac	1	03B	311	3	1	1	1	1	1	1	1	1
Cadillac	1	06B	305	6	1	1	1	1	X	X	X	X
Cadillac	1	06E	318	6	1	1	1	1	1	1	1	1
Cadillac	1	06G	316	6	1	X	X	X	X	X	X	X
Lincoln	2	03E	313	3	1	1	1	1	1	1	1	1
Lincoln	2	03F	332	3	1	1	1	1	1	1	1	1
Lincoln	2	06A	412	6	1	1	1	1	X	X	X	X
Lincoln	2	06B	413	6	X	X	X	X	1	1	1	1
Lincoln	2	08A	433	8	1	1	1	1	X	X	X	X
Lincoln	2	08B	434	8	X	X	X	X	1	1	1	1
Mustang	3	03B	108	3	1	1	1	1	1	1	1	1
Mustang	3	03C	105	3	1	1	1	1	1	1	1	1
Mustang	3	06D	303	6	X	X	X	X	1	1	X	X
Mustang	3	06F	304	6	X	X	X	X	X	X	X	X
Mustang	3	08E	311	8	X	X	X	X	1	1	1	1
Mustang	3	08F	310	8	1	1	1	1	X	X	X	X

Key: ASSGN = Assignment (task and student work)
SCHID = School ID
TCHRID = Teacher ID
WRTG1 = Writing Assignment 1 *(first collection for study year)*
CHWRTG1 = Challenging Writing Assigment/student work *(first collection for study year)*

1 = Assignment Submitted by Teacher
X = Assignment Not Submitted *(Teacher doesn't teach the subject, no longer participating, assignment not collected)*

Document 13: Electronic Version of Teacher Instructions

The Chicago Annenberg Research Project
Writing Task Cover Sheet

TASK ID: 2C08BW1Z

INSTRUCTIONS TO STUDENTS:

We are going to write a narrative essay describing a time when you were surprised. Remember, a narrative describes a personal experience or something you have seen. When you write your paper describe feelings. Use words that will help the reader to imagine how you felt when you were surprised. Also, don't forget to use transitional words to move from one point to the next.

Document 14: Student Work Sample Database

Task ID/Student ID	Grade	Coll Year
1C06AW1B 32094678	6	1
1C06AW1B 35632808	6	1
1C06AW2B 31576148	6	1
2C06AW1B 33409338	6	1
2C06AW1B 33676476	6	1
2C06BW1B 32900909	6	1
3C06BW1B 32391001	6	1
3C06BW2B 28694407	6	1
4C06CW1B 13161410	6	1
4C06CW1B 32452132	6	1
4C06DW2B 34060320	6	1

Document 15: Mathematics Rubric Manual

The following information has been reformatted for this Appendix. When using these rubrics for scoring, the Overview and General Rules sections were repeated with each standard to remind scores of the correct process to use. The rubric booklet used during scoring also was formatted so pages had considerable amounts of white space and scorers were encouraged to take notes and write in the booklet.

MANUAL FOR SCORING TASKS and STUDENT WORK IN MATHEMATICS

August, 2001

The writing portion of this manual was prepared by Eric Gutstein, Fred Newmann, Jean Biddulph, and Judy Merlau It is based on and quotes from F. M. Newmann, W. G. Secada, and G. G. Wehlage (1995), *A Guide to Authentic Instruction and Assessment: Visions, Standards and Scoring*. Madison, WI: Wisconsin Center on Education Research.

Material in this manual is under continuing development and may be revised based on the results of research and training.

INTRODUCTION

The standards in this manual are being used to describe the quality of teachers' assignments and students' work in writing and mathematics as part of the Chicago Annenberg Research Project. The standards are intended to measure intellectual activities that reflect analysis and extended communication in these subjects.

Teachers in participating Annenberg schools have cooperated in sharing assignments and students' work to advance this research. Other area teachers have also contributed to this effort by using these standards to score assignments and student work. The broad-based participation among Chicago schools and teachers is helping us learn about student performance in non-test settings and will help us describe how schools can advance school development in ways that support teachers.

While this manual is being used only for research purposes, educators have expressed interest in the standards, and we are distributing them for their information. However, these standards should not be mechanically used or adopted to judge teachers' practices or students' achievement. If the standards are to be beneficial to teachers and students, they require extensive study, discussion, and possible revision to meet the unique circumstances of individual schools and student groups.

By disseminating these standards, we hope to encourage further dialogue and discussion of criteria for the pursuit of intellectual quality in our schools.

Overview and General Rules

Our responsibility is to estimate the extent to which successful completion of the task requires the kind of cognitive work indicated by each of three standards: Knowledge Construction, Written Mathematical Communication, and Connections to Students' Daily Lives.

Each standard will be scored according to different rules, but the following apply to all three standards:

a. If a task has different parts that imply different expectations (e.g., worksheet/short answer questions and a question asking for explanation of some conclusions), the score should reflect the teacher's apparent dominant or overall expectations. Overall expectations are indicated by the proportion of time or effort spent on different parts of the task and by criteria for evaluation if stated by the teacher.

b. Scores should take into account what students can reasonably be expected to do at the grade level.

c. When it is difficult to decide between two scores (e.g., a 2 or a 3), give the higher score only when a persuasive case can be made that the task meets minimal criteria for the higher score.

d. If the specific wording of the criteria is not helpful in making this judgment, base the score on the general intent or spirit of the standard described in the introductory paragraphs of the standard.

Standard 1: Knowledge Construction

The task asks students to mathematically organize and mathematically interpret information in addressing a mathematical concept, problem, or issue.

Consider the extent to which the task asks the student to mathematically organize and mathematically interpret information, rather than to retrieve or to reproduce fragments of knowledge or to repeatedly apply previously learned algorithms and procedures.

Possible indicators of requiring mathematical organization are tasks that ask students to decide among algorithms, to chart and graph data, or to solve multi-step problems.

Possible indicators of requiring mathematical interpretation are tasks that ask students to consider alternative solutions or strategies, to create their own mathematical problems, to create a mathematical generalization or abstraction, or to invent their own solution method.

These indicators can be inferred either through explicit instructions from the teacher or through a task that cannot be successfully completed without mathematical organization and/or mathematical interpretation.

3= The task calls for mathematical interpretation of information. Tasks that require mathematical interpretation are assumed to require mathematical organization.

2= The task calls for mathematical organization of information, but minimal or no mathematical interpretation.

1= The task calls for very little or no mathematical organization and mathematical interpretation of information. Its dominant expectation is for students to retrieve or reproduce fragments of knowledge or to repeatedly apply previously learned algorithms and procedures.

Standard 2: Written Mathematical Communication

The task asks students to demonstrate and/or elaborate their understanding, ideas, or conclusions through written mathematical communication.

Consider the extent to which the task requires students to elaborate on their understanding, ideas, or conclusions through written mathematical communication.

Possible indicators of requiring written mathematical communication are tasks that ask students to generate prose (e.g., write a paragraph), symbolic representations (e.g., graphs, tables, equations) , diagrams, or drawings. The score depends on how the task requires students to elaborate their understanding, ideas, or conclusions.

To define some terms we use below, a *solution path* is a trace of work done to answer the problem; an *explanation* is a justification and/or presentation of the reasons for a student's choices.

4 = *Analysis/Persuasion/Theory*. The task requires the student to show his/her *solution path* and to *explain* the solution path with evidence.

3 = *Report/Summary*. The task requires the student to show her/his *solution path* but does not require *explanation* of why.

2 = *Short-answer exercises.* The task requires little more than giving a result or following a worked-out example. Students are asked to show some work.

1 = *No extended writing.* The task requires no or minimal written mathematical communication, for example, only giving mathematical answers or definitions.

Standard 3: Connections to Students' Lives

The task asks students to address a mathematical question, issue, or problem that may be similar to one that they have encountered or are likely to encounter in daily life.

Consider the extent to which the task presents students with a question, issue, or problem that they have actually encountered, or are likely to encounter in daily life.

Possible indicators of connections to students' lives are real-world tasks like estimating personal budgets or computing discounts (but completing a geometric proof generally would not be considered a real-world task).

Certain kinds of mathematical knowledge may be considered valuable in social, civic, or vocational situations in daily life (e.g., knowing basic arithmetic facts, or percentages). However, task demands for "basic" knowledge will not be counted here unless the task requires applying such knowledge to a specific problem likely to be encountered in daily life.

3 = The mathematical question, issue, or problem clearly resembles one that students have encountered or are likely to encounter in daily life. The connection is so clear that teacher explanation is not necessary for most students to grasp it.

2 = The mathematical question, issue, or problem bears some resemblance to one that students have encountered or are likely to encounter in daily life, but the connections are not immediately apparent. The connections would be reasonably clear if explained by the teacher, but the task directions need not include such explanations to be rated 2.

1 = The mathematical question, issue, or problem has virtually no resemblance to one that students have encountered or are likely to encounter in daily life. Even if the teacher tried to show the connections, it would be difficult to make a persuasive argument.

STUDENT WORK SCORING MANUAL

Overview and General Rules

Our task is to estimate the extent to which the student's performance illustrates the kind of cognitive work indicated by each of three standards: Mathematical Analysis, Mathematical Concepts, and Written Mathematical Communication.

Each standard will be scored according to different rules, but the following apply to all three standards:

a. Scores should be based only on evidence in the student's performance relevant to the criteria. Matters such as whether the student followed directions, neatness, correct spelling, etc. should not be considered unless they are relevant to the criteria.

b. Scores may be limited by tasks which fail to demand mathematical analysis, mathematical conceptual understanding, or written mathematical communication, but the scores must be based only upon the work shown.

c. Scores should take into account what students can reasonably be expected to do at the grade level.

d. Scores should be assigned only according to the criteria in the standards, not relative to other papers that have been previously scored.

e. When it is difficult to decide between two scores (e.g., a 2 or a 3), give the higher score only when a persuasive case can be made that the paper meets minimal criteria for the higher score.

f. If the specific wording of the criteria is not helpful in making this judgment, base the score on the general intent or spirit of the standard described in the introductory paragraphs of the standard.

g. Completion of the task is not necessary to score high.

Standard 1: Mathematical Analysis

Student performance demonstrates thinking about mathematical content by using mathematical analysis.

Consider the extent to which the student demonstrates thinking that goes beyond mechanically recording or reproducing facts, rules, and definitions or mechanically applying algorithms.

Possible indicators of mathematical analysis are organizing, synthesizing, interpreting, hypothesizing, describing or extending patterns, making models or simulations, constructing mathematical arguments, or inventing procedures.

The standard of mathematical analysis calls attention to the fact that the content or focus of the analysis should be mathematics. There are two guiding questions here:

- First, has the student demonstrated mathematical analysis? To answer this, consider whether the student has organized, interpreted, synthesized, hypothesized, invented, etc. or whether the student has only recorded, reproduced, or mechanically applied rules, definitions, algorithms.
- Second, how often has the student demonstrated mathematical analysis? To answer this, consider the proportion of the student's work in which mathematical analysis is involved.

To score 3 or 4, there should be no significant conceptual mathematical errors in the student's work, however, the analysis does not need to be at a high conceptual level to score a 3 or 4.

If the student showed work indicating significant analysis, but the answer was incorrect, score it a 2.

If the student showed only the answer to a problem, and it is incorrect, score it 1**.**

If the student showed only the answer to a problem, and it is correct, decide how much analysis is involved to produce a correct answer, and score according to the rules below.

4 = Mathematical analysis was involved throughout the student's work.

3 = Mathematical analysis was involved in a significant portion of the student's work.

2 = Mathematical analysis was involved in some portion of the student's work.

1 = Mathematical analysis constituted no part of the student's work.

Standard 2: Mathematical Concepts

Student performance demonstrates understanding of important mathematical concepts central to the task.

Consider the extent to which the student demonstrates understanding of mathematical concepts. A score of 1 or 2 may be due to a task that fails to require demonstration of substantial or exemplary understanding of mathematical concepts. For example, a task that requires students to mechanically record or reproduce facts and definitions or mechanically apply algorithms may not provide students the opportunity to demonstrate substantial or exemplary understanding of the mathematical concepts central to the task.

Possible indicators of understanding important mathematical concepts central to the task are expanding upon definitions, representing the concept in alternate ways or contexts, or making connections to other mathematical concepts, to other disciplines, or to real-world situations.

Correct answers can be taken as an indication of the level of conceptual understanding if it is clear to the scorer that the task or question requires conceptual understanding in order to be completed successfully. Thus, even if no work is shown, a score of 3 or 4 may still be given.

The score should *not* be based on the proportion of student work central to the task that shows conceptual understanding but on the *quality* of that understanding *wherever* it occurs in the work.

4= The student demonstrates an exemplary understanding of the mathematical concepts that are central to the task. Their application is appropriate, flawless, and elegant.

3= There is substantial evidence that the student understands the mathematical concepts that are central to the task. The student applies these concepts to the task appropriately; however, there may be minor flaws in their application, or details may be missing.

2= There is some evidence that the student understands the mathematical concepts that are central to the task. Where the student uses appropriate mathematical concepts, the application of those concepts is flawed or incomplete.

1= There is little or no evidence that the student understands the mathematical concepts that are central to the task, or the mathematical concepts that are used are totally inappropriate to the task, or they are applied in inappropriate ways.

Standard 3: Written Mathematical Communication

Student performance demonstrates and/or elaborates her or his understanding, explanations, or conclusions through written mathematical communication.

Consider the extent to which the student elaborates on their understanding, ideas, conclusions through written mathematical communication.

Possible indicators of written communication are diagrams, drawings, or symbolic representations (e.g., graphs, tables, equations), as well as prose.

The score should not be based on the proportion of student work central to the task that contains explanation/argument/representation but on the quality of written mathematical communication, wherever it may occur in the work.

To score high on this standard, the student must communicate in writing an accurate, clear, and convincing explanation or argument that justifies the mathematical work.

4= Mathematical explanations or arguments are clear, convincing, and accurate, with no significant mathematical errors.

3= Mathematical explanations or arguments are present. They are reasonably clear and accurate, but they may be less than convincing, slightly flawed, or incomplete in minor ways.

2= Mathematical explanations, arguments, or representations are present. However, they may not be finished, may omit a significant part of an argument/explanation, or may contain significant mathematical errors.

1= Mathematical explanations, arguments, or representations are absent or, if present, are seriously incomplete, inappropriate, or incorrect. This may be because the task did not ask for argument or explanation (e.g., fill-in-the-blank, multiple-choice questions, or reproducing a simple definition in words or pictures)

Document 16: Rubric for Writing Manual

The following information has been reformatted for this Appendix. When using these rubrics for scoring, the Overview and General Rules sections were repeated with each standard to remind scores of the correct process to use. The rubrics booklet used during scoring also was formatted so pages had considerable amounts of white space and scorers were encouraged to take notes and write in the booklet.

MANUAL FOR SCORING TASKS and STUDENT WORK IN WRITING

July, 2001

The writing portion of this manual was prepared by David Jolliffe, Fred Newmann, Anna Chapman, Carmen Manning, and Kendra Sisserson. It is based on and quotes from F. M. Newmann, W. G. Secada, and G. G. Wehlage (1995), *A Guide to Authentic Instruction and Assessment: Visions, Standards and Scoring*. Madison, WI: Wisconsin Center on Education Research.

Material in this manual is under continuing development and may be revised based on the results of research and training.

INTRODUCTION

The standards in this manual are being used to describe the quality of teachers' assignments and students' work in writing and mathematics as part of the Chicago Annenberg Study Project. The standards are intended to measure intellectual activities that reflect analysis and extended communication in these subjects.

Teachers in participating Annenberg schools have cooperated in sharing assignments and students' work to advance this research. Other area teachers have also contributed to this effort by using these standards to score assignments and student work. The broad-based participation among Chicago schools and teachers is helping us learn about student performance in non-test settings and will help us describe how schools can advance school development in ways that support teachers.

While this manual is being used only for research purposes, educators have expressed interest in the standards, and we are distributing them for their information. However, these standards should not be mechanically used or adopted to judge teachers' practices or students' achievement. If the standards are to be beneficial to teachers and students, they require extensive study, discussion, and possible revision to meet the unique circumstances of individual schools and student groups.

By disseminating these standards, we hope to encourage further dialogue and discussion of criteria for the pursuit of intellectual quality in our schools.

STANDARDS AND SCORING CRITERIA FOR WRITING

Overview and General Rules: Tasks

The main point here is to determine the extent to which successful completion of the task requires the kind of cognitive work indicated by each standard.

A. If a task has different parts that imply different expectations (e.g., worksheet/short answer questions and a question asking for explanation of some conclusions), the score should reflect the teacher's apparent dominant or overall expectations. Overall expectations are indicated by the proportion of time or effort spent on different parts of the task and by criteria for evaluation, if stated by the teacher.

B. Scores should take into account what students can reasonably be expected to do at the grade level.

C. When it is difficult to decide between two scores (e.g., a 2 or a 3), give the higher score only when a persuasive case can be made that the task meets minimal criteria for the higher score.

D. If the specific wording of the criteria is not helpful in making judgments, base the score on the general intent or spirit of the standard described in the guidelines of the standard.

STANDARDS AND SCORING CRITERIA FOR WRITING

A. TASKS

Standard 1: Construction of Knowledge

GUIDELINES:

The task calls for interpretation, analysis, synthesis, or evaluation of information.

CRITERIA:

3 = The task's dominant expectation is for students to interpret, analyze, synthesize, or evaluate information, rather than merely to reproduce information.

2 = There is some expectation for students to interpret, analyze, synthesize, or evaluate information, rather than merely to reproduce information.

1 = There is no or virtually no expectation for students to interpret, analyze, synthesize, or evaluate information. The dominant expectation is that students will merely reproduce information gained by reading, listening, or observing.

B. TASKS

Standard 2: Elaborated Written Communication

GUIDELINES:

The task asks students to draw conclusions or make generalizations or arguments AND support them through extended writing.

CRITERIA:

4 = Explicit Call for Generalization AND Examples. The task asks students, using narrative or expository writing, to draw conclusions or to make generalizations or arguments, AND to substantiate them with illustrations, details, or reasons.

3 = Call for Generalization OR Examples. The task asks students, using narrative or expository writing, either to draw conclusions or make generalizations or arguments, OR to offer illustrations, details, or reasons, but not both.

2 = Short-answer exercises. The task or its parts can be answered with only one or two sentences, clauses, or phrasal fragments that complete a thought.

1 = Fill-in-the-blank or multiple-choice exercises

STANDARDS AND SCORING CRITERIA FOR WRITING

C. TASKS

Standard 3: Connection to Students' Lives

GUIDELINES:

The task asks students to connect the topic to their lives.

CRITERIA:

3 = The task explicitly asks students ,to connect the topic to experiences, or situations in their lives.

2 = The task offers the opportunity for students to connect the topic to experiences, feelings, or situations in their lives, but does not explicitly call for them to do so.

1 = The task offers no or virtually no opportunity for students to connect the topic to experiences, feelings, or situations in their lives.

STANDARDS AND SCORING CRITERIA FOR WRITING

Overview and General Rules: Student Work

A.Scores should be based only on evidence in the student's writing relevant to the criteria. Matters such as following directions, neatness, correct spelling, etc. should not be considered unless they are relevant to the criteria.

B. Scores may be limited by tasks which fail to call for construction of knowledge or elaborated written communication, but the scores must be based only upon the work shown.

C. Scores should take into account what students can reasonably be expected to do at the grade level. However, scores should still be assigned only according to "absolute" criteria in the standards, not relative to other papers that have been previously scored.

D. When it is difficult to decide between two scores (e.g., a 2 or a 3), give the higher score only when a persuasive case can be made that the paper meets minimal criteria for the higher score.

E. If the specific wording of the criteria is not helpful in making judgments, base the score on the general intent or spirit of the standard described in the guidelines.

STANDARDS AND SCORING CRITERIA FOR WRITING

A. STUDENT WORK

Standard 1: Construction of Knowledge: Interpretation, Analysis, Synthesis, or Evaluation

GUIDELINES:

The writing demonstrates interpretation, analysis, synthesis, or evaluation in order to construct knowledge, rather than merely to reproduce information. Such interpretation, analysis, synthesis, or evaluation must appear to be reasonably original.

This standard is intended to measure the extent to which the student writing goes beyond mechanically recording, reporting, or otherwise reproducing information. The essential question is whether students demonstrate construction of knowledge by means of thinking and organizing information, versus reproduction of knowledge by means of restating what has been previously given to them.

CRITERIA:

4 = Substantial evidence of construction of knowledge. Almost all of the student's work shows interpretation, analysis, synthesis, or evaluation.

3 = Moderate evidence of construction of knowledge. A moderate portion of the student's work shows interpretation, analysis, synthesis, or evaluation.

2 = Some evidence of construction of knowledge. A small portion of the student's work shows interpretation, analysis, synthesis, or evaluation.

1 = No evidence of construction of knowledge. No portion of the student's work shows interpretation, analysis,
synthesis, or evaluation; OR virtually all construction of knowledge is in error.

STANDARDS AND SCORING CRITERIA FOR WRITING

B. STUDENT WORK

Standard 2: Elaborated Written Communication

The writing demonstrates an elaborated, coherent account that draws conclusions or makes generalizations or arguments and supports them with examples, illustrations, details, or reasons.

GUIDELINES:

Elaboration consists of two parts: a conclusion, generalization, or argument AND support for it, in the form of at least one example, illustration, detail, or reason. Elaboration is coherent when the examples, illustrations, details, or reasons do indeed provide appropriate, consistent support for the conclusions, generalizations, or arguments.

To use the criteria, the scorer should identify specific points in the student work that are elaborated and should make a judgment about their coherence.

CRITERIA:

4 = Substantial evidence of elaboration. Almost all of the student's work comprises an elaborated, coherent account.

3 = Moderate evidence of elaboration. A moderate portion of the student's work comprises an elaborated, coherent account.

2 = Some evidence of elaboration. A small portion of the student's work comprises an elaborated, coherent account.

1 = No evidence of elaboration. No portion of the student's work comprises an elaborated, coherent account.

STANDARDS AND SCORING CRITERIA FOR WRITING

C. STUDENT WORK

Standard 3: Grammar, Usage, Mechanics, and Vocabulary

The writing demonstrates proficiencies with grammar, usage, mechanics, and vocabulary appropriate to the grade level.

GUIDELINES:

This standard is intended to measure the degree to which students attempt to, and succeed at, using language structures at the sentence and word level to make their meaning understandable to readers.

Scorers should take into consideration the efforts students might make at trying out new language structures that represent a "stretch" for someone at their grade level and not fault students if these "stretch" efforts are not carried off with complete success.

Scorers should assess the quality of the actual written work and not take into consideration possible effects of a student's possible linguistic background or learning disability.

Illegible handwriting could result in a score of 2 or 1.

CRITERIA:

4 = The student writing is an excellent demonstration of grammar, usage, mechanics, and/or vocabulary appropriate for the grade level. There are no errors, or if there are a few errors, the errors present no problem for understanding the student's meaning, nor does the performance compromise the student's credibility.

3 = The student writing is a satisfactory use of grammar, usage, mechanics, and/or vocabulary for the grade level. There are some errors, but they present no problem for understanding the student's meaning.

2 = There are many errors in grammar, usage, mechanics, and/or vocabulary, OR the errors in grammar, usage, mechanics, and/or vocabulary make it difficult, but not impossible, to understand the student's meaning.

1 = The use of grammar, usage, mechanics, and/or vocabulary is so flawed that it is not possible to understand the student's meaning.

Document 17

Chicago Annenberg Research Project:
Quality of Intellectual Work Scoring Process
Amount of Assignments to be Scored Summer 2001

Writing:

	YEAR 5 Expected year-end	YEAR 1 Rescore	YEAR 2 Rescore	YEAR 3 Rescore
Grade 3	190	50	50	50
Grade 6	167	50	50	50
Grade 8	127	50	50	50

Total Writing 484 new + 450 rescored = **934 total writing assignments**
(With double scoring of half, scorers handle 1401 pieces)

Math:

	YEAR 5 Expected year-end	YEAR 1 Rescore	YEAR 2 Rescore	YEAR 3 Rescore
Grade 3	190	50	50	50
Grade 6	131	50	50	50
Grade 8	105	50	50	50

Total Math 426 + 450 rescored = **876 total math assignments**
(With double scoring of half, scorers handle 1314 pieces)

1810 assignments to be scored total (With double scoring 2715 handled)

Amount of Student Work to be Scored Summer 2001

Writing:

	YEAR 5 Expected year-end	YEAR 1 Rescore	YEAR 2 Rescore
Grade 3	500	50	50
Grade 6	480	50	50
Grade 8	320	50	50

Total Writing 1300 new + 300 rescored = **1600**
(With double scoring of half, scorers handle 2400 pieces)

Math:

	YEAR 5 Expected year-end	YEAR 1 Rescore	YEAR 2 Rescore
Grade 3	500	50	50
Grade 6	380	50	50
Grade 8	280	50	50

Total Math 1160 + 300 rescored = **1460**
(With double scoring of half, scorers handle 2190 pieces)

3060 pieces of student work to be scored total (With double scoring 4590 handled)

Document 18

Schedule for scoring sessions

Mathematics Schedule for Scoring Summer, 2001

Day 1

8:30	Orientation
9:00	Train Task Std 1
10:50	Break
11:05	Score Task Std 1
11:35	Refresher training
11:55	Score
12:15	Lunch
1:00	Review training Task Std 1
1:10	Score
1:45	Train Task Std 2
3:15	Break
3:30	Score Task Std 2
4:00	Refresher training
4:20	Score
5:00	Adjourn

Day 2

8:30	Review Training Task Std 2
8:45	Score
9:00	Train Task Std 3
10:30	Break
10:45	Score Task Std 3
11:15	Refresher Training
11:35	Score
Noon	Lunch
12:45	Review Training Task Std 3
12:50	Score
1:20	Train Student Work Std 1
3:00	Break
3:15	Score Student Work Std 1
3:45	Refresher Training
4:05	Score
5:00	Adjourn

Day 3

8:30	Review Training Student Work Std 1
8:50	Score
10:00	Break
10:15	Train Student Work Std 2
11:45	Lunch
12:30	Review Training Student Work Std 2
12:40	Score
1:10	Refresher Training
1:30	Score
3:30	Break
3:45	Train Student Work Std 3
5:00	Adjourn

Day 4

8:30 Training Student Work Std 3
8:50 Score Student Work Std 3
9:20 Refresher Training
9:40 Score
10:15 Break
10:30 Score
Noon Lunch
12:45 Debrief
2:00 Adjourn

Document 19

Annenberg Scoring Project: Year 3 Tentative Schedule

Action	*Days Reqd*	*Responsibility*	*Dates*	*Days*
Continuing				
Cleaning and labeling check	cont	Loretta/Nick/Verity	now through June/July	
Entering data to AssignYr3 database	cont	Loretta	now through June/July	
Cover sheets (writing)	cont	Loretta	now through June/July	
Cover sheet with conceptual domain	cont +10[1]	Loretta/Trainers	April/May and July/August	
Checking deliverables (tasks and other items) collection is up-to-date	cont	Loretta	now through July	
Editing rubrics	cont	Fred/Trainers	now? through August	
Yr1 tasks back from UIC storage		Mark	before June 1	
Prep for Pilot[2]				
Selection of scorers for pilot (teacher hiring)		CTC? with David & Rico?	April/May	
Selection of tasks for training packets	10[3]	Trainers	by Wednesday 4/28 (NB: AERAs 4/18-23)	
Schedule rooms and catering	0	Loretta	completed – will require final check Monday 5/17	
Stationery supplies	cont	Loretta		
Double-check for 'clean' tasks	0.5	Sabrina/Nick	April 29	Thursday
Photocopying selected tasks (x c.10 per packet)	5	Sabrina/Nick/Loretta/ Verity	April 30 – May 7	Friday - Friday
Labeling packets (including set-up)	2	Verity/Nick/Sabrina	May 5-7	Wed – Friday
Compiling scoresheets per trainers' selection	2	Verity	May 12-14	Wed – Friday
Editing rubrics	1	Fred/Trainers/Verity	May 13	Thursday
Packing tasks and scoresheets into packets	4	Sabrina/Nick/Loretta/ Verity	May 14-19	Friday - Wed
Double-check packets' contents for accuracy	1	Sabrina/Nick/Loretta/ Verity	May 20	Thursday
Pilot Scoring: Math	**1**		**May 21**	**Friday**
Pilot Scoring: Writing	**1**		**May 22**	**Saturday**
Re-filing originals	2	Sabrina	May 21-24	Friday - Monday
Unpacking and discarding photocopies	1	Sabrina	May 25	Tuesday
Prep for Final Scoring				
Selection of scorers (final teacher hiring)		CTC (with David & Rico?)	June	
Selection of Yr1 and Yr2 tasks for re-scoring	0.5	Jenny	June 21	Monday
Pulling Yr1 and Yr2 tasks from boxed and drawer files	4	Sabrina	June 22-25	Tues - Friday

[1] Conceptual domains on math tasks are decided by each trainer in a block system. Prior to the Pilot and Scoring sessions, they will need to come in to the CARP office, go through their grade's files and allocate a domain for each task. This then goes to Loretta who will enter in the domain, re-print the cover sheet and she (or Sabrina) will then re-file the original task.

[2] UC Spring Quarter ends Friday 6/11. As Sabrina and Verity will still be working part-time, all tasks requiring their involvement will necessarily take a longer time-span to complete

[3] Each trainer takes approximately half a day to select tasks for the Training session. In a perfect world the process would take one or two days, but trainers come in at different times. Two weeks is a general time-span.

Schedule rooms and catering	0.5	Loretta	completed – will require final check Monday 8/2	
Stationery supplies	cont.	Loretta	continuing	
Final task collection tally	0	Loretta	July 9	Friday
Inclusion of Yr1 and Yr2 task ids into Yr3 data	1	Verity	July 9	Friday
Random sort of all task id's	2	Verity/Nick	July 12-14	Monday – Wed
Matrices (Excel sheets)	4	Verity	July 14-19	Wed – Monday
Final conceptual domains for Math	15[4]	Trainers/Loretta	by July 21	Wed
Compiling scoresheets (Access)	4	Verity	July 20-23	Tues – Friday
Double-check for 'clean' tasks	2	Sabrina[5]/Nick	July 22-23	Thurs – Friday
Labeling packets	2	Verity/Nick/Sabrina	July 26-27	Monday – Tues
Packing tasks into packets	5	Sabrina/Nick/Loretta/ Verity	July 27 – August 2	Tues – Monday
Editing rubrics	2	Fred/Trainers/Verity	August 2-3	Monday – Tues
Double-check packets' contents for accuracy	2	Sabrina/Nick/Loretta/ Verity	August 3-4	Tues – Wed
Final Scoring: Writing	**2**		**August 5-6**	**Thurs – Frid**
Final Scoring: Math	**2**		**August 9-10**	**Mon – Thurs**
Logging scores on to Excel database	5	Verity[6]/Nick/Sabrina	August 8/6, 10-13	Frid, Tues-Friday
Double-check of scores	5	Verity/Nick/Sabrina	August 16-20	Mon – Friday
Filing away all packets for Yr3 scoring	2	Sabrina/Nick/Loretta/ Verity	August 23-24	Mon – Tuesday
Filing away Yr1 and Yr2 tasks back into correct packets	2	Sabrina/Nick	August 25-26	Weds - Thursday

[4] As per task selection (note 3), this process is not immediately completed.
[5] Sabrina may be away for a research trip during July/August. This schedule presumes her place will be taken by a temp employee (who has had a couple of weeks with the team to prepare)
[6] Verity may be away for a fieldwork research trip. If I do go, I won't leave until after the scoring sessions are completed. Nick helped me complete this task last year and knows all that is involved.

Document 20

Year 5 Scoring: The How To Manual

Note: This is written for a complete novice – if you're an old hand at Annenberg, my apologies for the superfluous, even pedantic, bits.

Vocab

"CARP" *Chicago Annenberg Research Project – the people you're working with; the project that looks at around 20 schools within the whole Annenberg Challenge.*

"pieces" *the generic name for the student work or teacher tasks that are being scored (sometimes "documents")*

"student work" *actual home or classwork produced by a student*

"tasks" *actual home or class assignments created by a teacher*

"challenging" *collected twice during the year, the teacher is asked by the researcher for something that they consider is a challenging assignment rather than the..*

"typical" *ordinary piece, collected four times in the year*

"rubrics" *the guideline manual against which all the pieces are scored*

"matrix" *the configuration of pieces in separate packets, your raison d'être*

"packets" *the manila folders into which bundles of pieces, along with their scoresheets, are packed*

What's being scored?

In year 5 student work and teacher tasks ("pieces") in both math and writing will be scored from Grades 3, 6, 8 and 9/10. Researchers during the school year have collected these pieces. The scoring process is overseen by Fred Newmann, two training group leaders: Rico Gutstein (math) and David Jolliffe (writing), six trainers (one for each grade in each subject – Rico and David also take a grade), and a recruited team of teachers from the Chicago area (who are not from the CARP schools). Fred and the trainers are responsible for preparing the rubric against which the work is scored, and for training over overseeing the teachers in the scoring process. Pieces are scored against three standards – see the rubrics. There are separate sets of standards for writing and math, and for student work and teacher tasks.

You have two main points of contact: Loretta (for all data management questions, such as 'how many pieces of work do we have?') and Fred (for all methodological questions, such as 'does this matrix for scoring look good?'). Also talk with Stacy ('how many scorers do we have?'), and Rico and David ('when is your team coming to choose Pilot tasks?').

When is it being scored?

There are two (four, to be precise if we take the subjects separately) sessions for scoring. In the first, the Pilot Sessions (Math and Writing, in May), the rubrics are being tested and a group of teachers gathered for the second, Scoring session. Although the teachers will receive training on the scoring process, this is not meant as a training session – although it can be used to 'weed out' potential scorers, its primary aim is to train the trainers and make any necessary changes in the training format. In the second, Scoring sessions (Math and Writing, in August), the teachers will be trained as they score the work. Math and Writing each have their separate sessions in both Pilot and Scoring.

In the Pilot session, the trainers choose sample pieces that represent a good cross-section of scoring possibilities. For this, they will delve into the year's collections and will give you a list of *about* six to eight pieces per standard (hence, e.g., 6 pieces per standard, 18 pieces per grade, 72 pieces per subject, 144 total Tasks). You will need to collect these lists, produce them on a table and coordinate with Loretta in taking those pieces out of the files, photocopying them (the originals will stay in the files) and producing the packets. You will also need to produce scoresheets for these tasks. The Pilot session looks much the same as the Scoring session, but it (1) does not require you to produce a matrix and (2) does not use all the pieces collected during the year. It is a good opportunity to understand something of what the whole process entails.

The following refers to the Scoring process itself.

Things you will have to produce

- The schedule (with Loretta)

- The matrix – which shows what pieces will go in which packets, and which scorers will score them for particular standards. The most important, and the most mind numbing of everything you do.
- The scoresheets

1. The Schedule of Events

As many different things need to be coordinated, a schedule helps to make sure everyone involved is doing what they need to be doing when they need to be doing it. See the Schedule example from 1999 for the time frames. Important things to include:

- Deadlines for receipt of collections from researchers – and double-checking (even triple-checking) that all pieces of work are clean (i.e. anonymous – by date, school name, locality, student or teacher name – anything that could identify the piece of work to its origin), properly labeled with their ID number and ready for scoring. Loretta deals with and will help you with much of this.
- Following the deadline for collections, constructing the matrices – at least two weeks as it will take some time to understand what you are doing, and to check back with Fred at various stages of the process.
- Rubrics (i.e. the training manuals) need to be edited and prepared – by the training group leaders (probably Rico (math) and David (writing)) – both for the Pilot and for the Scoring sessions.
- Pieces for the Pilot session will need to be chosen by the trainers – approximately 6 per subject per grade per standard.
- Training pieces for the Scoring session will also need to be chosen by the trainers, as well as 'refreshers' which are other samples used throughout the scoring process to check that everyone is on the right track.
- Time for you to produce scoresheets – including figuring out how to do it, and printing them all out (as of 1999, each standard should be printed on differently colored paper to make life a little easier for scorers)
- Labeling and preparing packets for Pilot and for Scoring. In the Pilot, all the pieces will have to be photocopied as each scorer is trained on the same pieces. In the Scoring, you will use the

originals of everything (but don't forget that the training pieces will need to be photocopied as they are also scored). All Years 1-3 pieces have been filed away according to the packet they were scored in – for the purposes of future analysis it is pointless to dismantle the packets and put everything back in their school files.

2. The Matrix

You will need to know:

- The final tally of pieces and their ID numbers on an Excel spreadsheet (get these from Loretta)[7]
- The final number of scorers for each grade in each subject (from Stacy)
- The approximate number of pieces in each packet – usually 10/11, but check with Fred before you go too far down the line

The general idea of the matrix is to decide who will score what packet on what standard, and then who will double-score. Double scoring (to ensure some reliability) is usually done on half of the packets across each standard.

The first step is to randomly sort all of the ID numbers. Once you get them all from Loretta, I suggest dividing them into four separate Excel files: (1) Math Student Work, (2) Math Tasks, (3) Writing Student Work, (4) Writing Tasks. Note that Student Work will have two ID numbers, one that relates to the actual Task involved, the other the actual document the student created. These two ID numbers on Student Work need to be kept beside each other but in separate columns so the Task ID can be pulled during analysis.

In each of the Excel files you can then create separate worksheets for each grade.[8] Each sheet will need the following columns:

ID #	Grade	Original sort number	Randomly sorted number	Packet Number	Standard 1 Scorer 1	Standard 1 Score 1	Standard 1 Scorer 2	Standard 1 Score 2	Standard 2 Scorer 1	etc to	Standard 3 Score 2

[7] Note that researchers will often collect more than 10 pieces of student work from each teacher. They will however choose 10 pieces to work with using a random number list. Not much of an issue for you, but the answer to any confusion you might have when you look in the files and see a lot of pieces that don't have an ID number.

[8] Use a header that automatically gives the name of the file and the sheet (the date's also useful) – as reference for everyone involved.

ID# here means Task ID in the Teacher Tasks sheets, and signifies two columns (Task ID and Student ID number) in the Student Work sheets.

Most of these columns can be 'hidden' until you need them, especially the score columns as these won't be needed until you have the scores back at the end of the sessions.

Hence, you will have a long list of ID numbers divided according to grade and subject. These ID numbers however will not be randomly sorted, so that's your first task:

- First, fill in the 'Original Sort Number' column, numbering each row 1 to *n* (where *n* is the total number of pieces) so that you have a record of their original order (this is especially important for the Student Work files). [9]
- Now create a list of randomly sorted numbers – even if it's just to take say 40 pieces of paper, number them 1-40 and draw them randomly from a bag, whatever works. This randomly sorted list can be used over and over for each grade, for each subject etc. It's useful to leave that list on a spare sheet on each of your Excel files so you have a quick record to copy and paste from.
- Go through each sheet, filling in the random sorted list on the 'Randomly sorted number' column.
- Now highlight the whole sheet, and sort on that 'Randomly sorted number' column
- Your IDs will now be in a random sort, BUT they don't have an actual randomly sorted number. Go to the column and number each ID 1 to *n*, all the way down.

This random sort is what you will work with from now on.

Now you have to work out the packet numbers. This means deciding how many packets there will be in each grade so that you can allocate the pieces to their packets, with an eye on how many scorers you have. In a perfect world you would have say 160 pieces and 10 pieces per packet and 4 scorers and everything works just perfectly (16 packets, 4 packets per scorer) but odds are it just won't be that way, so...

[9] Fast way to do this? Start with the first, enter '1'. Enter '2' on the second. Now highlight both those boxes. You will see a small black box on the bottom right of the highlighted area. Click and drag this all the way down to the end of the column and Excel will automatically number each box for you. Note that you have to do at least 1 and 2. If you only enter '1' and then click and drag, Excel will input a '1' into each box all the way down. This latter method is useful in the next stage for filling in Packet Numbers.

Packet Numbers

First, find out the essential figures for each subject in each grade. For an example, let's say you're doing Grade 3 Math Tasks. Loretta's list (your Excel sheet) tells you there are 175 tasks in grade 3 math. Stacy tells you there will be 7 scorers for grade 3 math. Fred says he wants about 10 pieces per packet.[10]

So, in this example you have 175 tasks, 7 scorers, and you want to aim for around 10 pieces per packet.

175 tasks with 7 scorers = 25 tasks per scorer

Find the nearest common multiple, assuming 10 pieces per packet:

17 packets of 10 will use up 170 pieces – you have five left over, so re-arrange to:

12 packets of 10 and 5 packets of 11, uses up all 175 – but this is not very equal, so see if there's a better way to work it by adjusting the packet sizes to 9 pieces or 11 pieces. In this example, 11 pieces will work a little more evenly – at 15 packets of 11 and 1 packets of 10.

Follow this formula for all of the grades, all of the subjects, in both tasks and student work – you will end up with 16 packet configurations (4 x 2 x 2). Send these to Fred for his approval. Fred will be working out the amount of time there is for each packet to be scored, and will also want to make sure there is a fairly even spread of labor across the scorers.

Once Fred gives his OK, you can go back to your spreadsheets with your packet figures.

Now allocate the packet numbers to the IDs. To continue with the example, go to your worksheet for Grade 3 Math Tasks. In the 'Packet Number' column, fill in the packet numbers, starting with '1' and fill that in for the first 11 IDs. Then fill in the next 11 IDs with '2'. And so on, until the last packet number (16) which will be for the last 10 IDs. It's crucial, as with all stages, to get this right, and a quick way to make sure is to check that as you drag and fill the boxes, Excel sums up

[10] At this stage you will find that Grade 3 packets are always going to be a little bigger than the other grades (Grades 9/10 will be a little smaller), simply because we have more elementary schools than high schools in our survey and as a result, a lot more pieces to be scored. Consequently if you're working with 'about' 10 pieces per packet, you can almost be certain that Grade 3 will need 11 pieces per packet, and Grades 9/10 will probably have 9 pieces per packet.

your entries to the 'right' figure. So in this example, as you drag and fill your first 11 boxes (see end of footnote 2), Excel will sum to 11, on Packet 2 it will give you a sum of 22 (11 x 2), and so on, through Packet 15 (where your total should be 165). On Packet 16 it should sum to 160 as you are only using the last 10 pieces.

And follow the same routine for all the other worksheets.

Now you're ready for...

The Matrix Itself!

Hide every column and row you don't need for this stage. That will include: score columns (you do, very desperately, need all the scorer columns), the grade column, original sort # and randomly sorted # columns, and then only keep the rows which contain the first ID number of each packet – i.e. hide pieces 2 to 11 of Packet 1 (in the above example), pieces 2 to 11 of Packet 2, etc. Using our example, you will be left with 16 rows showing that contain the first ID number in each packet. Having the ID number showing is not strictly necessary either – hide that column now if you feel so inclined. Following the example, you should now have something that looks like this (I've just invented realistic ID numbers for this example):

ID #	Packet Number	Standard 1 Scorer 1	Standard 1 Scorer 2	Standard 2 Scorer 1	Standard 2 Scorer 2	Standard 3 Scorer 1	Standard 3 Scorer 2
6T03BM3	1						
12T03AM2	2						
3T03CM1	3						
	etc to 16						

Each packet will be scored on each standard once. Half the packets will be scored a second time (this is where the 'Scorer 2' columns come in). First you need to allocate packets for the first scoring to each scorer. In this example, you have 7 scorers, A through G. Randomly sort these 7 letters, and fill them in, in order, down through each column, top to bottom, for each standard, first scorer:

ID #	Packet Number	Standard 1 Scorer 1	Standard 1 Scorer 2	Standard 2 Scorer 1	Standard 2 Scorer 2	Standard 3 Scorer 1	Standard 3 Scorer 2
6T03BM3	1	D		B		F	
12T03AM2	2	A		E		D	

3T03CM1	3	C		G		A	
	etc to 16						

i.e., take your random listing of letters and keep using them up in order from top to bottom, then left to right, across the standards so that each column begins with a different letter. The essential things to keep in mind are:

1. that ***no scorer should score the same packet twice***,[11] to ensure general reliability
2. that everyone should be 'kept busy' so make sure they are evenly distributed throughout – to make sure no one is sitting on their thumbs while everyone else is madly working
3. that each scorer should be paired with each other scorer in each standard, or at least as far as possible – to reveal the 'heavy' and 'light' scorers.[12] (OK, this is more relevant to the double scoring stage, but it's good to keep these checks together.)

Although trainers tend to ad-lib a little on the day as they hand out the packets to the scorers, try to make sure that allocations aren't bunched up too much, for instance that Scorer B does Packet 12 Standard 1, first scoring, and then immediately goes to Packet 13 Standard 1 first scoring. These errors are contained in this example layout:

ID #	Packet Number	Standard 1 Scorer 1	Standard 1 Scorer 2	Standard 2 Scorer 1	Standard 2 Scorer 2	Standard 3 Scorer 1	Standard 3 Scorer 2
6T03BM3	1	D		D		B	
12T03AM2	2	A		F		G	
3T03CM1	3	C		F		E	
	etc to 16						

A *very* useful thing to do at this stage is to tally every scorer's commitments at the end of each column, getting something that looks a little like this (I'm not following the grade 3 example here, but it's pretty easy to figure out what your tallies should look like):

	Scorer							Total
	A	4		2		4		10

[11] One caveat to this: you may have less than 5 scorers in one matrix. Here you encounter a problem: to make sure that no one scores a packet more than once, you need at least five scorers (three standards, and two of them double-scored – this may seem confusing, but once you see the end result you'll understand how this works). At this stage, do the best you can – at least make sure that if, for instance scorer B scores on standard 1 first scoring, s/he doesn't see the packet again until standard 3.

[12] i.e. in a five-scorer set you should have A score with B and with C and with D and with E (and so on BC, BD, BE, CD, CE, and DE) at least once during each standard. You may not have enough opportunities for everyone to be paired off, but if this is so, don't let one pair score together twice if another pair are to be left out.

	B	4		4		2		10
	C	4		4		3		11
	D	4		3		4		11
	E	4		2		4		10
	F	2		4		4		10
	G	2		4		4		10

Sometimes things just aren't even, but try to distribute out again at this stage so that it improves. For instance in this example, there could have been 6 scorers doing 10 and 1 doing 12, but I changed it to 5 doing 10 and 2 doing 11.

Now you need to allocate the double scoring. It's a matter of choice where you begin (Packet 1 or Packet 2), but two standards will be double-scored on one set of packet numbers and the other will be double-scored on the other set. Confused? The result will look a little like this:

ID #	Packet Number	Standard 1 Scorer 1	Standard 1 Scorer 2	Standard 2 Scorer 1	Standard 2 Scorer 2	Standard 3 Scorer 1	Standard 3 Scorer 2		
6T03BM3	1	D	G	B		F	E		
12T03AM2	2	A		E	B	D			
8TO3CD1	3	C	E	G		A	F		
2T03AC3	4	B		F	A	C			
10T03CD2	5	G	A	D		B	C		
7T03PN1	6	D		A	F	G			
12T03AB2	7	E	D	C		F	B		
3T03CM1	8	D		B	G	E			
	etc to 16	A	C	G		F	D		
AB		BC	0, 0, 1	CD		DF	0, 0, 1	FG	
AC	1,	BD		CE	1,	DG	1,	FH	
AD		BE	0, 1,	CF		DH			
AE		BF	0, 0, 1	CG		EF	0, 0, 1		
AF	0, 1, 1	BG	0, 1,	CH		EG			
AG	1,			DE	1,	EH			

I find it useful to gray-shade the boxes that don't need a scorer. You can see that there are no duplications across the columns. At this stage you will probably find that a scorer has to do a double scoring almost immediately after they do a first scoring (look at B in Standard 2), just try to keep these occurrences to a minimum.

In addition it is helpful to keep track of the pairs for each standard at the bottom of each matrix. Fill in the times pairs are together in each of the standards to see that the pairing of work is evenly distributed. Notice in the above sample that AF are paired in Standard 2 and Standard 3. While a number of pairs (such as AB, AD, AE) have so far not been paired at all (Of course, we've only bothered to show 8 of the 16 rows of this matrix).

As you go through each standard, fill in your tally so you know that everyone is getting a fair share of labor. You will find that as you get to Standard 3 Scorer 2, things will get very tricky. Knowing who can spare what, and who needs a little extra work, is invaluable at this stage. Your tally will hopefully look something like this:

	Scorer							Total
	A	4	2	2	3	4	2	17 (7)
	B	4	2	4	2	2	3	17 (7)
	C	4	2	4	2	3	2	17 (6)
	D	4	2	3	2	4	2	17 (6)
	E	4	2	2	3	4	2	17 (7)
	F	2	3	4	2	4	2	17 (7)
	G	2	3	4	2	4	2	17 (7)

The figures in the total column here indicate that Scorer A has scored a total of 17 times, of which 7 where double-scorings. Scorers C and D just got lucky on this run and had less to do than everyone else. Otherwise, things were well distributed and no one had to 'sit on their hands': if they had less single-scorings to do, they had an extra double-scoring to make up for it.

This is a long and mind-boggling process – especially as you begin to even out the allocations and correct any errors. But once you get the hang of how it all fits together (concentrate on one matrix and understand how it all works before you go to the next one) it really does get easier. Once your matrix is complete double check for all the possible errors: tally up everything one more time, make sure there are no duplications across the columns, make sure no one is bunched around a few packets, etc. Then check it again.

After all the matrices are done (reckon on the better part of a week for the tasks and at least a week for student work) and you've gone through plenty of checking, send them to Fred and Tony Bryk.

Once Fred gives his approval to everything, print out a set for the trainers' use, as well as any that you feel might be useful in the present format. You're about to unhide all the rows: things are about to look very different.

Unhide all the rows and fill in all the Scorer information (using the drag and fill function) so that you have something like this:

ID #	Packet Number	Standard 1 Scorer 1	Standard 1 Scorer 2	Standard 2 Scorer 1	Standard 2 Scorer 2	Standard 3 Scorer 1	Standard 3 Scorer 2
6T03BM3	1	D	G	B		F	E
12T03AM1	1	D	G	B		F	E
7T03BM3	1	D	G	B		F	E
8C03AM2	1	D	G	B		F	E
9C03BM1	1	D	G	B		F	E
5T03DM4	1	D	G	B		F	E
3T03CM1	1	D	G	B		F	E
8C03AM1	1	D	G	B		F	E
4T03EM3	1	D	G	B		F	E
12T03AM2	2	A		E	B	D	
	etc	A		E	B	D	

You need your matrices in this format for packing the packets and for creating scoresheets. Once every necessary column and row is showing (you may also be helpful to the packing team if you unhide the Grade column) print them all out and coordinate with Loretta to get everything packed into the right packets.

This is a critical stage where the right Task or Student Work MUST go into the right packet, in the right order. It involves many hands and a lot of tedium but requires a lot of concentration. Once everything is packed, you'll need to make sure it all gets double-checked carefully, you can be 99% sure there'll be some slip-up somewhere.

Loretta will coordinate with you about getting everything ready for the scoring sessions – putting the packets into large boxes in order of standard. Boxes will contain scoring packets, training packets, scoresheets (in packets, clearly labeled), matrices long and short, see below.

As well as the 'short' version of the matrix (the one with any superfluous information hidden and concentrating on the packet allocation) it is useful to provide the trainers with a 'long', complete, version of the matrix, one that shows all the ID numbers, just in case anything goes terribly wrong.

3. The Scoresheets

There are two options here, depending on your knowledge of software. You will need to merge your Excel sheets with a scoresheet template in either Word or Access. Either option has its advantage and disadvantage, personally I prefer Word because the logic of Access remains a mystery to me. Having used both options, I find merging in Word faster and simpler. So here's how to do it in Word.

The main document should follow the template for all previous years' scoresheets – see attached. You will first need to specify the Excel spreadsheet you are using (note that when Word looks for a linked file, it defaults to looking for Word files, you will have to tell it to look for an Excel file). To prepare for this, make sure that you save the Excel file you need with the correct worksheet on the front, so if you're doing Grade 3, make sure it is listed ahead of all the other sheets (not just showing on your screen). When Word looks into that Excel file, it will choose the first spreadsheet it sees, and if you don't have Grade 3 showing, it will choose whatever grade is at the top of the 'line'.

Once you have your main document linked to the matrices, you can then specify the fields it needs. This is where Word has a disadvantage over Access. You will need to create a table in the main document in which it can place the ID numbers, but unlike Access, the number of rows won't be automatically adjusted to fit the amount of pieces on each scoresheet. So when there's an uneven distribution of pieces over the packets you'll be left with some extra lines showing, and you're going to have to make adjusting empty rows in the Excel spreadsheet – for instance, if your packet composition is 15 packets of 10 and one packet of 11, fifteen of the scoresheets are going to have a blank line at the end of the table, and you're going to have to go into the Excel spreadsheet and insert a blank line at the end of each set of the first fifteen packets. (The advantage with Word is that the whole process is a lot easier to understand and get going than it is with Access, unless you're an Access wizard.) (It's a matter of aesthetics versus practicality.)

When you're ready (having double-checked everything to make sure that the IDs are grouped together correctly on the right scoresheets – take a look at the end of the pack) print out the sheets on colored paper according to the standard.

Training and Refresher Pieces

This process looks a little like the Pilot Session – the trainers will come in, choose some good examples as training pieces and 'refreshers' (used during the session if the scoring group takes a break in the middle of scoring one Standard, for the next morning, etc). You need to (1) coordinate with Loretta on pulling these pieces from the files, making copies of everything requested for the scorers, the trainer, and the office copy; and (2) make up scoresheets for each training set. Refreshers can be listed with the training pieces on the same scoresheet, just leave a space to differentiate between the two. Note that the scoresheet doesn't just have a space for the 'score', it looks more like this, which gives more space for discussion and understanding:

Task ID	Your Score	Consensus Score	Comments

Receiving the scores

As the sessions get going, you will need to take care of two processes: (1) checking the reliability of double-scoring (i.e. how much are scorers giving a piece the same score); and (2) entering the scores on to the spreadsheet and sending them for data analysis.

Entering the scores is self-explanatory: reveal the hidden columns and start punching in the figures. Begin this job as soon as you start getting scores back during the sessions – you'll be getting one standard at a time from each grade. It is *extremely* important to be accurate here, so do the job with someone sitting with you who call out the scores. Then double-check everything.

Fred is looking for a reliability check on the rate at which scorers gave a piece the same score, differed by one point, or differed by two. This should be expressed in percentages – so if there were 100 pieces being double-scored and 70 were given the same score by both scores, the reliability rate on 0 difference was 70%. You can start to do this as the scoresheets come back, either using the paper copies or doing it on the spreadsheets with an extra column. The latter is a little easier to keep track of, and you're doing two jobs at one time: entering the scores, and checking their accuracy.

Once all of this is done you need to (1) send the scores data off to the data analysis people, and (2) organize the filing of all the packets.

There are a few extra columns you can add to the sheets to facilitate the analysis stage:

- School Number (taken from the Task ID)
- Teacher ID (taken from the Task ID)
- Network Number (see attached)

Remove the tallies and combine all the worksheets on to one so that the Excel file you send them has only one worksheet containing all the data they need. A printout of Year 2's Writing Student Work – the final copy – is attached

to give you some idea of what to produce. This one contains the Room Number, originally from Loretta's data, but this doesn't seem to be necessary.

Be intelligent about what needs to be filed – only one copy of the training packets, rubrics, all of the scoresheets – make sure the packets are in order, and follow the example of previous years' filing.

Network Numbers

School Number (get this from Task ID)	Network Number
1	7
2	4
3	9
4	5
5	4
6	5
7	7
8	8
9	8
10	9
11	2
12	2
13	10
14	11
15	12
16	11
17	13
18	8
19	12
20	14
21	13
22	14
23	10

Document 21

August 2001

Grade 3 CONNECTION TO STUDENTS' LIVES

Scorer Name: _____

When assigning a score please check the ID number on the score sheet with the item you are scoring. Keep all items in the packet in order of their listing on the score sheet.

Task ID Student #	Your Score	Consensus Score	Comments
1T03BM3Z			
1T03BM4Z			
2T03FM6Z			
23C03FM1Z			
4T03CM6Z			
1C03IM2Z			
Refresher			
12T03BM1Z			
4T03BM4Z			

Document 22

APPENDIX FACETS CONTROL FILE

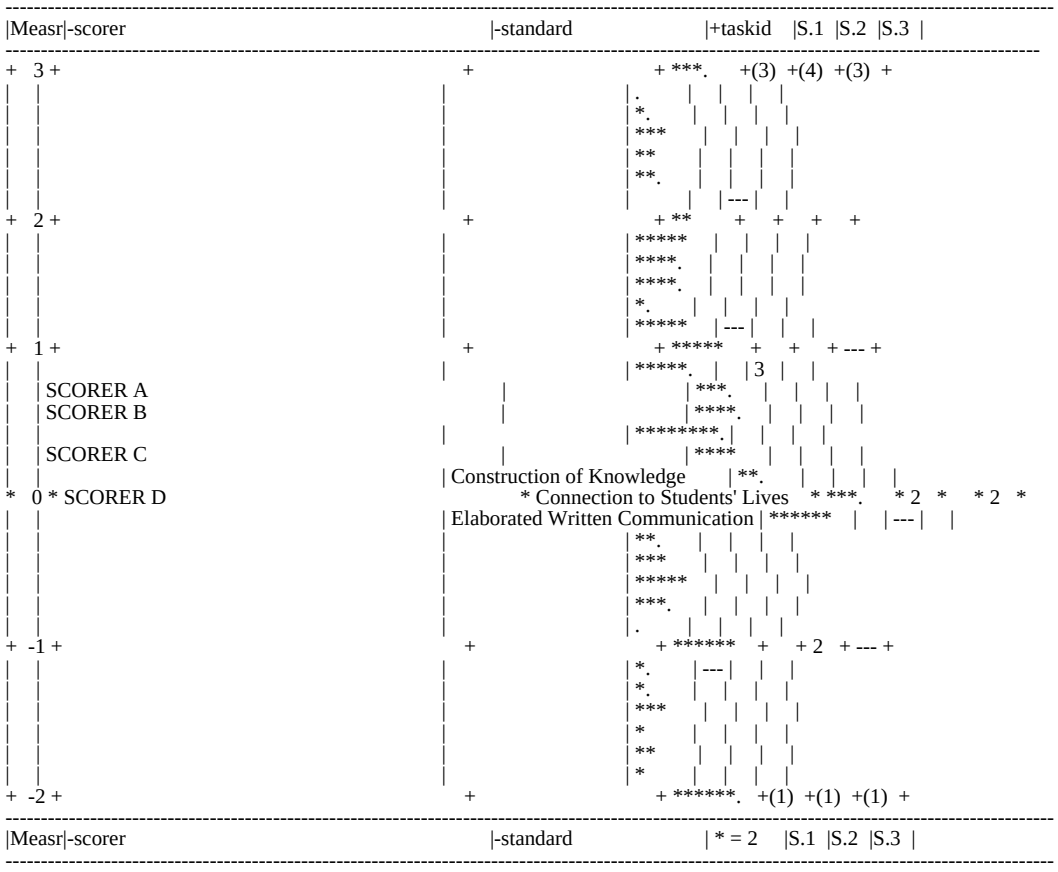
```
Title= Math for Grade 6 tasks
Output=t6mth_99.out
facets=3
positive=3
noncentered=3
arrange=m,N
missing=.
Query=YES
model=?B,1B,?,stan1
model=?B,2,?B,stan2
model=?,3,?,stan3

rating scale=stan1,r
1=no/little math org or interpretation
2=math org & no/minimal interpretation
3=both math org & interpretation
*
rating scale=stan2,r
1=no extended writing
2=short answer
3=report/summary
4=analysis/persuasion/theory
*
rating scale=stan3,r
1=no connection
2=some connection
3=clear connection
*

Labels=
1,scorer,
31=scorer A,0,3
32=scorer B,0,3
33=scorer C,0,3
34=scorer D,0,3
35=scorer E,0,3
*

2,standard
1=Knowledge Construction
2=Written Mathematical Communication
3=Connection to Students' Lives
*
3,taskid
1 = 11C06EM1Y
2 = 11C06EM2Y
3 = 11T06DM4X
4 = 11T06EM1Y
5 = 11T06EM2Y
*
data=
```

35 , 1 , 1 , 2
. , 1 , 1 , .
34 , 2 , 1 , 2
33 , 2 , 1 , 1
31 , 3 , 1 , 1
32 , 3 , 1 , 1
33 , 1 , 2 , 1
34 , 1 , 2 , 1
31 , 2 , 2 , 1
. , 2 , 2 , .
32 , 3 , 2 , 1
. , 3 , 2 , .
35 , 1 , 3 , 1
33 , 1 , 3 , 1
32 , 2 , 3 , 2
. , 2 , 3 , .
31 , 3 , 3 , 1
. , 3 , 3 , .
32 , 1 , 4 , 1
33 , 1 , 4 , 1
35 , 2 , 4 , 1
. , 2 , 4 , .
33 , 3 , 4 , 1
33 , 1 , 5 , 1
34 , 1 , 5 , 1
31 , 2 , 5 , 1
. , 2 , 5 , .
32 , 3 , 5 ,

Document 23

APPENDIX: Table 6.0 for CUTPOINTS

Table 6.0 All Facet Vertical "Rulers".
Vertical = (1A,2A,3*) Yardstick (columns, lines) = 0,8



Document 24

Writing for Grade 3 tasks
Table 8.1 Category Statistics.

Model = ?B,1B,?,STAN1
Rating scale = STAN1,R,G,O

DATA			FIT		STEP		EXPECTATION	MOST	THURSTONE	Cat	Response
Category	Counts	Cum.	Avge	OUTFIT	CALIBRATIONS		Measure at	PROBABLE	THRESHOLD	PEAK	Category
Score	Used	%	Meas	MnSq	Measure	S.E.	Category -0.5	from	at	Prob	Name
1	112	28%	-.99	1.0			(-2.02)	low	low	100%	no or little expectation
2	157	40%	.09	1.1	-.83	.14	.00	-1.15	-.83	-.98	53%	some expectation
3	126	32%	100%	1.04	1.2	.83	.13	(2.03)	1.17	.83	.97	100%	dominant expectation
							(Mean)	(Modal)	(Median)		

Writing for Grade 3 tasks
Table 8.2 Category Statistics.

Model = ?B,2,?B,STAN2
Rating scale = STAN2,R,G,O

DATA			FIT		STEP		EXPECTATION	MOST	THURSTONE	Cat	Response
Category	Counts	Cum.	Avge	OUTFIT	CALIBRATIONS		Measure at	PROBABLE	THRESHOLD	PEAK	Category
Score	Used	%	Meas	MnSq	Measure	S.E.	Category -0.5	from	at	Prob	Name
1	52	13%	-1.48	.9			(-2.80)	low	low	100%	fill in blank or mult-choice
2	102	26%	39%	-.45	.9	-1.56	.18	-.98	-1.99	-1.56	-1.77	48%	short answer
3	161	41%	80%	.80	.8	-.29	.13	.87	-.11	-.29	-.19	58%	generalization or example
4	81	20%	100%	1.55	1.1	1.85	.15	(3.01)	2.10	1.85	1.95	100%	generalization and example
							(Mean)	(Modal)	(Median)		

Writing for Grade 3 tasks
Table 8.3 Category Statistics.

Model = ?,3,?,STAN3
Rating scale = STAN3,R,G,O

DATA			FIT		STEP		EXPECTATION	MOST	THURSTONE	Cat	Response
Category	Counts	Cum.	Avge	OUTFIT	CALIBRATIONS		Measure at	PROBABLE	THRESHOLD	PEAK	Category
Score	Used	%	Meas	MnSq	Measure	S.E.	Category -0.5	from	at	Prob	Name
1	117	30%	30%	-1.16	.8			(-1.83)	low	low	100%	no/minimal opportunity
2	136	34%	64%	.23	.8	-.59	.14	.00	-1.03	-.59	-.80	47%	some opportunity
3	143	36%	100%	1.13	1.1	.59	.13	(1.85)	1.04	.59	.80	100%	explicit opportunity
							(Mean)	(Modal)	(Median)		

Document 25

Research Assistant Positions Available

Chicago Annenberg Research Project at the Consortium on Chicago School Research

Research Assistants are required for the upcoming academic year, September 2000 to June 2001. A variety of half-time and quarter-time appointments are available.

The Chicago Annenberg Research Project is a longitudinal study of the Chicago Annenberg Challenge, a $50 million grant to selected Chicago Public Schools. With qualitative and quantitative work conducted in a range of these schools, the Project analyses the changes and improvements influenced by the Challenge.

In the fifth and final year of the Project's fieldwork, Research Assistants are needed to help with the collection, management and analysis of data. Suitable applicants are likely to be graduate students in the social sciences, have an interest in urban school development, and willing to undertake a variety of tasks. Responsibility and reliability are vital. Spanish speakers would be a bonus.

Two main areas of work need to be covered: fieldwork in the schools and/or administrative work in the Project's office (located at the University of Chicago). Fieldwork in the schools can involve classroom observations, interviews, and collection of student assignments and work products. The exact requirement for each Research Assistant will depend on the Project's needs and applicant's interests and experience, and will be discussed on further appointment. Applicants must be available during the day to visit schools and attend project meetings. Individuals with experience working in schools or conducting field-based research are preferred. Administrative work at the Project's Offices will be as an assistant to the Data Manager, responsible for data entry and storage. As abilities and needs allow, this work may be combined with fieldwork in the schools.

This is an excellent opportunity for a graduate student interested in urban school development. Data from the Chicago Annenberg Research Project offer individual research possibilities for students interested in curriculum and instruction, student learning and development, school organization, anthropology, sociology, psychology and public policy. Field positions offer experience in ethnographic research methods.

Applicants should forward a letter of interest and a resume, with names of two referees, to:

Verity Elston, Assistant Director of Fieldwork, Chicago Annenberg Research Project,
Consortium on Chicago Schools Research, 1313 E 60th Street, Chicago IL 60615
Fax: 773 702 2010
Enquiries: Tel 773 834 ----